International Series in
Operations Research & Management Science

Brian T. Denton *Editor*

# Handbook of Healthcare Operations Management

## Methods and Applications

Springer

International Series in Operations Research & Management Science

Volume 184

For further volumes:
http://www.springer.com/series/6161

Brian T. Denton
Editor

# Handbook of Healthcare Operations Management

## Methods and Applications

Springer

*Editor*
Brian T. Denton
Department of Industrial
and Operations Engineering
University of Michigan
Ann Arbor, MI, USA

ISSN 0884-8289
ISBN 978-1-4614-5884-5 ISBN 978-1-4614-5885-2 (eBook)
DOI 10.1007/978-1-4614-5885-2
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013930226

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Preface

The unprecedented availability of data now affords the opportunity to improve decision making and inform the scientific discovery of best practices for healthcare delivery. This book will serve as a valuable reference for researchers interested in a survey of the state of the art in healthcare operations research and methods that can exploit the opportunities afforded by the available data. It is also intended to be a resource for practitioners interested in identifying opportunities for the implementation of operations research methods to improve healthcare operations. It is suitable for use as a supplementary text for educators offering graduate or senior undergraduate level classes in industrial engineering departments, schools of public health, and business schools.

This book builds on a long history of research and practice involving the application of operations research methods to healthcare delivery. Early work includes inventory planning for blood banks, appointment scheduling at outpatient clinics, and the deployment of emergency vehicles over a geographic region. In recent years there has been a resurgence of interest in healthcare operations management. This has been driven in part by rising costs of healthcare in many countries and concerns about timely access and the quality of care. This book is a unique compilation of chapters on emerging topics including optimization in resource-constrained settings, modeling behavioral aspects of patient care, advances in supply chain management, and the coordination of decision making among multiple parts of an integrated health system. These new applications are, in turn, driving the development of new simulation, optimization, and stochastic models and enriching the methodological foundations of operations research.

Collectively, the chapters in this book address application domains including inpatient and outpatient services, public health networks, supply chain management, and resource-constrained settings in developing countries. Many of the chapters provide specific examples or case studies illustrating the applications of operations research methods across the globe, including Africa, Australia, Belgium, Canada, the UK, and the USA. Chapters 1–4 review operations research methods that are most commonly applied to healthcare operations management including queuing, simulation, and mathematical programming. Chapters 5–7 address challenges related to inpatient services in hospitals such as surgery, intensive care units, and hospital wards. Chapters 8–10 cover outpatient services, the fastest growing part of many health systems, and describe operations research models for primary and specialty care services and how to plan for patient no-shows. Chapters 12–16 cover topics related to the broader integration of health services in the context of public health, including optimizing the location of emergency vehicles, planning for mass vaccination events, and the coordination among different parts of a health system. Chapters 17–18 address supply chain management within the hospitals, with a focus on pharmaceutical supply management, and the challenges of managing inventory for nursing units. Finally, Chaps. 19–20 provide examples of important and emerging research in the realm of humanitarian logistics.

Ann Arbor, MI, USA Brian T. Denton

# Acknowledgments

There are many people who deserve credit for their help in editing this book. First, I am indebted to several people who reviewed the chapters of this book including: Sarang Deo (Indian School of Business), Ayca Erdogan (Stanford University), Ken Klassen (Brock University), Scott Levin (Johns Hopkins), Edi Pinker (University of Rochester), Ed Pohl (University of Arkansas), Steve Roberts (North Carolina State University), Randy Robinson (Cornell University), and Cem Saydem (University of North Carolina at Charlotte). I would like to thank Fred Hillier for inviting me to serve as editor for this book and Springer Staff including Neil Levine and Matthew Amboy for their help in producing the final version of this book. I would also like to thank my colleagues at North Carolina State University, Julie Ivy, Steve Roberts, and Reha Uzsoy for many engaging conversations that helped form my opinions about how to organize the content of this book. Finally, and most importantly, I would like to thank the authors of the chapters for the contributions of their time and effort to make this book a success.

# Contents

1 **Improving Access to Healthcare: Models of Adaptive Behavior** ...... 1
Carri W. Chan and Linda V. Green

2 **Queueing Models for Healthcare Operations** ........................ 19
Diwakar Gupta

3 **Applications of Agent-Based Modeling and Simulation to Healthcare Operations Management** .................................. 45
Sean Barnes, Bruce Golden, and Stuart Price

4 **Optimization in Healthcare Delivery Modeling: Methods and Applications** .................................................... 75
Sakine Batun and Mehmet A. Begen

5 **Operating Room Planning and Scheduling** ............................ 121
Erik Demeulemeester, Jeroen Beliën, Brecht Cardoen, and Michael Samudra

6 **The Modeling, Analysis, and Management of Intensive Care Units** ........................................................... 153
Theologos Bountourelis, M. Yasin Ulukus, Jeffrey P. Kharoufeh, and Spencer G. Nabors

7 **Improving the Flow of Patients Through Healthcare Organizations** . 183
Steven M. Thompson, Robert Day, and Robert Garfinkel

8 **Capacity Allocation and Flexibility in Primary Care** ........................................................ 205
Hari Balasubramanian, Ana Muriel, Asli Ozen, Liang Wang, Xiaoling Gao, and Jan Hippchen

9 **Improving Scheduling and Flow in Complex Outpatient Clinics** ..... 229
Craig M. Froehle and Michael J. Magazine

10 **No-Show Modeling for Adult Ambulatory Clinics** ........ 251
Ayten Turkcan, Lynn Nuti, Po-Ching DeLaurentis, Zhiyi Tian, Joanne Daggy, Lingsong Zhang, Mark Lawley, and Laura Sands

11 **Simulation and Real-Time Optimised Relocation for Improving Ambulance Operations** ........ 289
Andrew James Mason

12 **Planning and Managing Mass Prophylaxis Clinic Operations** ........ 319
Rachel L. Abbey, Katherine A. Aaby, and Jeffrey W. Herrmann

13 **Emergency Departments: "Repairs While You Wait, No Appointment Necessary"** ........ 349
Kenneth N. McKay, Jennifer E. Engels, Sahil Jain, Lydia Chudleigh, Don Shilton, and Ashok Sharma

14 **Location Models in Healthcare** ........ 387
Bjorn P. Berg

15 **Models and Methods for Improving Patient Access** ........ 403
Jonathan Patrick and Anisa Aubin

16 **Coordinating Health Services: An Operations Management Perspective** ........ 421
Thomas R. Rohleder, David Cooke, Paul Rogers, and Jason Egginton

17 **Managing Supply Critical to Patient Care: An Introduction to Hospital Inventory Management for Pharmaceuticals** ........ 447
Anita R. Vila-Parrish and Julie Simmons Ivy

18 **The Challenges of Hospital Supply Chain Management, from Central Stores to Nursing Units** ........ 465
Sylvain Landry and Martin Beaulieu

19 **Overcoming the Challenges of the Last Mile: A Model of Riders for Health** ........ 483
Jessica H. McCoy

20 **Allocating Scarce Healthcare Resources in Developing Countries: A Case for Malaria Prevention** ........ 511
Jacqueline Griffin, Pinar Keskinocak, and Julie Swann

**Index** ........ 533

# Chapter 1
# Improving Access to Healthcare: Models of Adaptive Behavior

**Carri W. Chan and Linda V. Green**

## 1 Introduction

Demand for healthcare is increasing due to a growing and aging population, making access to care more difficult. Beyond anecdotal evidence, there is increasing empirical evidence of access problems, most notably through overcrowding in emergency departments (EDs) (Burt and Schappert 2004; Committee on the Future of Emergency Care in the United States 2007). While demand is increasing, the supply of hospital beds, physicians, nurses, and other health resources remains relatively stagnant or, worse, is potentially decreasing. It is already the case that the supply of nurses is insufficient to meet demands (Chagaturu and Vallabhaneni 2005), and there are predictions of severe physician shortages in the coming years (Cooper et al. 2002; Merritt et al. 2004; Salsberg and Forte 2002).

As a consequence of high demand and insufficient supply, many patients experience delays in receiving treatment. The overall median wait to see an ED physician increased from 22 min in 1997 to 30 min by 2004. Perhaps even more alarmingly, the median wait for patients diagnosed with acute myocardial infarction (AMI) (heart attacks) increased from 8 min in 1997 to 14 min in 2004 (Wilper et al. 2008). In one study of patients and their primary care physicians, 33% of patients cited inability to get an appointment soon as a significant obstacle to care (Strunk and Cunningham 2002). The average wait for a primary care appointment in the US in 2001 was over 3 weeks (Murray and Berwick 2003). Sixty percent of physicians reported being dissatisfied with delays (Poon et al. 2004).

C.W. Chan (✉)
Columbia Business School, 410 Uris Hall, 3022 Broadway, New York, NY 10027, USA
e-mail: cwchan@columbia.edu

L.V. Green
Columbia Business School, 423 Uris Hall, 3022 Broadway, New York, NY 10027, USA
e-mail: lvg1@columbia.edu

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_1, © Springer Science+Business Media New York 2013

Delays can result in adverse patient outcomes such as increased mortality rates and an overall reduction in quality of outcome (Rincon et al. 2010). For emergent patients, such as those suffering AMI, timely access to care is imperative as even delays on the order of minutes can increase mortality (Luca et al. 2004; Chan et al. 2008; Buist et al. 2002; Yankovic et al. 2010). Delays can also result in increased length of stay (LOS), resulting in patients consuming more resources and further intensifying the problem. For example, delays in transfers from the ED to the intensive care unit (ICU) have been shown to increase ICU and hospital LOS (Chalfin et al. 2007; Renaud et al. 2009; Rivers et al. 2001).

As in other service environments, access problems may be due to uncontrollable variability which can stem from arrival times of patients, differing treatment types and times, staffing shortages, demand surges due to an epidemic, etc. The ability to effectively react to and navigate through periods of high congestion is imperative to ensuring timely patient access to care. Operations research models and methods can be useful in doing just that.

There are a number of behavioral factors in the healthcare setting which exacerbate access problems. One such factor is planned variability in capacity due to physician preferences. For example, surgeons often have significant ability to influence their own operating schedules. Most surgeons prefer operating in the morning so they can see new patients in the afternoon. This often results in surgeries being scheduled within a tight time window without adequate attention to the variability of their durations. Not surprisingly, many surgeries get delayed and recovery rooms get congested causing cancelations of subsequent surgeries. Since inpatient beds are often reserved for surgical patients, these surgical delays can translate into ED congestion due to the inability to move ED patients into inpatient beds. In one noted hospital study, the level of ambulance diversions (ambulances turned away from the ED) was better correlated with the variability in the *scheduled* surgical load than with emergency admissions (McManus et al. 2003). While some variability in the surgical schedule is certainly unavoidable, there is potential to utilize better scheduling of elective admissions to smooth load variability (Litvak et al. 2005). For instance, using stochastic linear programming, Denton et al. (2010) consider how to assign surgeries to various specialties and how to determine the number of operating rooms (ORs) to open given unavoidable uncertainty in the duration of various surgeries. Price et al. (2011) use integer programming methods to improve scheduling the OR and reduce boarding of patients in the postanesthesia recovery room due to ICU congestion. In fact, there has been considerable operations literature dealing with surgical scheduling (see Cardoen et al. (2009) and related references).

In this chapter, we will focus on a distinctive and prevalent characteristic of healthcare delivery systems-adaptive behavior. There has been growing evidence that patients and providers dynamically alter their behavior based on congestion and backlogs. These adaptive behaviors have been observed in both outpatient and inpatient settings. For instance, if patients have to wait a long time for an appointment with a physician, they may cancel at the last minute or just not show up (Galucci et al. 2005). When delays in the ED are long, patients are more likely to leave without being seen, even though they require care (Fernandes et al. 1997). Hospital EDs sometimes adapt to increasing backlogs by diverting ambulances away from the ED, effectively reducing patient arrivals and ED load (Kolker 2008). Though some of these behaviors may reduce the system workload, some adaptive behavior may actually worsen the situation. In one study of a hospital ED, nurses were found to be more likely to not show up for work when the anticipated patient load was higher, creating an even larger imbalance between supply (nurses) and demand (patients) (Green et al. 2011).

In the inpatient environment, providers are often faced with the difficult task of caring for more patients than their resources allow and, hence, adopt practices to attempt to mitigate these high-stress periods. For instance, physicians may discharge patients early from an ICU when it is full and space is needed for new patients (Kc and Terwiesch 2009). If there is no room in a hospital stroke unit at the time of a stroke patient's arrival, the patient may be placed in a less specialized unit which could result in a longer LOS and a poorer clinical outcome (Yankovic 2009). Indeed, patients are often assigned to less appropriate clinical units due to congestion in the desired unit.

Adaptive behavior can sometimes amplify system workload and/or variability creating additional problems; alternatively, adaptive behavior may alleviate congestion when it is most critical to do so. In any case, it is clear that adaptive behavior can significantly affect patient access, operational efficiency, and clinical outcomes. Yet the potential impact of adaptive behavior has not generally been explicitly considered in the operations research literature.

There is a need to develop models to account for adaptive behavior by patients and physicians. These enhanced models can provide vital insight which can lead to better policies and operational guidelines. The first step is to identify the adaptive phenomenon and quantify its impact on patient care. Such an understanding will provide a foundation to develop models and analyze operational policies which are better able to deal with adaptive behavior.

The remainder of this chapter is organized as follows. In Sect. 2, we discuss how to quantify the impact of adaptive behavior. Section 3 examines how to account for this adaptive behavior when making decisions. Section 4 discusses dynamic decision-making in the presence of this dynamic human behavior. Finally, Sect. 5 provides some closing remarks.

## 2 Quantifying the Impact of Adaptive Behavior

To develop models that allow us to ultimately identify policies and practices to better manage healthcare systems that are subject to adaptive behavior, we must first understand the nature and degree of adaptation. This requires empirical data to quantify the manner in which patient and physician behavior adapts to slight changes in a patient's health status or in the presented workload of the healthcare delivery system in question.

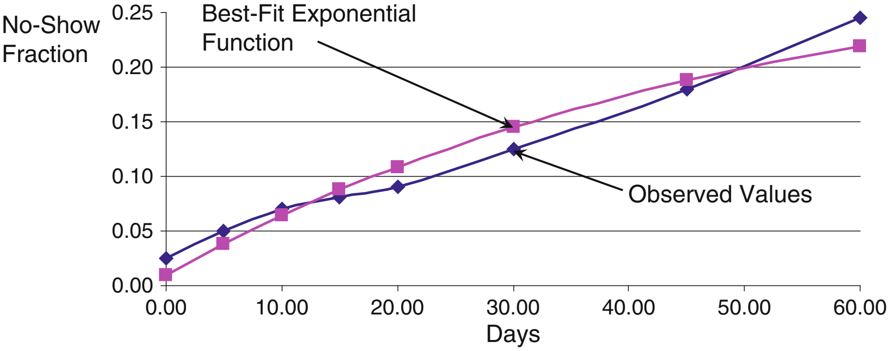


**Fig. 1.1** Observed no-show fraction values and the best-fit exponential functions for Columbia MRI data as reported in Green and Savin (2008) and as reported in Galucci et al. (2005) (Reprinted by permission, L.V. Green and S. Savin, Reducing delays for medical appointments: a queueing approach, Operations Research, volume 56, issue 6 (November/December 2008). Copyright 2008, the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 300, Hanover, MD 21076, USA)

## *2.1 Empirical Evidence: Adaptive Behavior of Patients*

There has been growing empirical evidence of adaptive behavior in a number of settings where patients react to delays. Using patient data, one can measure these effects via statistical analysis such as linear regression.

There are a growing number of healthcare practices and outpatient facilities that operate on an appointment basis. One of the difficulties faced by these facilities is patients who make last-minute cancelations or fail to arrive to their scheduled appointments. These patients are classified as “no-shows.” No-shows often waste already limited physician availability since it is usually impossible to fill a last-minute cancelation with another patient. This can result in significant monetary losses (up to 14% of annual revenues) for the clinic (Moore et al. 2001) (See Chap. 10 for more discussion on no-shows).

Empirical evidence has shown that the rate of no-shows increases with the increase in appointment backlog, i.e., the longer a patient has to wait until their appointment, the more likely he is to fail to show up. This phenomenon was observed at a mental health clinic, an MRI facility, and a family practice clinic (Galucci et al. 2005; Green and Savin 2008; Liu et al. 2010). In Green and Savin (2008), data on the connection between the appointment backlog and the likelihood of a patient no-show from both a mental health clinic and an imaging facility were fit to an exponential function as depicted in Fig. 1.1. The percentage of no-shows is monotonically increasing in backlog in these two independent data sets, though the rates are quite different, as would be expected with such different patient characteristics across the two facilities.

In another setting, there has been growing evidence that increased crowding in the ED has resulted in an increase in patients who leave the ED without being seen

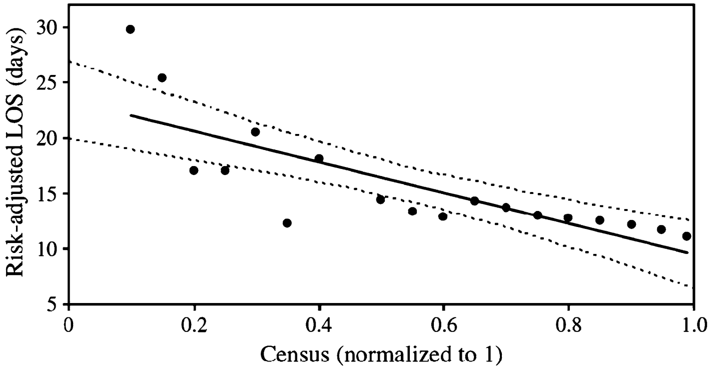


**Fig. 1.2** Length of stay (LOS) as a function of census. Note: census is defined as the number of patients in the cardiac unit at the time a patient is admitted. Length of stay is the total number of days a patient spends at the hospital. *Dashed lines* represent 95% confidence intervals (Reprinted by permission, D. Kc and C. Terweisch, Impact of workload on service time and patient safety: An econometric analysis of hospital operations, Management Science, volume 55, issue 9 (July 2009). Copyright (2009), the Institute for Operations Research and the Management Sciences (INFORMS), 7240 Parkway Drive, Suite 300, Hanover, MD 21076, USA)

(Derlet and Richards 2000; Green et al. 2002; Fernandes et al. 1997). This often means that patients who require care do not have the access they need (Baker et al. 1991).

## 2.2 Empirical Evidence: Adaptive Behavior of Physicians

Not only do patients react to the supply and demand mismatch, but physicians do as well. Ideally, physicians should make decisions for the provision of care based entirely upon medical and physiologic factors. Unfortunately, this is not always possible due to resource constraints. With the increase in sophistication of electronic medical record (EMR) systems and, subsequently, the increase in available patient data, econometric tools can be used to estimate how capacity constraints influence physician behavior. The general methodology begins with fitting a regression model to the available data and examining the relationships between key variables.

Kc and Terwiesch (2012, 2009) examine the relationship between occupancy levels and patient care. Using data from an ICU unit for cardiothoracic surgery patients, they demonstrate that, after controlling for patient severity, a patient's LOS decreases as the unit occupancy level increases. This is illustrated in Fig. 1.2. This supports anecdotal reports that patients are sometimes discharged prematurely in order to accommodate new, more critical patients. We refer to such a discharge as a *demand-driven* discharge.

More generally, there is evidence that patients' LOS in ICUs is influenced by bed availability. There are a number of important research questions surrounding such adaptive behavior:

1. Under what circumstances do physicians adapt LOS based on occupancy levels?
2. Does reduction in LOS adversely affect patient outcomes?
3. What policies might be employed to guide such adaptability so that operational and clinical performance is improved?

In this section, we will focus on the first two questions; the third will be addressed in Sect. 4. In addition to the effect of high occupancy levels on patient LOS, it may also affect patient readmission likelihood due to the early discharge of patients. However, there is an inherent endogeneity bias, since more severe patients are likely to have longer LOS in the ICU and have higher readmission risks, which could lead to a positive bias in estimating the effect of LOS on readmissions. The exogenous factors affecting LOS variables that affect the time spent in the ICU, but otherwise do not directly affect patient outcomes, constitute potential instrumental variables (IVs) to mitigate the endogeneity bias. In particular, an indicator variable which specifies whether or not the ICU is busy (i.e., at high occupancy levels) upon discharge becomes a valid IV (see Wooldridge (2002) for details on this methodology). Because operational factors are unlikely to be correlated with patient medical factors, such as severity, which may affect patient outcomes, they can often be used as instrumental variables to generate unbiased estimates of these outcomes.

In a study of cardiac surgical patients in a single hospital, a 10% increase in occupancy level corresponded to a 20% decrease in ICU LOS a reduction of nearly 2.5 days (Kc and Terwiesch 2009). This shortened LOS corresponded to increases in the likelihood of readmission. Specifically, being discharged 1 day earlier than one's expected LOS translated to an increase of 60% in the odds of being readmitted to the ICU (Kc and Terwiesch 2012). This modification of patient LOS due to congestion may initially free capacity in the ICU, but it can also negatively impact patient outcomes in the long run.

## 2.3 *Quantifying Effects via Modeling*

Empirical models are able to quantify adaptive behavior under the conditions of the particular patient setting in question. Randomized trials are generally not possible in hospital settings where it could result in patients being denied needed treatment. Hence, it can be difficult to empirically measure a variety of scenarios. By building and analyzing models which incorporate this behavior, the impact of adaptive behavior can be estimated for a wider range of scenarios.

Using the ICU described as an example, the first step is to build a stochastic model of an ICU which incorporates the fact that patients may be demand-driven

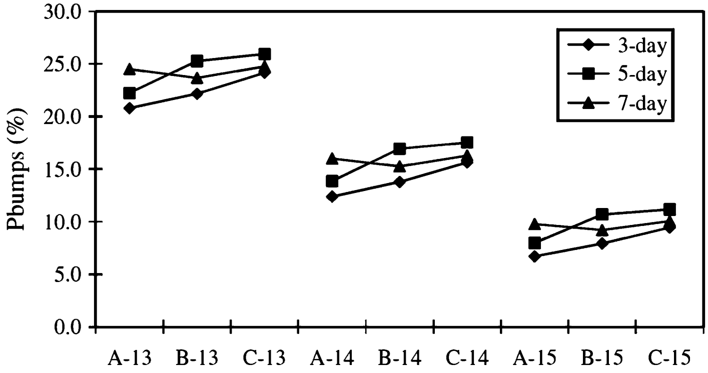


**Fig. 1.3** Average probability of being demand-driven discharged for (A) 70% (B) 50% and (C) 30% scheduled patients and an ICU of size 13, 14, and 15 beds as reported in Dobson et al. (2009) (Reprinted by permission, G. Dobson, H.-H. Lee, E. Pinker. A model of ICU bumping. Operations Research, volume 58, issue 6 (November/December 2010). Copyright (2010), the Institute for Operations Research and the Management Sciences (INFORMS), 7240 Parkway Drive, Suite 300, Hanover, MD 21076, USA)

discharged. Such a model can be used to consider how changes in arrival patterns, ICU capacity, and surgical schedules can affect the likelihood of being discharged early.

In Dobson et al. (2009), it is assumed that patients are either scheduled or unscheduled. The state of the system is given by the remaining LOS of the patients who occupy the ICU. If a new patient arrives and there is no space available, a current patient is demand-driven discharged to accommodate the new patient. Using an aggregation–disaggregation technique to reduce computational complexity, Dobson et al. calculate the desired performance metrics such as the probability of being *bumped* and the expected number of days remaining when a patient is bumped.

Figure 1.3 plots the probability of being discharged early for a number of different scenarios with increasing ratio of unscheduled to scheduled cases and increasing size of ICU. As expected, the probability of being bumped is lower when there are more beds. Additionally, the likelihood of being demand-driven discharged increases with the percentage of unscheduled patients who introduce higher variability. One can also vary the number of days in a week that scheduled patients can arrive (3, 5, or 7). Interestingly, when patients are scheduled on 3-day plans, the probability of a demand-driven discharge is the lowest. One possible explanation for this is that patient arrivals are more spread apart, allowing for more time to recover from busy periods. Such analysis is useful for understanding how various parameters and schedules affect the undesirable, yet unavoidable, phenomenon of demand-driven discharges.

## 3 Incorporating Adaptive Behavior into Decision-Making

Ignoring the impact of adaptive behavior may result in suboptimal operational decisions, which can further amplify the supply and demand mismatch rather than help alleviate it. We illustrate how to incorporate the impact of adaptive behavior in both the outpatient and inpatient setting.

### *3.1 Accounting for No-Shows When Determining Patient Panel Size*

As mentioned previously, no-shows are prevalent in many outpatient settings, particularly when the system is congested. Ignoring this phenomenon can hurt both providers and patients. One example of this is in determining how large a patient panel size a group of physicians can handle. Primary care practices and many specialty care practices, such as cardiology, have a "patient panel" a set of patients who receive their care from the practice on some regular basis. So in these practices, patient panel size is the primary lever to align demand and supply in order to offer timely access.

To identify a panel size that will result in short waits for appointments with high probability, it is necessary to explicitly consider the nature and impact of cancelations. Although some patients cancel their appointments far enough in advance of their scheduled time to allow for a new appointment request to be substituted, many practices experience a high level of patients who cancel too late for this to happen or who simply do not show up at the scheduled time. This results in the paradoxical situation where the physician may be idle for some significant amount of time during the day while patient backlogs for appointments are long. In addition, although some patients fail to appear at the appointed time because the original reason for the visit no longer exists, other no-shows are due to personal or work-related problems or to the patient's decision to seek treatment elsewhere rather than wait. In the latter situations, many no-shows schedule a new appointment with their original physician. This is true even when they have sought treatment elsewhere because it is common practice for clinics and emergency rooms to advise the patient to see their own physician as well.

Green and Savin (2008) model a single-physician practice via a modified M/D/1/K queue where patients arrive according to a Poisson process, service times are deterministic, and there is a finite appointment backlog limit $K$ such that any patients who arrive when the queue length is $K$ are "lost" in the sense that they are not given an appointment and so potentially seek treatment elsewhere. For more information on the standard M/D/1/K queue, we refer the reader to examine a book on queueing, such as the one by Kleinrock (1975). The modified M/D/1/K model approximates the no-show process by assuming that a customer who is scheduled to begin service has a state-dependent probability of being a no-show, resulting in

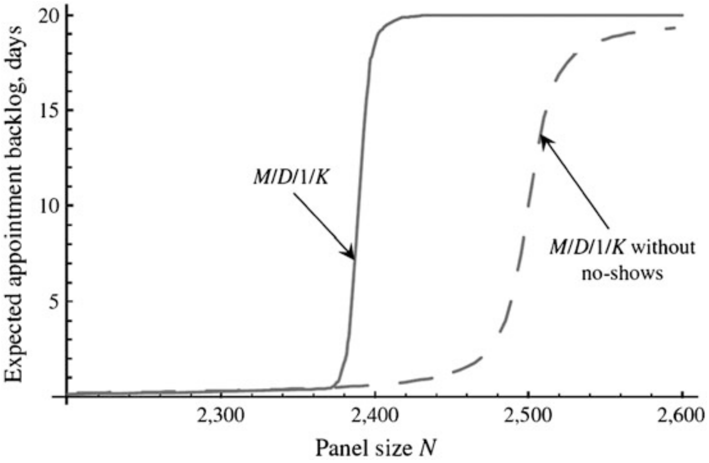


**Fig. 1.4** Expected appointment backlog as a function of the patient panel size for the M/D/1/K model with and without no-shows (using a no-show model based on an MRI facility data, assuming 20 slots per day, $K = 400$ appointment slots and a probability of rescheduling equal to 1) as reported in Green and Savin (2008) (Reprinted by permission, L. V. and S. Savin, Reducing delays for medical appointments: a queueing approach, Operations Research, volume 56, issue 6 (November/December 2008). Copyright 2008, the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 300, Hanover, MD 21076, USA)

an idle period for the server and, with a fixed probability, the customer rejoining the queue. The likelihood of no-show is nondecreasing in the number of patients who are still in the backlog upon the appointment (i.e., service) time of the patient in question. Such an approximation is able to capture wasted capacity by patient no-shows as well as the increased likelihood of such events when the system is more congested. Additionally, it allows for analytical tractability of the steady-state behavior of the system.

Figure 1.4 compares the expected appointment backlog of the M/D/1/K queue with and without no-shows. Using a no-show model calibrated from data of a MRI facility, one can see that the impact of patient no-shows is very significant (Green and Savin 2008). Since no-shows result in wasted appointment slots and rescheduled appointments, they result in longer appointment backlogs and hence more no-shows. Thus, there is an adverse feedback cycle and the backlog grows much more rapidly than in a model without no-shows. Ignoring no-shows in a model of a clinic may result in a physician electing to maintain a panel size that is too large to provide timely access to care for his/her patients.

An increasingly important performance metric for access in this setting is the probability of being able to get a same-day appointment. In fact, 33% of patients reported that the "inability to get an appointment soon" inhibited access to care (Strunk and Cunningham 2002). Figure 1.5 compares the probability of getting a same-day appointment with and without no-shows. If no-shows are not considered, the figure suggests that a panel size of 2,400 would provide timely

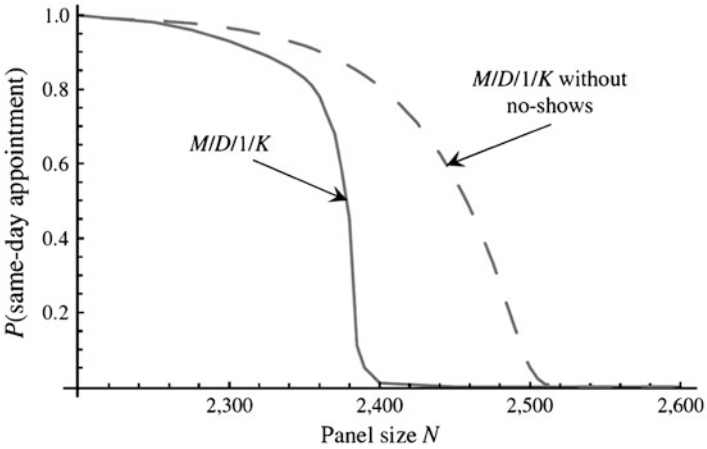


**Fig. 1.5** Probability of getting a same-day appointment as a function of the patient panel size for the M/D/1/K model with and without no-shows (using a no-show model based on an MRI facility data, assuming 20 slots per day, $K = 400$ appointment slots and a probability of rescheduling equal to 1) as reported in Green and Savin (2008) (Reprinted by permission, L. V. and S. Savin, Reducing delays for medical appointments: a queueing approach, Operations Research, volume 56, issue 6 (November/December 2008). Copyright 2008, the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 300, Hanover, MD 21076, USA)

access since patients will be able to get a same-day appointment 80% of the time. (This performance level would be consistent with data that suggests about 20% of appointments are for follow-up care and are scheduled weeks in advance.) However, in actuality, this probability is likely to be close to 0, due to the no-show phenomenon. Such an analysis highlights the importance of accounting for the adaptive behavior of patients when making operational decisions.

## 4 Dynamic Policies Which Account for Adaptive Behavior

Along with macro-level decisions such as patient panel sizing, staffing levels, and the number of beds, OR models can provide insights on how to dynamically account for adaptive behavior and unavoidable periods where demand exceeds supply.

### *4.1 Accounting for No-Shows When Scheduling Patient Appointments*

An important aspect of outpatient clinic management is scheduling patients as they call for appointments. As with panel size planning, patient no-shows are an important factor in crafting schedules so that the number of idle appointment slots is minimized. Liu et al. (2010) analyze a dynamic scheduling model which captures this no-show phenomenon. Each day, the appointment scheduler must determine which day to assign to each patient who calls for an appointment. The longer a patient waits for an appointment, the more likely she is to cancel or be a no-show. On any given day, a scheduled patient can show up or not show up for her appointment that day or cancel an appointment which may be on a future date.

This scheduling problem can be formalized as a dynamic optimization problem in which the objective is to maximize the number of patients cared for each day or, equivalently, minimize the number of idle slots without incurring high overtime costs. There is an inherent trade-off between providing timely access for patients and potentially incurring high overtime costs in order to ensure this versus allocating a large amount of initial capacity which may end up being wasted if there is not enough demand in a particular day. In principal, the optimal scheduling policy can be determined using dynamic programming and numerical methods. However, dynamic programming often suffers from the *curse of dimensionality*, and solving such a recursion for problem sizes of interest is practically infeasible. An alternative course of action is to develop heuristic algorithms.

Two simple heuristics are to optimize the scheduling policy assuming appointments depending on how quickly a patient must be seen (Liu et al. 2010). The first heuristic, referred to as *open access*, requires appointments to be provided on the current day. Hence, patients are guaranteed same-day appointments, even if this requires the physician to spend significant overtime to treat all patients beyond the initial allocated daily capacity. Another heuristic, the *two-day policy*, requires an appointment to be provided on the current or following day. In doing so, it tries to reduce overtime costs at the expense of immediate access.

These heuristics can serve as the basis for additional heuristics using policy improvement. The policy improvement heuristic selects the best scheduling decision in the current state under the assumption the suboptimal base policy is used for all subsequent decisions. Hence, it is a one-step policy improvement over the base heuristic (see Bertsekas (2005) for more details on this methodology). Finally, these heuristics are compared to the following benchmarks: a *threshold heuristic* where patients are scheduled on the earliest day with fewer than $M$ patients scheduled, a *load balancing heuristic* where patients are scheduled on the day with the fewest appointments, and a *random heuristic* where patients are scheduled on a random day.

Using data calibrated from empirical data of a family medicine clinic, Liu et al. (2010) compares the performance of the proposed algorithms via simulation. Table 1.1 summarizes the relative costs of the various heuristic policies in comparison to the open access scheduling policy for various daily capacities ($M$) and cost of scheduling one patient ($h$). Note that the number of patients scheduled in a day, $z$, can be greater or less than $M$. If there are more patients than appointment slots ($z > M$), all of these patients will be treated and overtime cost is incurred. This is in contrast to the $K$ in the M/D/1/K model of Sect. 3 as in that case, overtime was not allowed. Smaller $M$ suggests the clinic is more overloaded. The simulations

**Table 1.1** Simulation study results: percentage improvement of total reward (number of patients served less daily fixed cost and scheduling cost) compared to open access scheduling for daily capacity $M$ and cost $h$ for scheduling a patient as reported in Liu et al. (2010)

| | | PI 2-day | 2-day | PI open access | Threshold | Load balancing | Random |
|---|---|---|---|---|---|---|---|
| $M = 55$ | $h = 0$ | 2.11 | 0.78 | 2.18 | 2.11 | −6.30 | −3.28 |
| | $h = 0.2$ | 4.10 | 3.23 | 3.08 | 3.25 | −5.48 | −1.53 |
| | $h = 0.5$ | 12.74 | 12.14 | 3.72 | 4.39 | −5.48 | 2.68 |
| $M = 50$ | $h = 0$ | 6.77 | 2.75 | 5.42 | 6.45 | −2.22 | −1.20 |
| | $h = 0.2$ | 8.28 | 5.48 | 6.96 | 8.21 | −1.09 | 0.50 |
| | $h = 0.5$ | 18.56 | 15.29 | 9.25 | 12.11 | 0.72 | 5.31 |
| $M = 45$ | $h = 0$ | 10.63 | 6.23 | 9.25 | 5.24 | 4.11 | 1.81 |
| | $h = 0.2$ | 13.35 | 9.13 | 11.53 | 6.28 | 4.91 | 3.28 |
| | $h = 0.5$ | 25.01 | 20.32 | 21.78 | 10.40 | 8.10 | 9.16 |
| $M = 40$ | $h = 0$ | 9.84 | 9.12 | 10.21 | 2.79 | 2.99 | 4.23 |
| | $h = 0.2$ | 13.03 | 12.48 | 13.69 | 3.57 | 3.83 | 6.05 |
| | $h = 0.5$ | 27.41 | 26.82 | 28.13 | 6.79 | 7.32 | 13.58 |

Adapted with permission from N. Liu, S. Ziya, V. G. Kulkarni. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations, Manufacturing & Service Operations Research volume 12, issue 2 (October 2010). Copyright (2010), the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 300, Hanover, MD 21076, USA

suggest that the threshold heuristic, 2-day policy, and policy improvement heuristics based on open access and the 2-day policy generate more revenue compared to open access. Interestingly, even the Random heuristic sometimes outperforms open access. In an underloaded system, open access would be optimal. Under open access, all patients are scheduled on the current day, which minimizes their likelihood of being a no-show. However, as the practice becomes more heavily loaded, this will result in frequent overload, requiring physicians to work overtime, thus incurring high costs. Hence, open access is not the best scheduling policy to use in general.

### *4.2 Improving Demand-Driven Discharge Decisions*

Inpatient care is another setting in which dynamic policies can be useful to provide effective treatment. As an example, physicians are often faced with the difficult task of determining whether and when to discharge an ICU patient early due to limited bed availability. As seen in Sect. 2.2, there is empirical evidence that physicians adaptively alter patient discharge times based on congestion levels.

Various factors can affect how often demand-driven discharges must occur (Dobson et al. 2009). A natural question is how one should determine which patient to discharge. Section 2.3 assumed that patients were discharged in order of shortest remaining LOS. However, given the natural variability in patient stays, this quantity is not always known. Additionally, it ignores the potential impact of readmissions on ICU congestion. In Chan et al. (2011), a dynamic optimization model is developed to help guide such decisions.

The model in Chan et al. (2011) assumes that whenever a new patient arrives and there are no available beds, a physician must decide which patient to discharge in order to accommodate the new, higher acuity patient. Each patient is identified by type, which specifies the expected initial ICU length of stay, the likelihood of readmission upon a demand-driven discharge, and the expected ICU LOS upon readmission. Any cost function which accounts for a patient's disservice due to a demand-driven discharge can be incorporated. A cost function which is estimable from currently available data is given by the *readmission load*, i.e., expected ICU treatment time required by the demand-driven discharged patient following the initial discharge.

In principle the optimal policy can be computed numerically via dynamic programming. Unfortunately, the size of the state space makes it practically infeasible to solve. Utilizing properties of the optimal value function, the authors show that the performance of a greedy heuristic, which discharges the patient with the lowest readmission load, is a $(\hat{\rho}+1)$-approximation for the optimal policy, where $\hat{\rho}$ is a measure of utility (Chan et al. 2011). Such a bound is useful to quantify the worse-case performance of such a greedy policy.

Using patient data from seven different hospitals in a single hospital network, Chan et al. simulate the performance of the greedy discharge policy relative to several relevant benchmarks. In the medical community, the decision of which patient to discharge is made by assessing which patient is the "least critical" (see, for instance, Swenson (1992)) which can be somewhat subjective and is generally not based upon quantitative measures. Each of the discharge policies studied below can be interpreted as a measure of criticality:

- *Probability of readmission index:* Discharge the patient with the smallest probability of readmission. Readmitted patients tend to be more critical (see Durbin and Kopel (1993)), so that the rationale here is that a lower likelihood of readmission translates to lower patient criticality.
- *Length-of-stay (LOS) index:* Discharge the patient with the smallest remaining service time. This policy thus equates criticality with the nominal LOS of a patient. This policy is analyzed in Dobson et al. (2009) albeit for a model that is agnostic to readmission loads.
- *The greedy index:* This is the proposed heuristic from Chan et al. (2011) which prioritizes patients in increasing order of readmission load.

In addition to the preceding index rules, one can also consider a *random policy*.

Figure 1.6 compares the readmission load of the greedy heuristic compared to the other benchmark policies. The savings relative to the next best policy corresponds to 23.7 h over 1 week at a net patient arrival rate of $\lambda = 0.021$ (or 1 ICU bed out of 10 for 1 day per week). Figure 1.7 shows the number of deaths per week for the same discharge policies. One can see that the number of deaths is practically

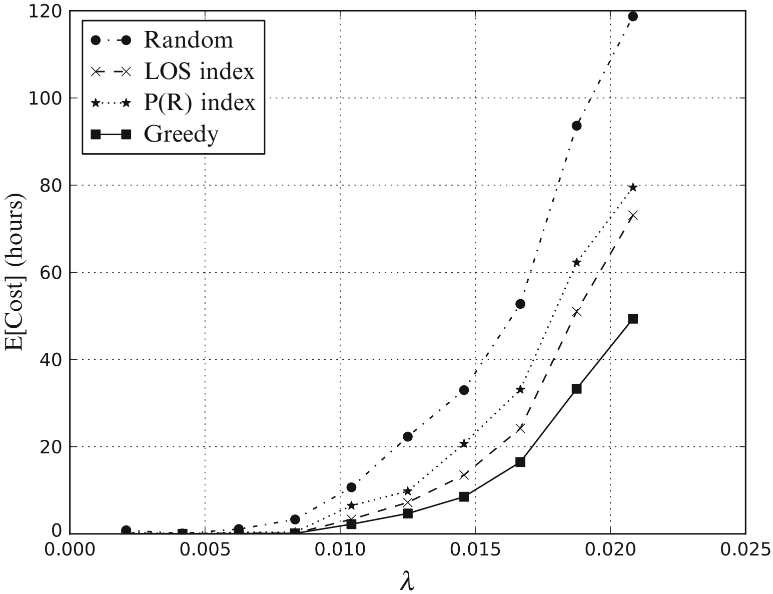


**Fig. 1.6** Performance of greedy policy compared to benchmarks for various arrival rates and distribution across patient types according to the proportions seen in the empirical data as reported in Chan et al. (2011)

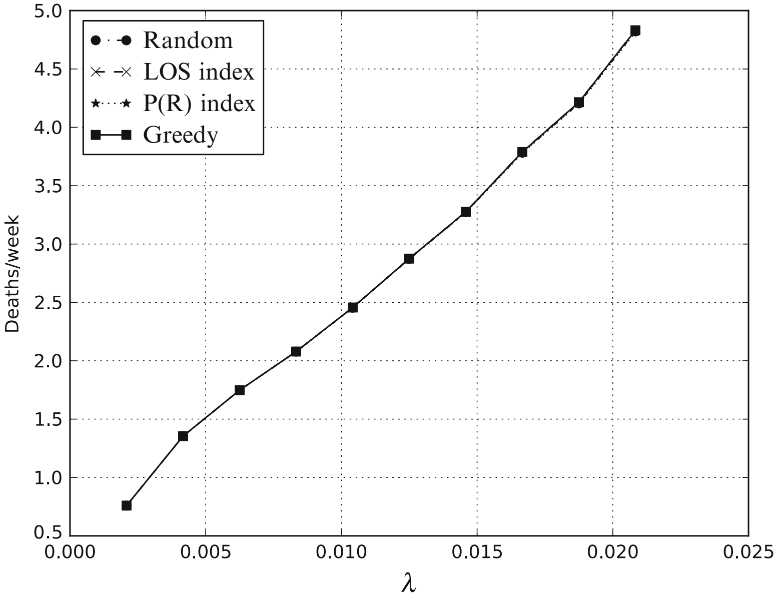


**Fig. 1.7** Number of deaths for greedy policy compared to benchmarks for various arrival rates and distribution across patient types according to the proportions seen in the empirical data as reported in Chan et al. (2011)

identical for all policies, while the readmission load is very different. Hence, without sacrificing patient quality, in terms of mortality, the greedy heuristic which incorporates readmission risks can significantly reduce the patient load on the ICU and subsequently increase the number of patients who receive critical care.

## 5 Conclusions and Future Research

As demonstrated in this chapter, behavior can have a significant impact on the efficiency and effectiveness of healthcare delivery and so must be considered in making both design and operational decisions. Operations research studies and methodologies are needed to both understand the nature of adaptive behaviors and identify policies that incorporate such behavior in order to improve access to care. In addition to the examples presented here, there are several other important areas of healthcare delivery where adaptive behavior is prevalent, providing potential opportunities for future research.

One consequence of adaptive behavior is that the true service requirements and arrival rates of patients may be censored. These potentially misleading measurements of the system load make it difficult to assess the actual required capacity requirements. Due to the chronic mismatch of supply and demand, much of the observed behavior of healthcare systems do not accurately reflect the true dynamics. Hence, there is a need to analyze the impact of adaptive behaviors on patient demands and treatment times when estimating required capacity for many healthcare resources (e.g., ICU beds, obstetric beds, surgical suites, primary care physicians, nurses). For instance, ignoring endogenous nurse absenteeism can result in understaffing (Green et al. 2011).

Adaptive behavior can also induce downstream effects which can create very complex decision-making environments. For instance, when making demand-driven discharges, one must account for the immediate impact on the discharged patient as well as the propagation effects due to his potential readmission. This could be expanded to consider how transferring patients to a less appropriate unit (i.e., an intermediary care unit rather than an intensive care unit) impacts patient LOS and outcomes.

There is significant heterogeneity in patient types. In Sect. 3, all patients were assumed to have identical characteristics. However, patients are likely to have different service requirements and/or no-show rates. Similarly, one could consider appointment scheduling for two patient types: routine versus urgent patients as in Dobson et al. (2011). This work does not account for the no-show phenomenon. An interesting research direction would be to consider how to combine heterogenous patients with the no-show phenomenon.

There has been a growing interest in the operations research community to understand how human behavior impacts operations. Accounting for "behavioral operations" when designing system can improve performance (Boudreau et al. 2003). A recent survey of this area emphasizes the need for experimental data to identify the impact of human behavior (Bendoly et al. 2006). While randomized controlled experiments are often not possible in healthcare settings, adaptive behavior can be measured from retrospective data as described in this chapter.

With an increase in the sophistication of electronic medical record systems, more patient data is becoming available. Combining operations research methodologies with real patient data will help facilitate the identification and modeling of adaptive behaviors in various healthcare settings. Models that use real data to demonstrate the impact of adaptive behavior and identify policies and practices that mitigate the potentially negative consequences of these behaviors can be extremely useful in improving access to healthcare. Moreover, it will provide further evidence and credibility to physicians who may be considering making policy and practice changes. There is a great deal of potential to significantly improve the operational performance of healthcare systems and enable better access to patient care by accounting for adaptive behavior when modeling, analyzing, and developing policies for such systems.

## References

Baker DW, Stevens CD, Brook RH (1991) Patients who leave a public hospital emergency department without being seen by a physician. Causes and consequences. JAMA 266: 1085–1090

Bendoly E, Donohue K, Schultz KL (2006) Behavior in operations management: Assessing recent findings and revisiting old assumptions. J Oper Manag 24(6):737–752

Bertsekas DP (2005) Dynamic programming and optimal control. Athena Scientific, Nashua, NH

Boudreau J, Hopp W, McClain JO, Thomas LJ (2003) On the interface between operations and human resource management. MSOM 5:179–202

Buist MD, Moore GE, Bernard SA, Waxman BP, Anderson JN, Nguyen TV (2002) Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: Preliminary study. Br Med J 324:387–390

Burt CW, Schappert SM (2004) Ambulatory care visits to physician offices, hospital outpatient departments, and emergency departments: United States, 1999–2000. Vital Health Stat 13(157):1–70

Cardoen B, Demeulemeester E, Belien J (2009) Operating room planning and scheduling: A literature review. Eur J Oper Res 201:921–932

Chagaturu S, Vallabhaneni S (2005) Aiding and abetting - nursing crises at home and abroad. New Engl J Med 353(17):1761–1763

Chalfin DB, Trzeciak S, Likourezos A, Baumann BM, Dellinger RP (2007) Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. Crit Care Med 35:1477–1483

Chan CW, Farias VF, Bambos N, Escobar G (2011) Maximizing throughput of hospital intensive care units with patient readmissions. Working Paper, Columbia Business School

Chan PS, Krumholz HM, Nichol G, Nallamothu BK (2008) Delayed time to defibrillation after in-hospital cardiac arrest. New Engl J Med 358(1):9–17

Committee on the Future of Emergency Care in the United States (2007) Emergency medical services at the crossroads. The National Academies Press, Washington, DC

Cooper R, Getzen T, McKee H, Laud P (2002) Economic and demographic trends signal an impending physician shortage. Health Aff 21(1):140–154

Denton BT, Miller AJ, Balasubramanian HJ, Huschka TR (2010) Optimal allocation of surgery blocks to operating rooms under uncertainty. Oper Res 58(4):802–816

Derlet R, Richards J (2000) Overcrowding in the nations emergency departments: Complex causes and disturbing effects. Ann Emerg Med 35:63–68

De Luca G, Suryapranata H, Ottervanger JP, Antman EM (2004) Time delay to treatment and mortality in primary angioplasty for acute myocardial infarction: Every minute of delay counts. Circulation 109(10):1223–1225

Dobson G, Lee H-H, Pinker E (2010) A model of ICU bumping. Oper Res 58:1564–1576

Dobson G, Hasija S, Pinker EJ (2011) Reserving capacity for urgent patients in primary care. Prod Oper Manag 20(3):456–473

Durbin CG, Kopel RF (1993) A case-control study of patients readmitted to the intensive care unit. Crit Care Med 21:1547–1553

Fernandes CM, Price A, Christenson JM (1997) Does reduced length of stay decrease the number of emergency department patients who leave without seeing a physician? J Emerg Med 15: 397–399

Galucci G, Swartz W, Hackerman F (2005) Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. Psychiatr Serv 56:344–346

Green LV, Savin S (2008) Reducing delays for medical appointments: A queueing approach. Oper Res 56:1526–1538

Green LV, Savin S, Savva N (2011) 'Nursevendor problem': Personnel staffing in the presence of endogenous Absenteeism. Working paper, Columbia Business School

Green RA, Wyer PC, Giglio J (2002) ED walkout rate correlated with ED length of stay but not with ED volume or hospital census [abstract]. Acad Emerg Med 9:514

Kc D, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. Manag Sci 55:1486–1498

Kc D, Terwiesch C (2012) An Econometric Analysis of Patient Flows in the Cardiac ICU. MSOM 14(1):50–65

Kleinrock L (1975) Queueing systems, volume I: Theory. Wiley Interscience, New York

Kolker A (2008) Process modeling of emergency department patient flow: Effect of patient length of stay on ED diversion. J Med Syst 32:389–401

Litvak E, Buerhaus PI, Davidoff F, Long MC, McManus ML, Berwick DM (2005) Managing unnecessary variability in patient demand to reduce nursing stress and improve patient safety. Joint Comm J Qual Patient Saf 31:330–338

Liu N, Ziya S, Kulkarni V (2010) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. MSOM 12:347–364

McManus ML, Long MC, Cooper A, Mandell J, Berwick DM, Pagano M, Litvak E (2003) Variability in surgical caseload and access to intensive care services. Anesthesiology 98: 1491–1496

Merritt J, Hawkins J, Miller PB (2004) Will the last physician in America please turn off the lights? A look at America's looming doctor shortage. Technical report, Practice Support Resources, Inc., Irving, TX

Moore CG, Wilson-Witherspoon P, Probst JC (2001) Time and money: Effects of no-shows at a family practice residency clinic. Fam Med 33:522–527

Murray M, Berwick DM (2003) Advance access: Reducing waiting and delays in primary care. J Am Med Assoc 289:1035–1039

Poon EG, Gandhi TK, Sequist TD, Murff HJ, Karson AS, Bates DW (2004) 'I wish i had seen this test result earlier!': Dissatisfaction With test result management systems in primary care. Arch Intern Med 164:2223–2228

Price C, Golden B, Harrington M, Konewko R, Wasil E, Herring W (2011) Reducing boarding in a post-anesthesia care unit. Production and Operations Management 20(3):431–441

Renaud B, Santin A, Coma E, Camus N, Van Pelt D, Hayon J, Gurgui M, Roupie E, Hervé J, Fine MJ, Brun-Buisson C, Labarère J (2009) Association between timing of intensive care unit admission and outcomes for emergency department patients with community-acquired pneumonia. Crit Care Med 37(11):2867–2874

Rincon F, Mayer SA, Rivolta J, Stillman J, Boden-Albala B, Elkind MSV, Marshall R, Chong JY (2010) Impact of delayed transfer of critically ill stroke patients from the emergency department to the neuro-ICU. Neurocrit Care 13:75–81

Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, Peterson E, Tomlanovich M (2001) Early goal-directed therapy in the treatment of severe sepsis and septic shock. New Engl J Med 345(19):1368–1377

Salsberg ES, Forte GJ (2002) Trends in the physician workforce, 1980–2000. Health Aff 21: 165–173

Strunk BC, Cunningham PJ (2002) Treading water: Americans access to needed medical care, Report, 1997–2001. Center for Studying Health System Change, Washington, DC

Swenson MD (1992) Scarcity in the intensive care unit: Principles of justice for rationing ICU beds. Am J Med 92:552–555

Wilper AP, Woolhandler S, Lasser KE, McCormick D, Cutrona SL, Bor DH, Himmelstein DU (2008) Waits to see an emergency department physician: U.S. trends and predictors, 1997–2004. Health Aff 27:w84–w95

Wooldridge JM (2002) Econometric analysis of cross section and panel data. MIT, Cambridge

Yankovic N (2009) Models for assessing the impact of resource allocation in hospitals. PhD thesis, Columbia Business School

Yankovic N, Glied S, Green LV, Grams M (2010) The Impact of Ambulance Diversion on Heart Attack Deaths. Inquiry. 47(1):81–91

# Chapter 2
# Queueing Models for Healthcare Operations

**Diwakar Gupta**

## 1 Introduction

Queues form when entities that request service, typically referred to as *customers*, arrive at a *service facility* and cannot be served immediately upon arrival. In healthcare delivery systems, patients are typically the customers and either outpatient clinics or diagnostic imaging centers or hospitals are the service facilities. There are also many atypical examples of customers and service facilities, as shown below.

| *Customers* | *Service facility* |
|---|---|
| Diagnostic images | Radiology department |
| Doctors' notes | Coding department (for billing purposes) |
| Prescriptions | Mail-order pharmacy |
| Transplant candidates | Organ procurement organization |

A service facility may consist of one or more service stations where customers are served. Further, each service station may consist of one or more servers. For example, the processing of diagnostic images may require two types of servers—radiologists who read the images and transcribers who input radiologists' dictated notes into patient charts. Servers are often grouped by their expertise to form service stations, although other configurations (e.g., multiple specialty teams of doctors and nurses) are also prevalent. A common feature of the vast majority of queueing models is that customers are discrete, and the number of customers waiting in the service facility is integer valued.

D. Gupta (✉)
University of Minnesota, 111 Church Street S. E., Minneapolis, MN 55455, USA
e-mail: guptad@me.umn.edu

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_2, © Springer Science+Business Media New York 2013

Queues are ubiquitous, particularly in healthcare delivery systems. At the same time, queues are undesirable because delay in receiving needed services can cause prolonged discomfort and economic loss when patients are unable to work and possible worsening of their medical conditions that can increase subsequent treatment costs and poor health outcomes. In extreme cases, long queues can delay diagnosis and/or treatment to the extent that death occurs while a patient waits. For example, there is a severe shortage of organs in the USA and many patients die while waiting for suitable organs for transplant.

Given the negative consequences of queues in healthcare delivery systems, the following questions naturally arise. Why do queues form? Why must customers wait to be served? Which features of system design affect queueing and by how much? What trade-offs must be considered by a service system architect when choosing system parameters? This chapter attempts to provide answers to questions such as these.

Although queues have existed as far back as historical records are available, mathematical study of queues, called queueing theory, has been around since the early 1900s. Works on the theory and applications of queueing systems have grown exponentially since the early 1950s. It is neither possible nor the intent to provide a summary of this vast body of literature in this chapter. For that, there are many excellent books, both at the introductory and advanced levels—see, for example, Bhat (2008), Cohen (1969), Cox and Smith (1961), Gross and Harris (1985), Morse (1958), Newell (1982), Takacs (1962) and Wolff (1989). A review of papers that attempts to tackle real queueing problems can be found in Worthington (2009). This chapter provides a review of a few basic queueing models and discusses their implications for healthcare operations management.

Borrowing terminology from the queueing literature, we shall henceforth use the terms *queueing system* and *service facility* interchangeably. A queueing system has the following elements:

1. Servicestations or workstations, their configuration, and routing protocols that determine flow of customers from one station to another.
2. Number of servers at each station.
3. Service protocol at each station—a commonly used protocol is first in first out (FIFO) because it is deemed to be fair (Larson 1987). However, it is not at all uncommon to give higher priority to certain types of customers (often referred to as *classes* in the queueing literature). For example, patients whose condition is deemed critical by medical professionals generally bypass queues.
4. Service time distribution by customer class, server, and station.
5. Arrival process—the distributions of inter-arrival times, number of arrivals at each arrival epoch, and arrival location.
6. Size of waiting room at each station. When waiting room is limited, either customers are turned away or congestion at a downstream station causes *blocking* at an upstream station.

7. Protocols governing server absences and distributions of server vacations. Vacation refers to a period of time when a server is not available, which could happen for a whole host of reasons including activities such as attending to other tasks and taking a break.

Suppose we observe arrivals and departures from a queueing system that starts empty. The system may consist of an arbitrary number of stations with an arbitrary number and configuration of servers at each station, customer classes, service protocols, and sizes of waiting rooms. We treat the entire system as a black box. An arrival to this system is either turned away on account of a full waiting room or the arrival enters the system. The stream of arrivals that enter the queueing system is characterized by arrival times $a_1 \le a_2 \le \cdots \le a_j \le \cdots$, where $a_j$ denotes the arrival epoch of the $j$th arrival in sequence, and corresponding service times $(s_1, \ldots, s_j, \ldots)$. Each customer is served either as soon as it arrives or according to some service protocol. Given this basic setup, queueing models are frequently used to characterize the following stochastic processes:

$$N_q(t) = \text{number of customers in queue at epoch } t, \tag{2.1}$$

$$N(t) = \text{number of customers in the queueing system at epoch } t. \tag{2.2}$$

Clearly, $N(t) = N_q(t) +$ the number of customers receiving service at time $t$. Similarly, for the $j$th customer, the quantities of interest are:

$$W_j = \text{time in queue of the } j\text{th customer}, \tag{2.3}$$

$$D_j = W_j + S_j = \text{total delay of the } j\text{th customer}. \tag{2.4}$$

Note that $S_j$ in the above expression denotes the random service time of the $j$th customer. For queueing systems with finite waiting rooms, we are also interested in the probability that an arrival is turned away.

The purpose of mathematical models of queues is to obtain closed-form or recursive formulae that allow system designers to calculate performance metrics such as average queue length, average waiting time, and the proportion of customers turned away. We say that mathematical models are tractable when closed-form or recursive formulae can be obtained, and in such cases the resulting expressions for the performance metrics are referred to as “analytical results.” Note that it is always possible to write equations that describe how the number of customers in each queue in the queueing system of interest changes over time. Such equations can be used to simulate a queueing system’s performance. In this chapter, the simulation-based results are also referred to as numerical solutions.

In the vast majority of cases, analytical results are possible only for limiting behavior (called steady state) of the above-mentioned performance metrics and in particular for time-average or customer-average metrics, when such averages exist. Specifically, steady-state analogs of $N_q(t)$, $N(t)$, $W_j$, and $D_j$ are obtained when either $t \to \infty$ or $j \to \infty$ and the limiting random variables exist. Loosely speaking, steady-state performance refers to the performance of a system with time-stationary parameters that has been in operation for a sufficiently long time such that time $t$ no longer affects the distributions of number in system, number in different queues, waiting times, and total delay. In contrast, transient queues arise when either system parameters are not time-stationary (therefore a steady state does not exist) or the queueing system does not remain in operation long enough to reach a steady state. If the purpose of the analysis is to obtain performance measures related to transient queues, then that often requires numerical analysis. Many healthcare facilities, such as outpatient clinics, are open for a fixed amount of time during the day and experience time-varying customer arrival patterns. Emergency departments, on the other hand, have demand that varies by the time of day, day of week, and month to month. In such instances, a steady-state may not exist. Still, analysis of steady state behavior can provide useful guidelines for making operational decisions.

Queueing systems in healthcare operations are complex. An example of patient flows through various units of a particular hospital is shown in Fig. 2.1. In this diagram, "out1" denotes the point of entry into the hospital and "out2" denotes the departure point. Ovals represent service stations, each of which is either an inpatient unit or a service department. The numbers shown in the ovals are ward numbers for inpatient units. The rest of the labels can be explained as follows. CL denotes the cath lab, DA refers to direct admits, ED is the emergency department, IR is the interventional radiology department, and PACU is the postanesthesia care unit. The numbers on the connecting arcs are the annual percent of patients that flow in and out of each service station, and arrows show the direction of flow. Each service station provides service to multiple customer classes with different service time distributions (referred to as lengths of stay among inpatient units), different service protocols, and different number of resources (e.g., beds, nurses, and physicians).

Queueing models for systems such as those shown in Fig. 2.1 are intractable unless one makes a number of simplifying assumptions. For these reasons, queueing systems as complex as those shown in Fig. 2.1 are not typically analyzed with the help of mathematical models. Instead, discrete-event simulation, where a computer samples values from different probability distributions to schedule events such as patient arrivals or service completions and keeps track of relevant statistics, is used to analyze such systems and obtain performance metrics. Discrete-event simulation techniques are discussed in Chapter 3 of this book. We focus on relatively simpler models that are tractable and provide useful insights for healthcare operations managers.

The organization of the rest of this chapter is as follows. Basic notation and terminology is introduced in Sect. 2. Single-station models are presented in two sections: Sect. 3 considers models in which there is a single server at each station, whereas Sect. 4 allows multiple servers. Basic results for queueing networks are presented in Sect. 5, and priority queues are discussed in Sect. 6. We conclude the chapter in Sect. 7.

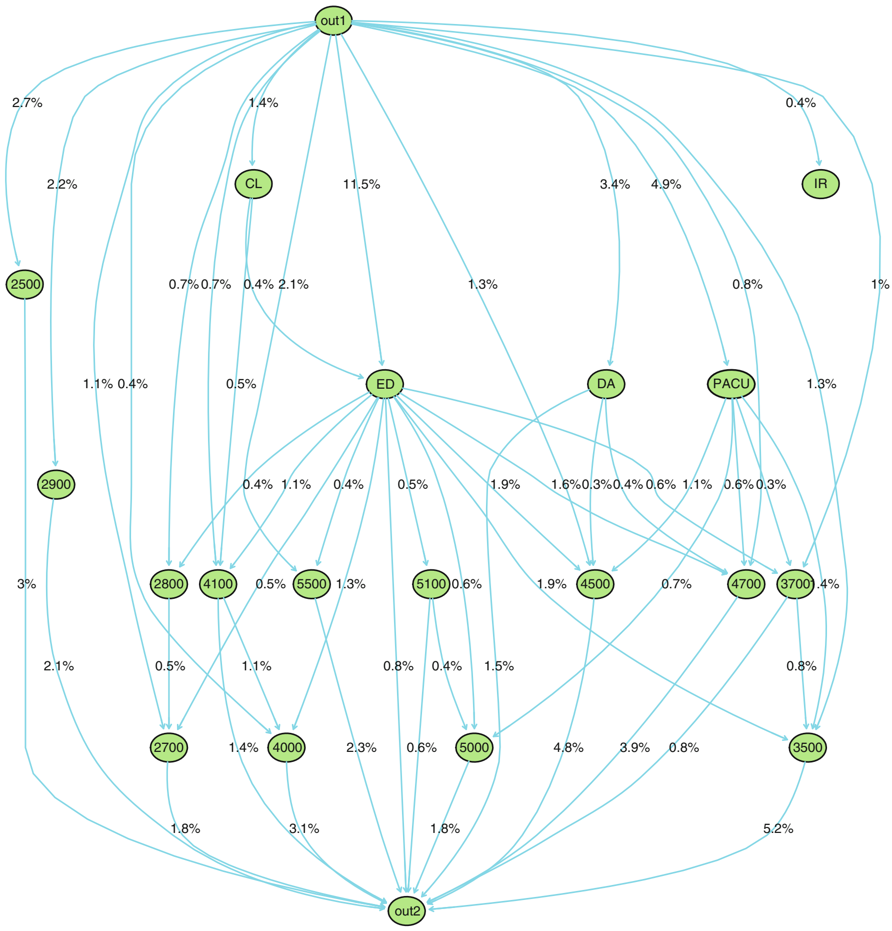


**Fig. 2.1** Patient flows through a general hospital

## 2 Basics

Let $N_q := \lim_{t\to\infty} N_q(t)$, $N := \lim_{t\to\infty} N(t)$, $W := \lim_{j\to\infty} W_j$, and $D := \lim_{j\to\infty} D_j$ denote steady-state distributions of quantities introduced earlier. It is assumed that such limits hold with probability 1. Additionally, we define

$$L = E[N], \tag{2.5}$$

$$L_q = E[N_q], \tag{2.6}$$

$$w = E[W], \qquad \text{and} \tag{2.7}$$

$$d = E[D] \tag{2.8}$$

as time or customer averages. We also define $A(t) =$ the number of customer arrivals during $(0,t]$ and $\lambda = \lim_{t\to\infty} A(t)/t$ as the mean arrival rate. Then, a key result in queueing theory, known as Little's law, is the following relationship:

$$L = \lambda w. \tag{2.9}$$

Little's law is extremely useful for carrying out rough-cut capacity calculations. Consider the following example. Suppose that an emergency department (ED) of a hospital receives on average 50 new patients in each 24 h period. Of the 50 patients, 22 are discharged after examination and treatment in the ED. The remaining 28 are admitted to the hospital as inpatients for further observation and treatment. The average length of inpatient stay is 3 days. Given this information, Little's law allows us to estimate that on average 84 inpatient beds would be needed to serve the needs of patients that are admitted via the ED. This comes from observing that $w = 3$ days, $\lambda = 28$ inpatients per day, and therefore, $L = \lambda w = 84$. Very few assumptions are made when arriving at this result. For example, Little's law remains valid regardless of the priority of service of arriving patients and differences in their service times in the ED.

Single-station queueing systems are often referred to by their four-part shorthand notation $A/B/m/K$, where $A$ and $B$ describe the inter-arrival and service time distributions, $m \in \{1,\ldots,\}$ is the number of servers, and $K \geq m$ is the size of the waiting room including customers in service. The fourth descriptor $K$ is omitted if there is no limit on the size of the queue. Typical values of $A$ and $B$ are as follows: $M$ for exponential, $D$ for deterministic, $E_k$ for Erlang with $k$ phases, $PH$ for phase type, $GI$ for general independent, and $G$ for general. In this notation, $M/G/2/20$ refers to a queueing system consisting of two servers at a single station, exponential inter-arrival times, general service times, and a waiting room capacity of 20.

Literature on queueing systems can be broadly divided into two categories: (1) models that focus on steady-state behavior, that is, stationary distributions $N_q$, $N$, $D$, and $W$, and (2) models that attempt to characterize transient behavior, that is time-dependent distributions of the number of customers and their waiting times. As mentioned earlier, the vast majority of analytical results pertain to steady-state behavior. Queueing models may be further classified into single- or multiple-station (network) models and those with single or multiple customer classes. In the remainder of this chapter, we provide a summary of key results pertaining to steady-state performance evaluation of single station (with single and multiple servers) and network models. In both types of models, we assume that there is a single customer class. Models with multiple customer classes and service priority are discussed briefly in Sect. 6.

# 3 Single-Station, Single-Server Models

Single-server queueing models with infinite queues are the workhorses of queueing theory, and among this class of models, the most commonly studied models are the $M/M/1$, $M/G/1$, and $GI/G/1$ models. The popularity of the first two models in this list is in part due to the fact that they are mathematically tractable, which in turn comes from the presence of Markovian property (the $M$ in the model descriptor), and in part from the fact that exponential distribution is a good fit for customer inter-arrival times in many real systems. We describe key results for each of these systems below. Details of analyses that lead to these results can be found in one of several queueing theory books cited earlier. All of these models assume independent and identically distributed inter-arrival and service times. We use the following notation for reporting the key results:

$p(n) = P(N = n),\ n = 0, 1, \ldots,$ the probability distribution of number in system
$F_S(x) = P(S \leq x),\ x \geq 0,$ the CDF of service time distribution
$\mu = 1/E[S],$ the mean service rate
$\rho = \lambda/\mu,$ the server utilization rate
$F_W(x) = P(W \leq x),\ x \geq 0,$ the probability distribution of customer wait time
$F_D(x) = P(D \leq x),\ x \geq 0,$ the probability distribution of customer delay
$F^*(s) = \int_0^\infty \mathrm{e}^{-sx} \mathrm{d}F(x),\ s \geq 0,$ the Laplace–Stieltjes transform (LST) of CDF $F(\cdot)$
$P(z) = \sum_{n=0}^\infty z^n p(n),$ the $z$-transform or probability generating function of $p(n)$

When waiting room is infinite and $\rho \geq 1$, queues can continue to grow over time. If $\rho > 1$, this happens because for each unit of time that the server is available, the average amount of work brought by new arrivals exceeds 1 unit. If $\rho = 1$, queues can still continue to grow because randomness in inter-arrival and service times can cause periods of server idling and the server can never make up for the lost work time. Note that in this instance, for each unit of time that the server is available, the average amount of work brought by new arrivals is exactly 1 unit. The effect of periods of idleness is cumulative and queues continue to grow. In such cases, stationary distributions of $N$ and $W$ do not exist. Therefore, we henceforth assume $\rho < 1$ in all models with infinite waiting room. Finally, we need to specify the service protocol in order to calculate the waiting time distribution. Throughout this section and in Sects. 4 and 5, we assume the first-in-first-out (FIFO) protocol.

## 3.1 Models with Infinite Waiting Room

In this section, we discuss $M/M/1$, $M/G/1$, and $GI/G/1$ queueing models.

### The M/M/1 Model

In healthcare settings, the $M/M/1$ model may prove to be useful either because it fits reality well or because it serves as a reasonable approximation for first-pass analysis. For example, it may be a reasonable choice for modeling walk-in clinics, pharmacy operations, and patient check-in and registration services at hospitals. Similarly, even in situations where customer arrival and service rates vary over time, $M/M/1$ models may be used to estimate capacity requirements to keep peak-period congestion within tolerable limits.

The following distributions and mean performance metrics can be calculated for $M/M/1$ queues from either Chapman–Kolmogorov equations, or an analysis of the embedded Markov chain at customer arrival and/or departure epochs:

$$p(n) = (1-\rho)\rho^n,\ n = 0,1,\ldots, \tag{2.10}$$

$$E[N] = \frac{\rho}{1-\rho}, \tag{2.11}$$

$$F_D(x) = 1 - \mathrm{e}^{-\mu(1-\rho)x},\ x \geq 0, \tag{2.12}$$

$$E[D] = \frac{1}{\mu(1-\rho)}, \text{ and} \tag{2.13}$$

$$E[W] = \frac{\rho}{\mu(1-\rho)}. \tag{2.14}$$

The analysis also utilizes an important property called PASTA—Poisson arrivals see time averages. Loosely speaking, this property ensures that if system state were observed at moments that coincide with Poisson arrivals, then system properties calculated from these observations are also time average system properties. The PASTA property greatly simplifies the analysis of Markovian queues.

Another important property of $M/M/1$ queues is that the distribution of the number of departures from such queues is also Poisson with parameter $\lambda$. It should be clear from the law of conservation of entities that the mean departure rate must equal the mean arrival rate. It is interesting to find that the distribution of departures is also identical to the distribution of arrivals. This property leads to a class of tractable queueing network models called Jackson networks (see Sect. 5).

From expressions (2.11), (2.13), and (2.14), we observe that the expected number in the system, the expected delay, and the expected waiting time are highly sensitive to the server utilization rate. In particular, all three quantities are increasing in $\rho$ at an increasing rate, and as $\rho \to 1$, all three quantities $\to \infty$. This helps explain the fundamental trade-off in queueing systems design. When service times and/or inter-arrival times are random, higher server utilization (efficiency) comes at the cost of increased congestion and customer waiting. High utilization rate may not be economical in situations where waiting cost is very high, and conversely, excess capacity may not be economical when resource cost is significantly higher than waiting cost. Upon knowing the cost of customer waiting and the cost of providing service resources, designers can find the economic balance between congestion and efficiency.

### The M/G/1 Model

In many healthcare settings, it is not appropriate to assume exponentially distributed service times. For example, service times for flu shots, lab services (blood draws), and magnetic resonance imaging (MRI) may not vary much from one patient to another. In the case of surgery practices, it is often found that the lognormal distribution provides a good fit for surgery durations. In such cases, service times may be better modeled by a distribution other than exponential, and the $M/G/1$ model may be more appropriate for calculating queue length and waiting time statistics. For the $M/G/1$ model, the following results have been established:

$$P(z) = \frac{(1-\rho)(z-1)F_S^*(\lambda - \lambda z)}{z - F_S^*(\lambda - \lambda z)}, \tag{2.15}$$

$$E[N] = \frac{\lambda^2 E[S^2]}{2(1-\rho)} + \rho, \tag{2.16}$$

$$F_D^*(s) = \frac{(1-\rho)sF_S^*(s)}{s - \lambda(1 - F_S^*(s))}, \tag{2.17}$$

$$E[D] = \frac{\lambda E[S^2]}{2(1-\rho)} + E[S], \tag{2.18}$$

$$E[W] = \frac{\lambda E[S^2]}{2(1-\rho)}. \tag{2.19}$$

Higher moments of the distribution of the number in system and customer delay can be obtained by differentiating $P(z)$ and $F_D^*(s)$. Also, distributions of $N$ and $D$ can be computed numerically by utilizing recent developments in the area of numerical inversions of transforms (see, e.g., Abate and Whitt 1992). However, closed-form expressions for the distribution of $N$ and $D$ are difficult to obtain except for some specific service time distributions.

Expressions (2.16), (2.18), and (2.19) can be recast by expressing $E[S^2]$ in terms of $C_s^2$, the squared coefficient of variation of service times. Note that the squared coefficient of variation of a random variable is the ratio of its variance to square of its mean. In particular, the expected waiting time expression for $M/G/1$ queues is

$$E[W] = \frac{\rho(1 + C_s^2)}{2\mu(1-\rho)}.$$

Upon comparing expressions (2.14) and (2.19), one can better understand the effect of service time variability. When $C_s^2 = 1$, we recover the expected waiting time in $M/M/1$ queues given in (2.14). All other parameters of the queueing system remaining unchanged, the mean waiting time increases linearly in $C_s^2$ at rate $\rho/(2\mu(1-\rho))$. This means that the negative effect of $C_s^2$ on customer waiting time is magnified nonlinearly as $\rho$ increases.

### The GI/G/1 Model

The $GI/G/1$ model requires the fewest assumptions about the shape of inter-arrival and service time distributions among the three models we discuss in this section. As such, it is useful in many settings. For example, in addition to examples mentioned before, appointment systems for non-urgent office visits can be modeled as queues in which clinics choose the inter-arrival time of patients. Queueing theory-based approaches for modeling appointment systems utilize either $D/M/1$ or $D/G/1$ queueing models; see surveys of literature in Cayirli and Veral (2003) and Gupta and Denton (2008). We present an application of such models for retail health clinics later in this section.

Analysis of $GI/G/1$ queues requires solving Lindley's integral equation (Lindley 1952). Closed-form solutions, which would be of interest to those interested in the design of healthcare delivery systems, are difficult to obtain except for some specific distributions of inter-arrival and service times. Therefore, many papers have studied approximate methods. Common approaches fall into two categories—(1) approximate either $GI$ or $G$ by a specific distribution leading to a tractable model and (2) assume a structural form of the distribution of $N$ and estimate its parameters. In the first group of methods, commonly studied approximations include Erlang, phase type, and generalized hyperexponential distributions (see, e.g., Neuts (1981, 1989), Li (1997) for details). In the sequel, we present an example of the second approach because it requires knowledge of only the first two moments of inter-arrival and service time distributions and provides greater insights into the key drivers of congestion.

Suppose we can estimate the first two moments of the inter-arrival and service time distributions, but the precise form of these distributions is unknown. In particular, we assume knowledge of $C_s^2$ and $C_a^2$, the squared coefficients of variation of service and inter-arrival times. A variety of approximations have been proposed for calculating the mean number in system in $GI/G/1$ queues given $C_a^2$ and $C_s^2$. The following is an example of a commonly used expression (see Buzacott and Shanthikumar 1993, p. 75):

$$E[N] \approx \left(\frac{\rho^2(1+C_s^2)}{1+\rho^2 C_s^2}\right)\left(\frac{C_a^2+\rho^2 C_s^2}{2(1-\rho)}\right)+\rho. \tag{2.20}$$

It is easy to verify that when $C_a^2 = 1$, the above expression reduces to the expression we obtained in (2.16) for an $M/G/1$ system. From (2.20), one can also derive expressions for mean waiting time and mean delay with the help of Little's law.

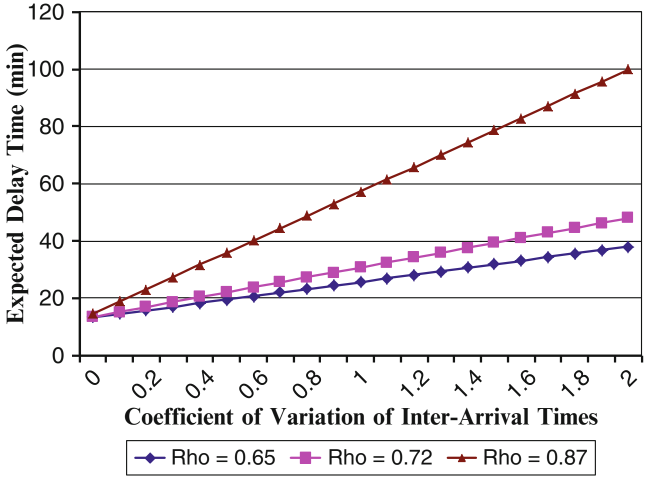


**Fig. 2.2** Effect of $C_a^2$ on expected delay. This figure shows that greater inter-arrival time variability causes expected customer delay to increase much more rapidly when the server has a higher workload (i.e., greater value of $\rho$)

The expression for $E[N]$ in (2.20) shows that both inter-arrival and service time variability contribute to the congestion in the system and that the effect of variability is magnified by server utilization—the higher the utilization, the greater the effect of variability on congestion. We illustrate the importance of this observation in healthcare operations with the help of an example next.

Retail healthcare clinics, such as MinuteClinic, RediClinics, and Target Clinics, promise to serve patients with routine diagnoses such as ear infections, flu, and minor injuries in a short amount of time without requiring appointments. The success of such clinics depends on providing timely service to a stream of customers that arrive randomly. The study of $GI/G/1$ queueing model can help shine light on the potential benefit to such service providers from reducing inter-arrival time variability by broadcasting time-of-day-dependent estimated waiting times on the web and other promotional media and encouraging customers to time their arrivals when the clinics are not too busy. For example, Target Clinics advise potential visitors about waiting times as follows: "… we are likely to be busiest before and after the average work/school day (8–10 a.m. and 5–7:30 p.m.)" (see response to the frequently asked question "How long will I have to wait to be seen?" at http://sites.target.com/site/en/spot/clinic_faqs.jsp#4). Such practices have the effect of making arrivals more uniform throughout the day, thereby reducing $C_a^2$. The author obtained data from a retail healthcare clinic chain and studied the effect of variability of arrival pattern on customer waiting times. The results are shown in Fig. 2.2 for three levels of server utilization commonly observed in the data at different clinics. This analysis also shows that if inter-arrival time variability cannot be reduced, clinics need to operate in a range where server utilization would be low in order to keep waiting times from becoming very long.

Without specifying the distributions of inter-arrival and service times, it is not possible to obtain expressions for the distributions of $N$ and $D$. Therefore, a variety of approximations have been proposed in the literature. In one such approximation, it is assumed that $p(n) = k\sigma^{n-1}$ for $n \geq 1$, where $k$ is a constant, and $p(0) = (1-\rho)$. Note that this was the form of the distribution of number in the system in $M/M/1$ queues. From the requirement that total probability must equal 1, we obtain that $k = \rho(1-\sigma)$. Furthermore, $\sigma$ can be estimated by equating the calculated mean of the approximate distribution, which equals $\rho/(1-\sigma)$, with the approximate value of $E[N]$ calculated in (2.20). Details of the accuracy of this approximation can be found in Buzacott and Shanthikumar (1993, Sect. 3.3.4).

## *3.2 Models with Finite Waiting Room*

Overcrowding in urgent care clinics and emergency departments is quite common. When waiting rooms become full, patients may leave without receiving service or the service facility may temporarily stop accepting new arrivals. To model such situations, we next consider models in which the maximum number of customers in the system is restricted to $K$, including the customer in service. When the waiting room limit is reached, one of two possibilities is typically modeled. Either additional arrivals are discouraged until the waiting room is no longer full, or additional arrivals continue to occur but depart immediately upon observing a full waiting room. The latter is identified in the literature as the lost sales model. It turns out that whether arrivals are discouraged or lost makes no difference when arrivals are Poisson. However, the pattern of arrivals when waiting room limit is reached does matter for queues with non-Poisson arrivals. We discuss each of the three basic models next and provide two examples of the usefulness of the $M/M/1/K$ model in healthcare setting. Note that $\rho < 1$ is no longer required for stability of queues. Stability is guaranteed because queue size cannot exceed the size of the waiting room $K$.

### The M/M/1/K Model

The following results are well known for $M/M/1/K$ queueing systems:

$$p(n) = \begin{cases} \dfrac{(1-\rho)\rho^n}{1-\rho^{K+1}} & n = 0, 1, \ldots, K, \\ 0 & \text{otherwise}, \end{cases} \tag{2.21}$$

$$E[N] = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}, \tag{2.22}$$

$$E[D] = E[N]/\lambda. \tag{2.23}$$

The expected waiting time can be calculated from (2.23) and the fact that an average customer spends $E[S]$ in service. Upon comparing (2.22) and (2.23) to their counterparts in (2.11) and (2.13), where the latter allow infinite waiting room, it is easy to see that when all parameters of the two types of queueing systems are identical, both the mean number in the system and the mean delay are smaller in situations where waiting room is limited. This should not come as a surprise because $\lambda p(K)$ fraction of arrivals is not served when waiting room is limited.

The $M/M/1/K$ model can be utilized to make capacity choices for emergency departments. One capacity parameter concerns the number of medical staff, which determines the service rate $\mu$. A second parameter concerns the number of ED beds, which determines the mean waiting time and the number of potential ED arrivals turned away. The latter is sometimes called ambulance diversion. Both types of capacities give rise to different fixed and operating expenses for the hospital. In addition, there are different implications for patient wait times.

For a fixed level of medical staff, that is, fixed service rate, if a hospital increases the number of ED beds, then this would result in greater mean waiting times for patients, but fewer ambulance diversions. Longer wait times can increase a hospital's risk from possible adverse events (e.g., poorer health outcomes and even deaths), and turning away more patients can lower a hospital's revenue because of reduced patient volume, giving rise to the trade-offs we discuss in detail below. Note that the model presented in this chapter is a highly stylized model. EDs are served by teams of multiple doctors and nurses working in parallel, diversions can be caused by a whole host of reasons including lack of availability of inpatient beds, and a variety of regulations may affect a hospital's decision (ability) to go on ambulance diversion. We smooth out such complexities in the discussion that follows.

For each fixed level of $\rho$, ED managers can develop trade-off curves between ambulance diversions and mean patient waiting times to identify the right combination of capacity parameters, as shown in Fig. 2.3. In this example, it is assumed that if patients are diverted on account of all ED beds being full, then they are able to find appropriate care at other hospitals located in geographical proximity to the hospital in question. If this were not the case, then a network model with strategic capacity choices by administrators of different hospitals will be required (Deo and Gurvich 2011). We do not discuss such models as they are beyond the scope of this chapter.

We consider an ED processing rate of ten patients per hour and two different peak-load scenarios, one with average patient arrival rate of 9 per hour and the other with average patient arrival rate of 9.9 per hour. These scenarios give rise to $\rho = 0.9$ and 0.99, respectively. Figure 2.3 shows that adding more beds (increasing $K$) increases delay, but reduces the average number of patients turned away, which is denoted as "loss" in Fig. 2.3b. Whereas $E[D]$ increases almost linearly in $K$, the benefit of having more beds in terms of reduction in expected loss exhibits diminishing returns. That is, each additional bed serves to reduce the expected number of patients turned away by a smaller amount. Service effectiveness can be improved by using triage (prioritizing patients) to identify and serve more critical patients first. We briefly discuss priority queues in Sect. 6.

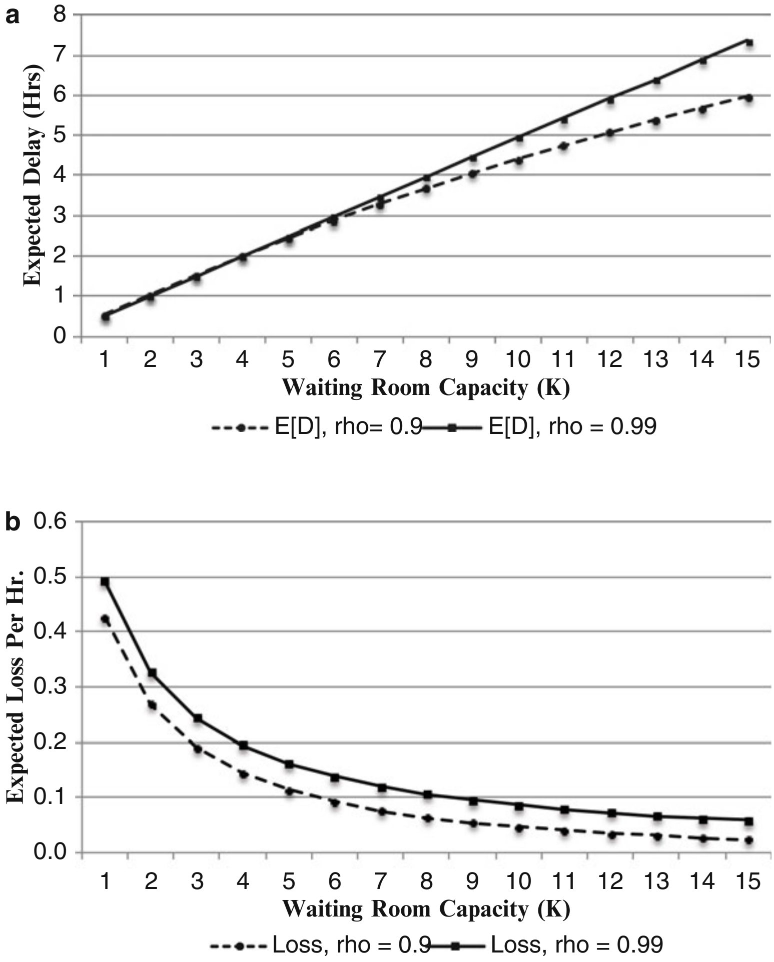


**Fig. 2.3** Trade-offs in ED capacity choices

In yet another example, the $M/M/1/K$ model can be used to perform first-cut capacity calculations for physician panel sizes. A panel refers to the list of patients who choose a particular primary care provider as their preferred provider. Suppose a physician implements the *Advanced Access* approach to servingpatients (Chap. 8). In this approach, patients are offered appointments on the day they call, eliminating queues. The physician can serve on average $K$ patients per day and arrival rate is $\lambda M$ where $M$ is the panel size and $\lambda$ is the incidence rate per patient. Assuming a national average visit rate of 3.356 office visits per patient per year (Hsiao et al. 2010), and

**Table 2.1** Panel size $(P)$ and service capacity $(K)$

| $K$ | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|
| $P$ | 1,208 | 1,296 | 1,386 | 1,463 | 1,540 | 1,617 | 1,701 | 1,779 | 1,857 |

260 working days per year (52 weeks, 5 days per week), we obtain $\lambda = 0.01291$. Suppose the physician is willing to accept the possibility that 5% of the patients who call on any given day would not be accommodated that day, that is, the overflow rate should not exceed 0.05. From the analysis of $M/M/1/K$ model, above, we know that the overflow rate is simply $p(K) = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}}$. Setting this quantity equal to 0.05, we can calculate $\rho$ and subsequently $M$ because $\mu = K$ patients per day. Upon performing these calculations, we obtain estimates of maximum panel sizes for different values of $K$ as shown in Table 2.1.

The modeling approach described above can be refined to include patient no-shows and advance-book appointments (see, e.g., Green and Savin 2008; Robinson and Chen 2010).

### The M/G/1/K Model

The classical analysis of $M/G/1/K$ queues relies on the embedded Markov chain observed at service completion epochs. This results in a series of equations relating state probabilities, which can be solved numerically along with the normalization equation (probabilities must sum to 1) to obtain the steady-state distribution of the number in the system. Closed-form expressions are often difficult to obtain. An alternative is to use the method of transforms, which was utilized in the section dealing with $M/G/1$ queues. That method is also primarily a numerical approach. Because our goal in this chapter is to shine light on the insights that queueing models provide for healthcare operations managers, we focus on a subclass of $M/G/1/K$ queueing models in which $\rho < 1$. Recall that $\rho < 1$ is not required for the existence of steady-state distributions of queue congestion and customer waiting times when waiting rooms are finite.

Given $\rho < 1$, let $p_\infty(n)$ and $\bar{P}_\infty(K)$ denote, respectively, the distribution of number in system and the CCDF of this distribution at $K$ in an $M/G/1$ queue. The subscript "$\infty$" emphasizes the fact that these quantities refer to the case in which there is no limit on the size of the waiting room. Then, the distribution of $N$ can be obtained as follows (see Buzacott and Shanthikumar 1993, p. 109 for details):

$$p(n) = \begin{cases} \dfrac{p_\infty(n)}{1-\rho\bar{P}_\infty(K)}, & n = 0, 1, \ldots, K-1, \\ \dfrac{(1-\rho)\bar{P}_\infty(K)}{1-\rho\bar{P}_\infty(K)} & n = K. \end{cases} \tag{2.24}$$

Expected number in system and mean delay can be calculated from the above expression. Remarkably, when $\rho < 1$, the distribution of number in system in an $M/G/1/K$ queueing system is proportional to the number in system in an $M/G/1/\infty$ system. The proportionality constant is $\frac{1}{1-\rho \bar{P}_\infty(K)}$ for $n = 0, 1, \ldots, K-1$, and $\frac{(1-\rho)}{1-\rho \bar{P}_\infty(K)}$ for $n = K$.

### The GI/G/1/K Model

The $GI/G/1/K$ model is more difficult to analyze, except when inter-arrival and service time distributions have some specific forms. Therefore, papers dealing with the analysis of $GI/G/1/K$ queueing systems propose a variety of approximations. A useful approximation that imposes the relationship between $M/M/1$ and $M/M/1/K$ models onto the relationship between $GI/G/1$ and $GI/G/1/K$ models is presented in Buzacott and Shanthikumar (1993, pp. 110–116). We omit the details in the interest of brevity.

## 4 Single-Station, Multiple-Server Models

In all of the examples mentioned in Sect. 3, for example, walk-in clinics, pharmacy and lab services, and emergency departments, situations involving multiple servers arise naturally. That is, service facilities in a healthcare setting often have multiple servers taking care of customers who queue up for similar services. In this section, we focus on queueing systems with unlimited waiting room, constant arrival rate $\lambda$, and $m$ identical servers. Servers process one customer at a time, and each server can process customers at rate $\mu$ when busy. In this case, the overall service rate is a function of the number in system. In particular, if there are $n$ customers in the system, then $\mu_n$, the state-dependent service rate is

$$\mu_n = \begin{cases} n\mu & \text{if } 0 \le n \le m, \\ m\mu & \text{otherwise.} \end{cases} \tag{2.25}$$

The overall server utilization in this instance is $\rho = \frac{\lambda}{m\mu}$, and stationary distributions of $N$, $N_q$, $D$, and $W$ exist if $\rho < 1$. Because it is more difficult to obtain closed-form expressions for performance metrics of multiple-server queueing systems, we focus in this section on $M/M/m$ models, that is, on situations in which the inter-arrival and service times are exponentially distributed. For $M/M/m$ models, it can be shown that

$$p(n) = \begin{cases} p(0)\dfrac{(m\rho)^n}{n!} & \text{if } n \le m, \\ p(0)\dfrac{\rho^n m^m}{m!} & n \ge m, \end{cases} \tag{2.26}$$

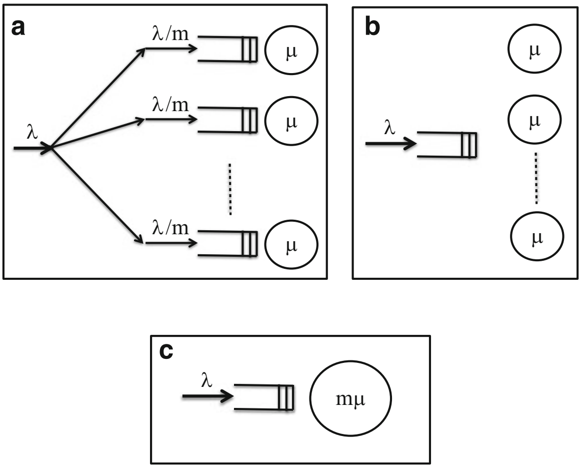


**Fig. 2.4** Comparison of server and queue configurations

where

$$p(0) = \left[ \sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left( \frac{(m\rho)^m}{m!} \right) \left( \frac{1}{1-\rho} \right) \right]^{-1}, \tag{2.27}$$

$$E[D] = \left( \frac{1}{m\mu(1-\rho)} \right) \left( \frac{(m\rho)^m}{m!(1-\rho)} \right) p(0) + \frac{1}{\mu}. \tag{2.28}$$

Expressions for mean delay in the system shown in (2.13) and (2.28) can be utilized to obtain a key result in the design of queueing systems—that combining queues and pooling servers reduces system delay. To demonstrate this, we show an example next in which three systems are compared. A schematic of these systems is shown in Fig. 2.4. Systems are labeled A, B, and C. System A consists of $m$ parallel queues each served by a single server. When a customer arrives, it is routed by a Markovian router that sends the customer to any one of the $m$ queues with equal probability $1/m$. Each server's processing rate is $\mu$. In system B, the queues are combined into a single queue. Customers wait for the next available server upon joining the queue. Each server's processing rate is $\mu$. In system C, queues are combined into a single queue, and the $m$ servers are replaced by a super server with a faster processing rate of $m\mu$. Note that the overall utilization rate remains $\rho = \lambda/m\mu$ in all three systems.

Choices similar to those depicted in Fig. 2.4 can arise in a number of different contexts in the healthcare setting. For example, the three configurations could represent choices for setting up patient registration and check-in counters at a hospital or clinic. System A represents a case in which each server specializes in serving a particular type of patient arrivals. In system B, each server can serve any arrival, and finally in system C, technology may be employed to assist a server,

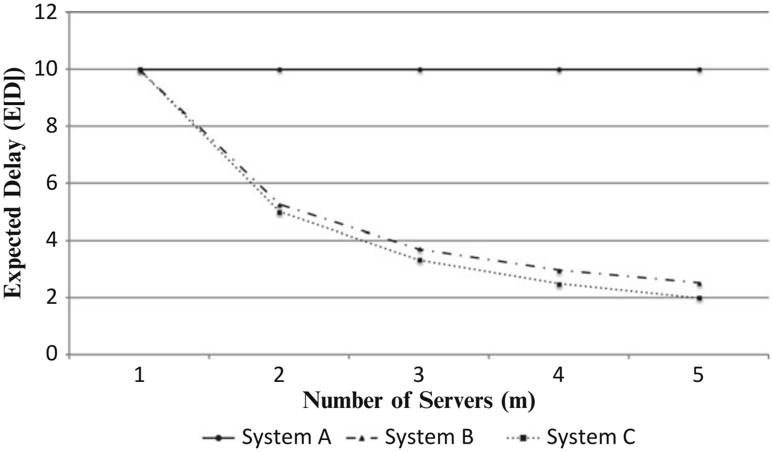


**Fig. 2.5** Effect of server/queue pooling

making that server faster. Alternatively, servers in systems A and B could represent the choice between specialized and general-purpose beds in an inpatient unit. In each example, the choice of configuration affects labor and capital costs as well as patient wait times. Queueing models can help compare the impact of these configurations on patient wait times.

In system A, an arbitrary arrival joins one of the $m$ separate queues with equal probability. Therefore, an arbitrary arrival's expected delay (using (2.13)) can be written as follows:

$$E[D^{(A)}] = \frac{\rho}{\mu(1-\rho)} + \frac{1}{\mu}. \tag{2.29}$$

The delay experienced by an arbitrary arrival in system B is as shown in (2.28). Finally, in system C, the expected delay equals

$$E[D^{(C)}] = \frac{\rho}{m\mu(1-\rho)} + \frac{1}{m\mu}. \tag{2.30}$$

To compare the mean delay in the three systems, consider an example in which $m$ is systematically varied from 1 through 5 while keeping $\rho = 0.9$ fixed. This can be achieved by setting $\mu = 1$ and varying $\lambda$ as $m$ varied. In particular, with $\mu = 1$, set $\lambda = m\rho$ to maintain a fixed $\rho$. Next, suppose we use the expressions derived above to calculate mean delay and plot the values in Fig. 2.5. Note that the delay in system A remains invariant in $m$. This is expected because customer waiting occurs in one of the $m$ queues, each of which is independent of $m$. We observe a significant improvement from combining queues, and a further improvement (though much smaller) from creating a single faster server. Most of the benefits appear to come from combining queues. This example serves to highlight a general principal for the design of healthcare service systems; namely, combining queues improves efficiency.

A common reason for creating separate queues is that each queue is served by a group of servers who possess special expertise to serve a particular type of customers. Combining queues requires that servers be trained to serve all types of customers. This can be expensive, and such considerations have motivated the study of partial pooling arrangements. Reasons why combining queues is not always beneficial can be found in Rothkopf and Rech (1987), and general principles of work design and pooling have been discussed in Buzacott (1996) and Mandelbaum and Reiman (1998).

Before closing this section, we briefly discuss the Erlang loss formula, which can be used to calculate overflow probability in $M/M/m/m$ systems. These are multiple-server queueing systems in which $K = m$, that is, waiting room size equals the number of servers. In such cases, a customer is lost if no server is available to serve this customer immediately. Erlang loss formula is of great interest in telephony, where it is used to calculate the number of telephone lines needed to accommodate a desired fraction of incoming calls. Erlang loss, or the probability of finding all $m$ servers busy, is given by

$$p(m) = \frac{\rho^m}{m!(\sum_{k=0}^{m} \rho^k/k!)}. \tag{2.31}$$

Erlang loss formula has been used to model capacity requirements of EDs where each ED bay (bed plus care team) is treated as a server.

# 5 Network Models

Hospitals and specialized treatment facilities for particular medical conditions (e.g., cancer, cardiovascular, or neurological services) perform a range of services, each with its own resources (servers) and queues. Such facilities are best modeled as networks of queues. The simplest network model from the viewpoint of obtaining analytical results is the multi-station (network) analog of the $M/M/1$ queueing system with $J$ service stations, each with a single server. Customers may arrive either from outside or move from one station to another. Suppose server $i$'s service rate is $\mu_i$, customers are routed from station $i$ to $j$ according to probability $r_{ij}$, and exogenous arrival rate at station $i$ is $\gamma_i$. Then, the effective arrival rate at station $i$ is

$$\lambda_i = \gamma_i + \sum_{1\le j\le J} \lambda_j r_{ji}, \qquad \text{for each } 1 \le i \le J. \tag{2.32}$$

Similarly, the server utilization rate is $\rho_i = \lambda_i/\mu_i$, for each $i$. A network is called *open* if it allows exogenous arrivals and departures. Departures can be modeled by designating a particular station to serve as a sink, say station indexed $J$, such that $r_{Ji} = 0$ for all $i$. Similarly, a network is called *closed* if no customers can enter or leave the network. In this case $\gamma_i = 0$ and there is no sink.

The state of the number in different stations of a network is a vector $n = (n_1, \ldots, n_J)$. Let $p_i(n_i)$ denote the distribution of number in system of a $M/M/1$ queue with parameters $\lambda_i$ and $\mu_i$. Then, a key result in the analysis of Markovian open networks is that the stationary distribution $p(n)$ has a product form. In particular,

$$p(n) = p_1(n_1) \times p_2(n_2) \cdots \times p_J(n_J), \tag{2.33}$$

where $p_i(n_i) = (1-\rho_i)\rho_i^{n_i}$. Networks for which the distribution of number in the system has a product form are also called product-form or Jackson networks (see Jackson (1954, 1957)). The product-form structure remains intact when there are $m_i$ servers at each station, that is, we have a network of $M/M/m_i$, $1 \le i \le J$, queues. Because of the existence of a product form, the results from the analysis of $M/M/1$ and $M/M/m$ queues can be applied directly to such networks. For example,

$$E[N] = \sum_{i=1}^{J} \frac{\rho_i}{1-\rho_i}, \tag{2.34}$$

$$E[D] = \frac{1}{\lambda} \sum_{i=1}^{J} \frac{\rho_i}{1-\rho_i}. \tag{2.35}$$

Queueing network models have been studied extensively (Walrand 1988), and there are numerous manufacturing applications of these models (Buzacott and Shanthikumar 1993). Many of these models are not directly applicable to health systems design. Each model is specific to a particular type of system (e.g., transfer lines with limited buffer) and typically requires either special techniques or approximations to derive system performance measures. Therefore, we focus only on papers that utilize queueing network methodology for modeling healthcare operations.

Whereas there are many attempts to represent networks of healthcare facilities as networks of queues, these networks are typically not analyzed using queueing-theoretic approaches. Instead, a common approach is to use computer simulation to obtain performance metrics of interest; for examples, see Taylor and Keown (1980), Harper and Shahani (2002), and Feck et al. (1980). Papers that use an analytic approach include Hershey et al. (1981) and Weiss and McClain (1987). Hershey et al. (1981) present a methodology for estimating expected utilization and service level for a class of capacity-constrained service network facilities operating in a stochastic environment. In this paper, queues are not allowed to form, that is, waiting room size equals the number of servers at each facility, and the authors use the Erlang loss formula to approximate the probability of overflow. They show that their calculation is exact for two cases and recommend its use as an approximation in the general case.

Weiss and McClain (1987) model the transition of care from acute to extended care (e.g., a nursing home or community care center). Inadequate capacity at downstream service facilities can lead to extra wait in the acute care facility, which is often referred to as "administrative days." The authors use a queueing-analytic approach to describe the process by which patients await placement. They model the situation using a state-dependent placement rate for patients backed up in the acute care facility. Using data from seven hospitals in New York State, the study also derives policy implications.

## 6 Priority Queues

There are many variants of the basic models described in the preceding sections. These may consider different types of service priority (Jaiswal 1968; Takagi 1986, 1990, 1994; Gupta and Gunalay 1997), server vacations (Tian and Zhang 2006), and bulk arrivals and batch service (Chaudhry and Templeton 1986). The literature on each of these topics is vast. In this section, we discuss an assortment of results from priority queues and discuss their implications for healthcare operations.

Suppose in a single-server queueing system, there are $k$ customer classes, indexed by $\ell = 1,\ldots,k$. Type $\ell$ customers arrive according to an independent Poisson process with rate $\lambda^{(\ell)}$, and their service time distribution is $S^{(\ell)}$. An arrival observes $N^{\ell}$ type-$\ell$ customers in the system upon arrival. Therefore, the total expected work in the system at an arbitrary arrival epoch is $\sum_{\ell=1}^{k} E[N^{(\ell)}]E[S^{(\ell)}] - \sum_{\ell=1}^{k} \frac{1}{2}\lambda^{(\ell)}E[(S^{(\ell)})^2]$, where the second term is the amount of work already completed on the customer in service, if any. After some simplification and using Little's law, the total expected work can be expressed as $\sum_{\ell=1}^{k} \rho^{(\ell)}E[D^{(\ell)}] - \sum_{\ell=1}^{k} \frac{1}{2}\lambda^{(\ell)}E[(S^{(\ell)})^2]$, where $\rho^{(\ell)} = \lambda^{(\ell)}E[S^{(\ell)}]$. If the service protocol is work conserving, that is, it neither creates nor destroys work, then the expected total work must be constant. This immediately implies that

$$\sum_{\ell=1}^{k} \rho^{(\ell)}E[D^{(\ell)}] = \text{constant} \tag{2.36}$$

because $\sum_{\ell=1}^{k} \frac{1}{2}\lambda^{(\ell)}E[(S^{(\ell)})^2]$ is independent of service protocol. The queue is stable so long as $\sum_{\ell=1}^{k} \rho^{(\ell)} < 1$, which we assume throughout this section.

The above relationship establishes an important property of priority queues. When service protocol is work conserving, a particular priority scheme may affect delays experienced by different customer classes, but reduction in the expected delay of one customer class is realized at the expense of increase in the expected delay of another class. The work conservation principle is violated when switching from one customer class to another requires setup or switchover time, and/or some amount of work is lost if service of one customer type is preempted by a higher priority customer. Because switchover or setup times are common when a server attends to customers with different service requirements, pseudo work conservation laws have been derived for queues with switchover times and certain service protocols, for example, cyclic priority (see Takagi 1986, 1990, 1994; Gupta and Gunalay 1997). Although queues with switchover times are not work conserving, the amount of additional work created by certain switching protocols can be fully characterized. The expected total work in the system is then the sum of two components—work associated with customer arrivals, which is independent of service protocol, and work associated with switching regime, which is dictated by the priority scheme. For this reason, such queues are said to satisfy pseudo conservation laws.

In healthcare applications, one finds examples of both preemptive and non-preemptive priority. For example, ED physicians often serve the most critical patients preemptively in order to save lives. In the outpatient setting, specialists reserve certain appointment slots for high-priority patients, but once a low-priority patient books an appointment, he or she is rarely preempted during service. In many situations, service protocols may not be work conserving because priority rules may increase service providers' walking time to reach patients located in different inpatient units. Finally, service protocol may be static or dynamic. In a static protocol, each customer class has a fixed priority, and its members receive a strictly higher priority over all lower-ranked classes. In contrast, in dynamic priority protocols, customers of a particular class may be higher ranked at one time and lower ranked at another. A common dynamic priority protocol is one in which customers of each class form a separate queue, the server moves from one queue to another, and when serving a particular queue, all customers of that class (previously waiting or new arrivals) have higher priority. Such a queueing protocol is observed in healthcare operations when a physician travels to different community-based clinics on different days of week.

Next, we provide some basic results that allow an operations manager to quickly calculate mean delay experienced by customers of different classes in a system with a single server, Poisson arrivals, and work-conserving, non-preemptive, and static-priority service protocol. Suppose customers classes are arranged in the order of priority, that is, class-1 customers have the highest priority, followed by class 2, and so on until class $k$. Then, an arbitrary class-1 customer waits only for class-1 customers already in the queue when it arrives. Moreover, because of PASTA property, we have

$$E[W^{(1)}] = E[N_q^{(1)}]E[S^{(1)}] + \frac{\lambda E[S^2]}{2}, \tag{2.37}$$

where $\frac{\lambda E[S^2]}{2} = \sum_{\ell=1}^{k} \frac{1}{2}\lambda^{(\ell)} E[(S^{(\ell)})^2]$ is the amount of work remaining to complete the service of the customer in service. Introducing new notation $w_0 = \frac{\lambda E[S^2]}{2}$, $\lambda_r = \sum_{\ell=1}^{r} \lambda^{(\ell)}$, and $\rho_r = \sum_{\ell=1}^{r} \rho^{(\ell)}$, and using the fact that $E[W^{(1)}] = \frac{E[N_q^{(1)}]}{\lambda^{(1)}}$, we can rearrange (2.37) to obtain

$$E[N_q^{(1)}] = \frac{\lambda^{(1)} w_0}{(1-\rho_1)}. \tag{2.38}$$

Consider next an arbitrary class-2 arrival. This customer will be served only after all earlier-arriving class-1 and class-2 customers are served, which causes an initial delay of $E[N_q^{(1)}]E[S^{(1)}]+E[N_q^{(2)}]E[S^{(2)}]+w_0$. In addition, it must also wait until all those type-1 customers who arrive during $E[N_q^{(1)}]E[S^{(1)}]+E[N_q^{(2)}]E[S^{(2)}]+w_0$, and additional class-1 arrivals during the service time of those, and so on, are served. It turns out that the corresponding delay has mean duration

$$(1/(1-\rho_1))\left\{E\left[N_q^{(1)}\right]E[S^{(1)}]+E\left[N_q^{(2)}\right]E[S^{(2)}]+w_0\right\}.$$

That is, $E[W^{(2)}]=(1/(1-\rho_1))\{E[N_q^{(1)}]E[S^{(1)}]+E[N_q^{(2)}]E[S^{(2)}]+w_0\}$. Upon using $E[N_q^{(2)}]=\lambda^{(2)}E[W^{(2)}]$ and simplifying, we obtain

$$E\left[N_q^{(2)}\right]=\frac{\lambda^{(2)}w_0}{(1-\rho_1)(1-\rho_2)}. \tag{2.39}$$

Continuing in the same fashion, we obtain for $\ell=1,\ldots,k$,

$$\begin{aligned} E[N_q^{(\ell)}] &= \frac{\lambda^{(\ell)}w_0}{(1-\rho_{\ell-1})(1-\rho_\ell)} \\ &= \frac{\lambda^{(\ell)}\lambda E[S^2]}{2(1-\rho_{\ell-1})(1-\rho_\ell)}, \end{aligned} \tag{2.40}$$

where $\rho_0=0$. Note the similarity between the above expression and the mean number in the queue for $M/G/1$ systems, where the latter can be calculated from (2.16).

Suppose $w_\ell$ is the cost of making a type-$\ell$ customer wait for service. What is an optimal priority rule that minimizes total waiting cost? This question has been addressed in the queueing literature, and it has been shown that class priority should be proportional to $w\mu$, that is priority index should be such that $w_1\mu_1 \geq w_2\mu_2 \geq \cdots \geq w_k\mu_k$. This means that customers with higher waiting cost per unit time and shorter mean processing time should be given higher priority. If all customer classes have the same per-unit time waiting cost, then customers with shorter mean processing time would be processed first. This is also called the shortest-processing-time-first rule.

## 7 Concluding Remarks

Many types of healthcare service systems are characterized by random demand (in timing and type of services required); time-varying and uncertain availability of service resources due to preferred work patterns, work rules, and planned or unplanned absences; and service protocols that assign different priorities to different customer classes (e.g., urgent versus nonurgent patients). These are precisely the types of environments in which queueing theory can be brought to bear to obtain useful insights for system design and for developing operating principles for service delivery systems. It is no surprise that queueing theory has been used extensively in healthcare operations. The following is a list of key topic areas and some recent papers on each of these topics:

1. Capacity calculations (matching supply and demand)
    (a) Panel size determination—See the discussion in Sect. 3.2, Green and Savin (2008), Robinson and Chen (2010), Gupta and Wang (2011).
    (b) ED beds—See the discussion in Sect. 3.2, Deo and Gurvich (2011) and references therein.
    (c) Network capacity—See Hershey et al. (1981) and Weiss and McClain (1987).
    (d) Nurse staffing—See Yankovic and Green (2011) and references therein.
2. Scheduling arrivals (appointment systems)—See Gupta and Denton (2008) for a review of queueing-analytic approaches
3. Priority queues (allocation of organs to transplant candidates)—See Su and Zenios (2004) and references therein.

Some of the above-mentioned problems were not discussed in this chapter because of the specialized institutional background necessary to introduce the key operations management challenges.

Notwithstanding the success of queueing models for addressing important questions in the delivery of healthcare services, there remain significant opportunities for new models and analytic tools. For example, hospitals can benefit from insightful network models of patient flow (recall Fig. 2.1). Hospitals have limited capacity within each inpatient unit, which leads to blocking at upstream units. This situation may be resolved by keeping patients longer in some units, placing patients in less-than-ideal units, transferring patients to other hospitals, or refusing admissions. Such decisions can affect lengths of stay and health outcomes (Rincon et al. 2011; Sinuff et al. 2004). Clearly, hospitals could benefit from knowing the performance implications of such practices and having access to models that allow them to factor nursing units' flexibility into capacity calculations. In the manufacturing setting, there are numerous dynamic job shop models that address similar problems; see, for example, Chap. 7 in Buzacott and Shanthikumar (1993). However, the number of similar models for hospital operations is quite limited and represents an opportunity for future research.

## References

Abate J, Whitt W (1992) The fourier-series method for inverting transforms of probability distributions. Queueing Syst 10:5–88

Bhat UN (2008) An introduction to queueing theory modeling and analysis in applications. Springer, Boston [distributor]

Buzacott JA (1996) Commonalities in reengineered business processes: Models and issues. Manag Sci 42:768–782

Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice Hall, Englewood Cliffs

Cayirli T, Veral E (2003) Outpatient scheduling in health care: A review of literature. Prod Oper Manag 12(4):519–549

Chaudhry ML, Templeton JGC (1986) Bulk queues. Department of Mathematics, McMaster University, Hamilton

Cohen JW (1969) The single server queue. North-Holland, Amsterdam

Cox DR, Smith WL (1961) Queues. Chapman and Hall, London; Distributed in the USA by Halsted Press, London

Deo S, Gurvich I (2011) Centralized vs. decentralized ambulance diversion: A network perspective. Manag Sci 57:1300–1319

Feck G, Blair EL, Lawrence CE (1980) A systems model for burn care. Med Care 18(2):211–218

Green LV, Savin S (2008) Reducing delays for medical appointments: A queueing approach. Oper Res 56(6):1526–1538

Gross D, Harris CM (1985) Fundamentals of queueing theory. Wiley, New York

Gupta D, Denton B (2008) Appointment scheduling in health care: Challenges and opportunities. IIE Trans 40:800–819

Gupta D, Gunalay Y (1997) Recent advances in the analysis of polling systems. In: Balakrishnan N (ed) Advances in combinatorial methods and applications to probability and statistics. Statistics in industry and technology series. Birkhauser, Boston

Gupta D, Wang WY (2011) Patient appointments in ambulatory care. In: Hall RW (ed) Handbook of healthcare system scheduling: Delivering care when and where it is needed. Springer, New York, chap 4

Harper PR, Shahani AK (2002) Modelling for the planning and management of bed capacities in hospitals. J Oper Res Soc 53(1):11–18

Hershey JC, Weiss EN, Cohen MA (1981) A stochastic service network model with application to hospital facilities. Oper Res 29(1):1–22

Hsiao CJ, Cherry DK, Beatty PC, Rechtsteiner EA (2010) National ambulatory medical survey report: 2007 summary. National Health Statistics Reports, Number 27. Available on the web at http://www.cdc.gov/nchs/data/nhsr/nhsr027.pdf. Cited 7 March 2011

Jackson JR (1957) Networks of waiting lines. Oper Res 5:518–521

Jackson RRP (1954) Queueing systems with phase-type service. Oper Res Q 5:109–120

Jaiswal NK (1968) Priority queues. Academic, New York

Larson RC (1987) Perspectives on queues: Social justice and the psychology of queueing. Oper Res 35(6):895–905

Li J (1997) An approximation method for the analysis of GI/G/1 queues. Oper Res 45(1):140–144

Lindley DV (1952) On the theory of queues with a single server. Proc Camb Philos Soc 48:277–289

Mandelbaum A, Reiman MI (1998) On pooling in queueing networks. Manag Sci 44(7):971–981

Morse PM (1958) Queues, inventories, and maintenance; the analysis of operational systems with variable demand and supply. Wiley, New York

Neuts MF (1981) Explicit steady-state solutions in stochastic models: An algorithmic approach. The Johns Hopkins University Press, Baltimore

Neuts MF (1989) Structured stochastic matrices of M/G/1 type and their applications. Marcel Dekker, New York

Newell GF (1982) Applications of queueing theory. Chapman and Hall, London

Rincon F, Morino T, Behrens D, Akbar U, Schorr C, Lee E, Gerber D, Parrillo J, Mirsen T (2011) Association between out-of-hospital emergency department transfer and poor hospital outcome in critically ill stroke patients. J Crit Care 26(6):620–625

Robinson LW, Chen RR (2010) A comparison of traditional and open-access policies for appointment scheduling. Manuf Serv Oper Manag 12:330–346

Rothkopf MH, Rech P (1987) Perspectives on queues: Combining queues is not always beneficial. Oper Res 35:906–909

Sinuff T, Kahnamoui K, Cook DJ, Luce JM, Levy MM (2004) Rationing critical care beds: A systematic review. Crit Care Med 32(7):1588–1597

Su X, Zenios S (2004) Patient choice in kidney allocation: the role of the queueing discipline. Manuf Serv Oper Manag 6:280–301

Takacs L (1962) Introduction to the theory of queues. Oxford University Press, New York

Takagi H (1986) Analysis of polling systems. MIT, Cambridge

Takagi H (1990) Queueing analysis of polling models: An update. In: Takagi H (ed) Stochastic analysis of computer and communication systems. North-Holland, Amsterdam, pp 267–318

Takagi H (1994) Queueing analysis of polling models: Progress in 1990–93. Institute of Socio-Economic Planning, University of Tsukuba, Japan

Taylor BW III, Keown AJ (1980) A network analysis of an inpatient/outpatient department. J Oper Res Soc 31(2):169–179

Tian N, Zhang ZG (2006) Vacation queueing models. Springer, New York

Walrand J (1988) An introduction to queueing networks. Prentice Hall, Englewood Cliffs

Weiss EN, McClain JO (1987) Administrative days in acute care facilities: A queueing-analytic approach. Oper Res 35(1):35–44

Wolff RW (1989) Stochastic modeling and the theory of queues. Prentice Hall, Englewood Cliffs

Worthington D (2009) Reflections on queue modelling from the last 50 years. J Oper Res Soc 60:s83–s92

Yankovic N, Green LV (2011) Identifying good nursing levels: a queuing approach. Oper Res 59(4):942–955

# Chapter 3
# Applications of Agent-Based Modeling and Simulation to Healthcare Operations Management

**Sean Barnes, Bruce Golden, and Stuart Price**

## 1 Introduction

Simulation techniques have been applied to problems in healthcare operations management for many years in several focus areas. Historically, research in this area has consisted of results derived from system dynamics (SD) and discrete event simulation (DES) methods (Fone et al. 2003; Jun et al. 1999; Koelling and Schwandt 2005). Both methods focus on system-level behavior, but they differ in how the system is modeled and how time is simulated. SD models represent entities as continuous variables whose states change continuously with time, whereas DES models contain individual components whose states only change at discrete moments in time. In either case, the goal is to aggregate the system behavior and draw conclusions about how the system evolves over time under internal and external forces. These techniques can provide valuable insight to problems in healthcare operations management and are ideally suited for many such problems.

In recent years, a new modeling and simulation methodology has gained momentum with respect to healthcare applications. The methodology is most commonly known as agent-based, or individual-based, modeling (ABM). In contrast to SD and DES methodologies, ABM focuses on modeling individuals, interactions between individuals, and in some cases, interactions with a physical or influential surrounding environment (Macal and North 2007b). This activity can then be aggregated to simulate how a system behaves over time. The focus on detailed, individual agents and their interactions makes ABM an ideal tool for analyzing complex systems such as healthcare facilities and organizations because there

S. Barnes (✉) • B. Golden • S. Price
Robert H. Smith School of Business, University of Maryland, College Park, MD, USA
e-mail: sbarnes@rhsmith.umd.edu; bgolden@rhsmith.umd.edu; sprice@rhsmith.umd.edu

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_3, © Springer Science+Business Media New York 2013

are many components to these systems and outcomes can be difficult to predict without adequate model representation. A comparison between the three modeling frameworks is shown in Table 3.1.

Agents can interact with each other in many ways. Interactions may occur in a spatial environment, which could be a simple one-dimensional ring, a two- or three-dimensional Cartesian grid, or a specific geographic location or region. Alternatively, interactions can lack spatial representation and instead be constrained by relational considerations, where agents only interact if there is an explicit connection between them, such as being a member of the same family, working together, or being cared for by the same healthcare worker. For these cases, physical space has no effect on the outcome of the interactions, and, therefore, it does not need to be modeled explicitly. Social network analysis (SNA) is often paired with ABM for problems in which the structure of the interaction network is not uniform, implying that each individual may only interact with a specific subset of the population. For these types of problems, agents are often represented by nodes in a network, and interactions are represented by edges, which can be weighted to represent the frequency or type of those interactions.

There are several advantages of ABM over SD and DES. First, ABM is a more realistic modeling approach for many problems, especially problems in which there are multiple types of actors that interact in different ways. For these cases, it is very straightforward to model these actors as agents that have distinct sets of behaviors and characteristics without making assumptions as to how the system would be affected by each type. Individuals are not typically represented in SD models, and in DES models, individuals are explicitly modeled, but their states are typically a function of their status in some type of predefined system process. In addition, agent-based models facilitate detailed analysis of both individual- and system-level behavior because metrics at each level can be updated with each interaction. Agent-based models are also easier to explain than most SD and DES models because of their direct correlation to reality, which is an important factor in gaining the confidence of healthcare professionals and ultimately having an impact. SD models often consist of mathematical models that can become quite complicated, and DES models are usually described by intricate flow diagrams. These abstractions can cause difficulty in explaining model concepts to healthcare professionals who are not trained in mathematical or computational disciplines.

As in all methodologies, there are disadvantages to ABM as well. ABMs can become very complex when they incorporate a lot of detail. When this happens, it becomes difficult to separate the actual effect of each input parameter in the model. In addition, agent-based models can become computationally expensive, requiring excessively long computer run times for simulations. This problem has been alleviated to some degree by high-performance and parallel computing techniques, but it demands additional developmental resources that are not typically required by SD or DES models. Agent-based models also face different challenges related to assumptions. SD and DES models often require more assumptions about system-level parameters (e.g., patient admission rates, average hand-hygiene compliance of healthcare workers), whereas ABM requires more assumptions about

**Table 3.1** Comparison of system dynamics (SD), discrete event simulation (DES), and agent-based modeling (ABM) methodologies

| Attribute | SD | DES | ABM |
|---|---|---|---|
| Model perspective | System level | System level | Individual level |
| Level of realism | Low | Moderate | High |
| Model flexibility | Low | Moderate | Very high |
| Time domain | Continuous | Discrete | Discrete |
| Run times | Very fast | Moderate | Slows with increasing system size and complexity |
| Model inputs | Rate parameters and flow characteristics | Entity types, arrival times, queuing parameters, resource scheduling, process flows | Agent characteristics and interactions, environment specification |
| Model outputs | Dynamics, steady-state values, analytic expressions | Wait times, resource utilization, throughput | Unlimited |
| Software | Spreadsheet (e.g., Microsoft Excel), any programming language or mathematical software (e.g., MATLAB[a], Mathematica[b]) | Object-oriented programming languages (e.g., C++, Python[c], Java), simulation software packages (e.g., Arena[d], AnyLogi[e], SIMUL8[f]), or mathematical software (e.g., MATLAB, Mathematica) | Object-oriented programming language (e.g., C++, Python, Java), open-source ABM software (e.g., NetLogo[g], Repast[h], MASON[i], Swarm) |

[a] http://www.mathworks.com/products/matlab/
[b] http://www.wolfram.com/mathematica/
[c] http://www.python.org
[d] http://www.arenasimulation.com/
[e] http://www.xjtek.com/
[f] http://www.simul8.com/
[g] http://ccl.northwestern.edu/netlogo/
[h] http://repast.sourceforge.net/
[i] http://cs.gmu.edu/~eclab/projects/mason/

**Table 3.2** Agent characteristics and definitions

| Characteristic | Definition |
|---|---|
| Autonomy | Agents act independently of other agents |
| Heterogeneity | Agent characteristics and evolution of state are sufficiently different for all agents |
| Awareness | Agents can have varying levels of knowledge of the system state, ranging from ignorance to omniscience |
| Memory | An agent can remember its state and/or the state of the system at earlier points in time |
| Adaptation | Agents can change their behavior over time based on the current state of the system or prior experience |
| Goal oriented | Agents act toward accomplishing an objective |
| Rationality | Agents act in their best interests |
| Interactivity | Agents can exchange information or resources with other agents |
| Reactivity | Agent state or behavior can change in reaction to the environment or changes in the behavior of other agents |
| Mobility | Agents can move within the environment |

individual-level parameters (e.g., probability of nurse-to-patient transmission for a given infectious disease) and the nature of interactions. Some of these requirements can be satisfied easily, by speaking with experts who have experience working in these environments, but others can be more difficult to quantify. However, the advantage of ABM is that few assumptions need to be made about the system as a whole, because system behavior is determined by the activities at the individual level.

Agents have several characteristics that can be used to distinguish ABMs from SD or DES models. These characteristics and their definitions are summarized in Table 3.2. For a more complete discussion on ABM, the reader is referred to Macal and North (2007a). There are additional characteristics associated with agent-based models, but this set is sufficient for identifying appropriate studies. Agent-based models do not necessarily possess all of these characteristics, and in some cases, the distinction between ABM and DES can be difficult. For the purposes of this review, each included research article must present the results of a simulation that models sufficiently heterogeneous agents that interact in a dynamic, nondeterministic manner. In general, articles that only describe the concept and implementation of an agent-based model without demonstrating insightful results were also excluded.

As with any modeling methodology, choosing an appropriate software package is critical during the early stages of developing an agent-based model. This selection process for healthcare applications depends primarily on three factors: the complexity of the model, the skill of the developer, and the level of interaction the consumer wants to have with the model. Developers with limited modeling experience and enthusiasm for modeling should identify the platform that is easiest to learn and provides the necessary functionality. More experienced developers are probably either skilled in a programming language that offers the required capability or willing to learn a new language if it is more appropriate to the nature of the model. On the delivery end, some consumers may only require the results from a model and do not need to interact with it in any way. In this case, visualization and animation would be helpful capabilities, but a graphic user interface would not be required. If the consumer will interact with the model regularly, a graphic user interface is necessary and proper training needs to be provided.

Software can range from commercial products to a variety of programming languages and open-source platforms. Commercial software is typically easy to learn and is accompanied by extensive documentation and customer support, but it may not provide the modeling flexibility necessary for modeling a complex system. Object-oriented programming languages are especially well suited for agent-based models because objects correspond well to agents and object methods correspond to agent behavior, which facilitates a relatively straightforward implementation. Mathematics-based software typically provides substantial built-in functionality, but computational times can be prohibitively slow for an agent-based model. Lower level programming languages typically execute significantly faster than the aforementioned software types, which may be critical for large models, but they lack the built-in functionality that helps to accelerate development times. There is a growing set of dedicated software packages that offer built-in functionality common to agent-based models. These tools vary significantly in terms of their user-friendliness, level of documentation, speed, and modeling flexibility, but their diversity demonstrates the variety of applications that can be modeled using agent-based methods.

In this chapter, we review and evaluate a selected body of research that has applied ABM and simulation techniques to healthcare operations management, which we interpret broadly. Specifically, research that applies agent-based methods in the following areas is included, based on surveys of simulation and system dynamics applications to healthcare (Fone et al. 2003; Jun et al. 1999; Koelling and Schwandt 2005):

- *Healthcare delivery*: studies that focus on a single facility (e.g., a hospital) or unit within a department (e.g., an emergency department) and emphasize patient flow characteristics, scheduling, or allocation of resources.
- *Epidemiology*: studies that focus on the spread of illness or disease or the physiological understanding of an illness or disease.
- *Healthcare economics and policy*: studies that focus on the financial and policy decisions made by hospitals or healthcare organizations which can include the purchase of equipment, pharmaceutical items, or other medical equipment.

We review the ABM literature for each topic and briefly describe the core methods, summarize the key results, and identify best practices. We will also highlight areas within each topic where ABM and simulation filled a significant gap that was not addressed previously by other methods. Finally, we will propose some new questions that may be of interest moving forward.

## 2 Healthcare Delivery

There are many different types of facilities that serve as healthcare providers. Each provider has its own set of objectives and constraints. Outpatient facilities tend to operate using appointments that are scheduled well in advance, whereas emergency rooms deal almost entirely with unscheduled arrivals. This section focuses on modeling hospitals and the individual departments therein. There are models of other facilities such as long-term care facilities, outpatient clinics, or diagnostics laboratories, but the literature is not as extensive as it is on hospitals.

Hospitals are complex service systems that operate under many constraints. Providing healthcare services to patients is their primary objective, but there are many obstacles to doing so effectively. First, almost every patient has a different set of medical needs. Therefore, hospitals need to have a robust strategy to accommodate their needs. In addition, some patients are scheduled, whereas other patients (e.g., emergency and ambulance patients) enter the hospital unexpectedly. Failure to meet these demands could adversely affect the health of the patients, a far greater consequence than just lost revenue.

Hospitals have to manage many resources—including physical and human—in order to meet the needs of the patients they serve. They employ physicians, nurses, technicians, administrative personnel, and many other support staff that help to provide the necessary care to patients. Employees must be scheduled in such a way as to properly staff the hospital in the presence of a highly variable workload. On the other hand, these personnel are limited in the number of hours they can work before patient safety becomes a concern. Medical equipment can also be a critical resource. Some equipment is expensive and must be employed at high utilization rates in order to be cost effective. However, shortages in availability can cause bottlenecks in the system, which lead to excess waiting times for patients. Ultimately, the goal is to maximize the utilization of medical personnel and equipment while maintaining a sufficient buffer to accommodate spikes in patient demands.

A substantial amount of work has been done using DES to model the delivery of healthcare in hospitals. DES is a natural extension to the traditional queuing theory models, which are popular analytic models for service-oriented problems. DES allows for the use of distributions of system parameters that might be intractable in a purely analytic model but better fit observed data. For a survey of DES in hospitals and healthcare clinics, see Jun et al. (1999). DES is the current standard, and some of the agent-based models presented have yet to take full advantage of their potential in providing unique insight into these types of problems. In addition, these preliminary models have the potential to address questions inaccessible to DES. ABM can generate heterogeneous patients that each have a unique set of medical needs. Patient arrival rates can be a function of time or other system parameters. Moreover, dynamic responses by the physicians and nurses in the system can be tested, allowing for them to respond to not only the needs of each patient but also the state of the entire system. In addition, ABM can represent many resource types, including medical equipment, to analyze the system efficiency under various patient demands and constraints. The following two sections deal with patient flow through the operating room and the emergency department, respectively. The third section examines the use of radio-frequency identification (RFID) technology in an emergency department.

## *2.1 Patient Flow*

With rising costs and an aging population, the demand for affordable healthcare is likely to increase significantly. In the USA, the Bureau of Labor Statistics predicts that the healthcare industry will generate more new jobs than any other industry between 2008 and 2018, but increasing the workforce is not enough and does little to help control expenses. Beyond increasing the capacity of the healthcare system, it is important that facilities are operated efficiently to minimize the idle time of operating rooms, imaging equipment, and healthcare workers.

Workflow problems are often modeled using queuing theory (see Chap. 1). Queuing theory facilitates analysis of complex systems that serve individuals with either stochastic arrival times, service times, or both. For a stationary arrival process, it is straightforward to determine the number of servers that can adequately meet the system demands. However, when the arrival process is less predictable, as it is in hospitals where patient arrival times typically vary by time of day and day of week, analytic solutions for these problems are not tractable. As a result, simulation becomes the best approach to evaluate potential solutions for improving the system throughput.

One example is a simulation model developed by Pearce et al. (2010), in which the authors seek to minimize delays in the first set of morning surgeries in an operating room (OR). They focus on delays at the start of a period because these delays can cause successively longer delays throughout the day. Patients scheduled for surgery are either inpatients or outpatients. All outpatients must first pass through registration and then receive some subset of tasks (e.g., triage, phlebotomy, etc.) before they join the inpatient track. Inpatients and processed outpatients see a combination of healthcare workers, which may include a patient care technician, surgeon, anesthesiologist, and certified registered nurse anesthetist before finishing their preoperative process. The tasks in both the initial outpatient track and the subsequent inpatient track may be subject to reordering in order to minimize wait times and maximize resource utilization. After finishing both tracks, the patient is then queued for the operating room.

The model developed by Pearce et al. treats patients, staff, and informational units (e.g., charts and laboratory results) as agents. Patients progress through four states: waiting room, preoperative room 1 (for the outpatient track), preoperative room 2 (for the inpatient track), and the OR itself. Hospital records are used to determine relative inpatient/outpatient loads. Data for the time each task requires was gathered by several days of shadowing patients through the process.

The expected length of preparatory time is a function of acuity level, with sicker patients taking more preparatory time on average than patients with lower acuity levels.

A variety of tests are run to improve the percentage of initial surgeries that began within 10 minutes of the scheduled time while keeping patient wait times as short as possible. Several scheduling strategies are evaluated. One strategy that was found to be effective is to schedule high-acuity patients and patients requiring blood work as early as possible because they have longer service times on average and more variance. The best improvement is achieved by signaling and coordinating the agents that needed to see the patient upon arrival to preoperative room 2. This strategy results in an increase of on-time surgeries from a baseline of about 60% to one of 85%. Signaling also allows the benefits of other changes to be more fully realized. For example, a policy of using the RN staff to perform triage without signaling results in only about a 1% improvement in on-time surgeries. Without signaling, patients under this policy spend less time in the waiting room but more time waiting in the preoperative stage. When signaling is used in addition to this policy, it allows for around an 89% on-time rate, about a 4% improvement over the base case with signaling. Future work will focus on resource utilization at each of the tasks for the inpatient and outpatient track.

## *2.2 Emergency Departments*

Emergency rooms have become a primary source of medical treatment for an increasing number of people. According to the CDC there were over 123 million visits to emergency departments in 2008, a 37% increase from 1996. Emergency rooms, more so than other parts of the hospital, have a highly variable workload because patients by their very nature are unscheduled. Therefore, one of the key problems is how to appropriately staff physicians and nurses in order to have a robust response to uncertain demand. Such a response, ideally, has neither excessive patient waiting nor idle physicians or nurses. Another key question is identifying bottlenecks in the system. For example, what should be the relative ratio of staffing of nurses to physicians? In addition to staffing, what procedures can be implemented to reduce waiting times? This might include steps taken within a single department, such as criteria for adding or subtracting triage nurses based on queue length. On a regional scale with multiple hospitals, how can ambulances and patients be diverted to prevent excessive congestion at any one hospital in the system?

Models of patient flow through an emergency department (ED) typically simulate a process similar to that in Fig. 3.1. The patients enter the system as either walk-ins or arrive by ambulance. Patients who arrive via ambulance circumvent the registration process and proceed directly to treatment as they have, in a sense, already been assigned a critical diagnosis. Walk-in patients first queue to register and then, after registration, proceed to queue for triage. At triage, a nurse will assign the patient a priority based on their emergency severity index (ESI). While the number of levels in an ESI need only be two or greater, the trend is toward a five-tiered

**Fig. 3.1** The basic workflow of patients through an emergency department. Note that queues may form prior to any rectangular station in the system

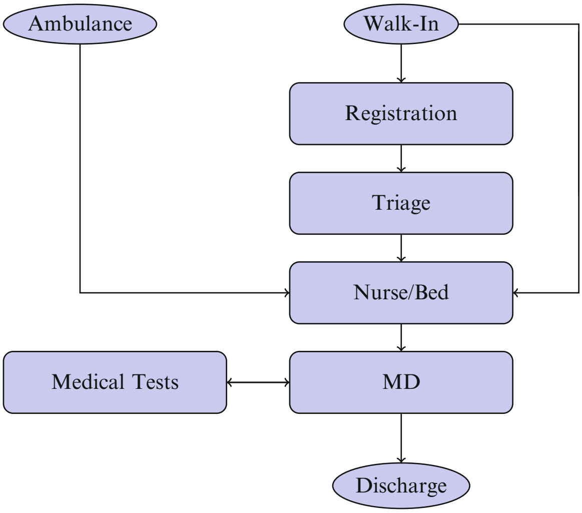


**Table 3.3** Summary of emergency department model parameters

| Model parameters | Kanagarajah et al. (2006) | Jones and Evans (2008) | Laskowski et al. (2009) | Wang (2009) |
|---|---|---|---|---|
| Software platform | Micro Saint Sharp | NetLogo | C++ | NetLogo |
| Adaptive agents | Yes | No | No | Yes |
| Medical tests | Yes | No | Yes | Yes |
| Ambulance arrivals | No | Yes | Yes | No |
| Triage levels | 3 | 5 | 3 | 1 |

system. ESI in a five-tiered system ranges from 1, typically labeled resuscitation which must be treated immediately, to 5 for nonurgent patients. Patients are then queued separately with those receiving the highest ESI typically being served first (as opposed to a purely FIFO system). A certain percentage of walk-in patients are treated as ambulance patients (in that they are visibly in such need that they skip the registration and triage processes and proceed directly to a room or bed to await physician care). After triage, patients queue for a room or a bed where they will then queue for treatment by a physician. Some models have a physician order tests for a percentage of patients, the results of which will be reviewed by a physician before discharging the patient. Patients can exit the ED at any step in the process if they leave against medical advice, leave before receiving treatment, or die. Patients complete the system after seeing a physician when they are either discharged from the hospital, transferred to another hospital, or admitted to the hospital. A summary of models using this basic framework can be seen in Table 3.3.

The Kanagarajah et al. model focused on balancing economic incentives, workload, and quality of care (Kanagarajah et al. 2006). These competing goals ultimately are a function of physician utilization and wait time. The simulation tests a variety of effects which included using on-call physicians to aid when arrival rates surged, adjusting the capacity of examination rooms where physicians serviced triaged patients, and having physicians spend less time with patients when lines were long. The amount of time spent with patients by a physician is an example of an adaptive behavior, which is unique to each individual physician and evolves over time. This adaptive behavior is a feature of ABM that cannot be easily replicated by other modeling techniques.

The Jones and Evans model, which allowed for a high degree of control in terms of inputs, was validated using data gathered from an urban hospital's ED (Jones and Evans 2008). Using a heterogeneous Poisson arrival process for patients, the authors observed how hourly and daily variation in patient arrival rates coupled with staffing levels affects patient waiting times. These distributions were gathered from hospital data and paired with actual staffing levels in order to validate their system's predicted wait times. In the model, the time required for a patient to be treated is an exogenous variable determined at the creation of the patient agent.

The Laskowski et al. model was used in the context of a single hospital to explore appropriate staffing levels for expected patient arrival rates (Laskowski and Mukhi 2009). The model, however, is also run in parallel to simulate multiple emergency rooms in order to explore policies that seek to distribute patient arrivals between hospitals in order to prevent excessive congestion in any one hospital in the system. The authors tested a policy known as random early detection, which reroutes ambulances with increasing probability as a function of the queue length of the intended destination. Reroutes are also applied to walk-in patients in the queue and can be thought of as either a suggestion that wait times might be shorter elsewhere or an offer for an ambulance to transfer them. This policy has the effect of greatly reducing congestion at any one hospital, essentially bounding the average queue length across hospitals and increasing physician utilization rate.

The Wang model focused on the total time spent in the system by patients (Wang 2009). The focus is on finding bottlenecks in the system that, if improved, could decrease total time spent in the system. The model explored the effect of adding triage nurses conditioned on the line length, then reducing staff once the line had reduced to one. The model also examined the system's sensitivity to radiology test times by having physicians order tests for a percentage of patients. The model which does not include patient acuity levels showed substantial improvements (30%) in total wait times by adding an extra triage nurse when the queue length reached ten patients. They also showed that small improvements in the average time spent in radiology (from 30 to 28 min) actually translated to much larger time savings, 10.38 min to all patients in the system.

The models described above define patient arrivals using a Poisson process, with Wang also having the option for a random uniform distribution for inter-arrival waiting times. The Jones and Evans model goes a step further in modeling patient arrival rates using a heterogeneous Poisson process with arrival rates varying each hour with daily and weekly cycles. All the models use empirical data to capture arrival rates and service times, drawing from either data available to them or data published from other ED studies. All models used between one day and one week of simulated time as a warm up for the system to enter an equilibrium state. One of the more novel applications of agent-based models in this area was the 2009 Laskowski and Mukhi paper, in which a method for using the model to test the effects of multiple staffing levels is described. The results from each simulation are input into a genetic algorithm that generates a new set of staffing levels to be tested.

Most of the models in the papers cited above were a first attempt to replicate more complicated real-world behavior and as such included simplifications that might be relaxed in future versions of the models. The Jones and Evans model, for example, showed that when using real-world staffing levels and variable arrival rates, their model simulated with a high degree of accuracy the wait time experienced by patients. Most of the papers stated a desire to better capture the many roles of nurses in the ED. Most of the workflow models still do not incorporate much in the way of spatial components. In addition, patient outcomes (as viewed by how they leave the ED) were independent of wait time. One can imagine that instead of this being exogenous to the system, it might be a function of wait time.

## 2.3 Modeling Radio-Frequency IDs and Electronic Health Records

RFID is a technology that allows for both passive and active tracking of goods or patients. RFID technology is being used to help implement a real-time location system (RTLS) for patients. This system allows for patients' progress through the ED to be tracked, allowing for more accurate record keeping and potentially shifting some of the burden of such record keeping away from nurses, especially when used in conjunction with electronic health records (EHR). Such systems, however, are not without their shortcomings. Active tags, which require batteries, have a greater range than passive tags, but can cost upwards of $50 per tag. Passive tags, which do not require batteries, can cost as little as five cents and are often preferred to active tags. Passive tags typically have a limited range of approximately one to two meters, with objects and other patients potentially impeding the signal further. If, however, passive readers are placed too densely in an area, there is the potential for interference between readers.

Laskowski et al. (2010) modeled the use of RFID technology in an RTLS implemented to track patients in an ED. Given the nature of the problem, the actual topology of the ED plays a much more important role than it did in the workflow models. The topology of the ED is in two dimensions with walls, which completely block RFID signals. Patients, nurses, and other agents that represent people are represented as circles with a 60 centimeter radius. A patient is thought to be at the last RFID reader triggered until another reader is triggered. Thus, when the patient is not within a reader's radius, but has not yet triggered another reader, there is an error in where the system believes the patient is located. The paper seeks to analyze the effect of passive RFID reader placement on system accuracy. The paper, which tests six reader configurations, runs each configuration 500 times to generate meaningful statistics (given the stochastic nature of arrival processes, this experimental design increases the likelihood that some high-volume days will be observed). There are two error terms calculated: the average Euclidean distance between a patient and the reader zone and the time a patient is reported as being in an incorrect zone. The paper then examines the effect of combining readers with variable ranges. The paper concludes that having too many readers leads to interference, which overwhelms any benefits from covering a larger area. The best performance is achieved by spacing readers as tightly as possible without overlap.

EHRs are another technology being integrated into many hospitals and emergency departments. The Poynton et al. model addresses a different question from the earlier ED workflow models, exploring how the spatial distribution of computer terminals for data entry affects workflow (Poynton et al. 2007). This model is unique in that between each stage of the patient's journey, the attending physician or nurse must enter the data from their most recent interaction with a patient. The model is clearly simplified because it assumes that caregivers are dedicated to a patient until they are discharged. The authors found that a policy based on having one computer per bed produces the best outcome. However, this strategy is often not feasible due to economic reasons; centralized computer clusters are preferable to computers distributed along the length of the hallway, depending on agent behavior.

### *2.4 Conclusions*

ABM is an intuitive framework to design and interpret models of healthcare delivery. Agents are well suited to navigate and interact with two- and three-dimensional spatial environments, as they were considered in the RFID model. Agents have the potential to learn and adapt their behavior as the situation changes. Adaptive behavior can represent different responses to different inputs, but also different behavior to the same inputs at different times, which allows for a richer modeling of policies. The scope of applications is quite broad, ranging from the scheduling of patients, medical diagnostics, and imaging equipment to the directing of ambulances to hospitals depending on urgency, specialty, and crowding.

## 3 Epidemiology

The field of epidemiology is primarily concerned with how disease spreads in a given population. Studies that model the transmission of infectious diseases can focus on increasingly large scales, ranging from single hospital units or wards to entire hospitals to communities, cities, and global pandemic scales. Vector-borne diseases, or those that are transmitted by way of an intermediate carrier, are the most commonly modeled types, but airborne, waterborne, food-borne, respiratory, and sexually transmitted diseases can also be modeled. These diseases can be modeled with different degrees of specificity to include incubation periods and periods where a colonized individual is infectious but asymptomatic. Understanding how diseases are transmitted and determining the best ways to control transmission are critical to preventing excessive spread, and epidemic modeling is an ideal method for experimenting with various strategies.

Epidemiological models have typically been compartmental models, which predict how proportions of several population states evolve over time using differential equations. The most well known of these models is the susceptible-infected-recovered (SIR) model (Kermack and McKendrick 1927), which laid the foundation for many other compartmental models (Austin and Anderson 1999; Austin et al. 1999; Beggs et al. 2008; Bootsma et al. 2006; Cooper et al. 2004; McBryde et al. 2007; Raboud et al. 2003; Robotham et al. 2007; Sebille et al. 1997). The SIR model equations are shown in (3.1), where $S$, $I$, and $R$ represent proportions of the population that are in the susceptible, infected, and recovered states, respectively. $\beta$ and $\gamma$ are the transmission and recovery rates. These equations can simply be integrated over time to generate population transmission dynamics and can provide interesting results for both deterministic and stochastic scenarios. An important measure for this type of model is the basic reproduction number, $R_0$, which is the average number of secondary infections (i.e., transmissions) per primary case in an entirely susceptible population. $R_0$ is a key metric in predicting the extent to which an infection is likely to spread. If $R_0 > 1$, then an epidemic is likely to grow because on average, each infected person transmits the disease to more than one other person. If $R_0 < 1$, then an epidemic is likely to become extinct. Some models use $R_0$ as an input to drive transmission within a population, whereas other models use a transmission rate or probability parameter and calculate the resultant $R_0$ value for the population based on the number of initially and secondarily infected individuals. These models have also provided a lot of insight into the effects of certain parameters on transmission dynamics and the effectiveness of various infection control measures as defined, for example, by the following differential equations:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI, \frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - \gamma I, \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I \tag{3.1}$$

Mathematical models such as the SIR model have several assumptions and limitations that prevent them from producing more valuable results. The first key assumption is that populations modeled by mathematical equations are well mixed, meaning that all individuals within the population interact with equal probability. For example, the SI term in the differential equations in (3.1) shows that transmission is proportional to the interaction between all susceptible and infected individuals, rather than a specific subset. All individuals within each compartment are assumed to be uniform, which prevents analysis of how mixed or extreme behavior, such as superspreaders or non-compliant healthcare workers (HCWs), can affect transmission dynamics. It can also be difficult to implement time-varying or conditional behavior in a mathematical model, and thus analysis of control measures is often performed by simply varying input parameters without modeling the interactions involved in that particular intervention. Some variables, such as hand-washing probabilities and screening test return times, are reasonable to approximate with simple parameters, but others such as isolation or the effect of staffing ratios are more difficult to simplify. Several advances in mathematical modeling have been made in recent years to account for these limitations, but the results are often only valid for particular applications (Bansal et al. 2007).

Agent-based, or computational, disease spread models have expanded on the research established by the mathematical models described above and have addressed many of their limitations. They have reinforced many conclusions from mathematical and simulation models and have provided additional detail about the nature of transmission. The key advantage of ABMs is that they simulate the interactions that serve as the primary mechanism for transmission and they are capable of implementing many infection control measures explicitly. In addition, ABMs are also more adept at simulating stochastic effects, which must be captured when modeling heterogeneous populations. As a result, they have contributed significantly to a better understanding of epidemics.

The articles reviewed in this section fall mainly into two categories: process-oriented models of transmission and network models of transmission. Process-oriented models typically simulate agents that are moving through a series of stages before being removed from the population, much like they would in a DES model. Network models simulate the spread of disease between individuals using relational connections as the primary mechanism of transmission. Intermediate carriers such as HCWs do not need to be modeled explicitly in this type of model because the connections between individuals are explicit, unlike in a process-oriented model. Exemplary methods and contributions for both types of models are summarized in the following subsections.

Pandemic modeling is another research area that has been affected positively by ABM. Diseases such as malaria, SARS, smallpox, and various strains of influenza are typically modeled for these applications. Control measures are most often concerned with logistics, such as distributing vaccinations, locating community clinics, delivering emergency rations and medical supplies, and evaluating the effects of social distancing (e.g., school closures). The initial work in applying ABM techniques to pandemic scenarios was done primarily by Carley et al. (2006) and Cummings et al. (2004), which both demonstrated the value of agent-based models in generating pandemic dynamics and evaluating response strategies. Since then, agent-based pandemic models have begun to incorporate massive data sets that reflect detailed demographic, social, transportation, and even climate characteristics of a particular geographic region. However, although many of the characteristics and advantages of these models are applicable to healthcare operations management, we do not provide additional coverage of this research area because these large-scale applications are more appropriately discussed within the context of public health.

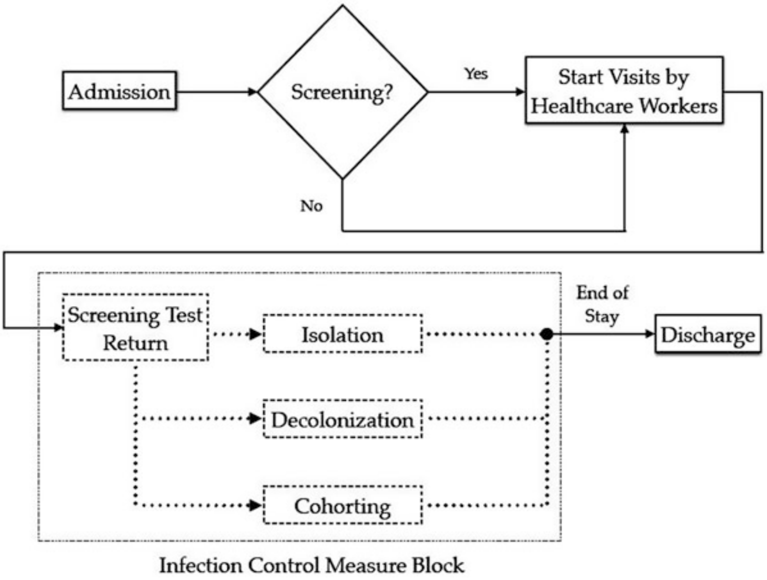


**Fig. 3.2** Sample patient flow diagram of disease spread model in a hospital. Optional infection control measures are indicated by *dashed lines* and *dotted arrows*

## *3.1 Process-Oriented Models of Transmission*

Process-oriented models of transmission are the most natural application of ABM to epidemiology. A common example is a simulation of patient-to-patient transmission in a hospital, in which patients are admitted, visited by HCWs, and discharged (see Fig. 3.2 for a sample patient flow diagram). What separates these models from traditional DES models is that there are often multiple patient and HCW types, and their behavior is often dynamic. Transmission typically occurs through HCWs, who spread an infection from one patient to another because they fail to wash their hands adequately. These types of transmissions, from an already infected (i.e., primary) patient to a newly infected (i.e., secondary) patient, are known as hospital-acquired, or nosocomial, infections. Hospital intensive care units (ICUs) are commonly modeled in an agent-based framework. ABM is ideally suited for this type of model because populations are small and diverse, patients are typically more susceptible to infection than in other hospital units, contacts between patients and HCWs are frequent and intimate, and stochastic effects are of considerable importance.

Within agent-based models, there is often an increased ability to track various simulation data and provide additional insight into the transmission dynamics. These models can evaluate the effectiveness of various infection control measures in order to offer recommendations of which measure or bundle of measures should be implemented. Typical control measures that are modeled include the hand-washing behavior of HCWs, diagnostic screening of patients, isolation of infected patients, vaccination, and decolonization, in which colonized patients undergo a therapeutic process that negates their ability to infect others.

There is a set of pathogens that are commonly modeled in these types of simulations, and many of these are resistant to antibiotic treatments. The most prevalent pathogens are methicillin-resistant Staphylococcus aureus (MRSA) and vancomycin-resistant enterococci (VRE), in which patients typically become colonized with the pathogen prior to developing an infection. This scenario is particularly difficult because colonized patients are often asymptomatic and, therefore, they can only become identified by using active surveillance techniques such as diagnostic screening. Consequently, these patients can spread the pathogen to HCWs and ultimately other patients before any intervention is initiated. In addition, treatment for these resistant organisms is often difficult, thus protecting patients from acquisition is the most effective approach for ensuring their safety.

The articles discussed in this section all follow a similar pattern for modeling transmission of an infectious disease in a hospital. They incorporate resistant pathogens and assess the potential effectiveness of specific control measures. However, these models differ slightly in their specific sets of experimental parameters and the nature of their results. These differences are summarized in Table 3.4.

In our review, we begin with two agent-based models that extensively explore the effects of external factors and infection control measures on disease transmission in a hospital. Both models highlight the specific level of detail afforded by ABM that is not possible using SD or DES methods. In the first example, Barnes et al. (2010) developed a detailed agent-based simulation of MRSA transmission in a hospital. The model incorporates many of the most common infection control measures and is able to reinforce the conclusions from mathematical models that one-to-one HCW-to-patient ratios and patient screening combined with isolation are most effective in reducing transmission (Austin and Anderson 1999; Austin et al. 1999; Bootsma et al. 2006; Cooper et al. 1999, 2004; McBryde et al. 2007; Raboud et al. 2003; Robotham et al. 2007). In addition, the authors highlight cases in which hospitals that have implemented infection control measures could still fall susceptible to MRSA outbreaks. Barnes et al. (2010) also conduct numerous simulations to explore the effects of individual HCW behavior on transmission, experimenting with the hand-washing probabilities of entire HCW classes (i.e., nurses and physicians) as well as rogue individuals that were less compliant than others in the same class. Many of these experiments are executed using parallel computing techniques, which alleviate the long run times caused by the model complexity. In the second example, Hotchkiss et al. (2005) also test the effects of several factors and infection control measures on transmission and demonstrate that early detection and subsequent isolation of infected patients, quick patient turnover, cohorting patients, and limiting the frequency of physician visits can all reduce the likelihood of a significant outbreak.

In addition to evaluating infection control measures and determining the most influential external factors, ABM can also provide more realistic transmission dynamics by introducing additional model complexity. This additional detail not

**Table 3.4** Comparison of parameters for several significant process-oriented agent-based models of infectious disease transmission

| Model parameters | Barnes et al. (2010) | D'Agata et al. (2007) | Hotchkiss et al. (2005) | Meng et al. (2010) | Ong et al. (2008) | Temime et al. (2010) |
|---|---|---|---|---|---|---|
| Software platform | Python | MATLAB | Mathematica | AnyLogic | PathoSim[a] | Java |
| Spatial representation | Yes | No | Yes | Yes | Yes | Yes |
| HCW types | 2 | 1 | 2 | 0 | 5 | 3 |
| HCW shifts | No | Yes | Yes | No | Yes | Yes |
| Variable hand hygiene | Yes | * | * | No | No | Yes |
| Hygiene efficacy | Yes | * | * | No | No | Yes |
| Patient screening | Yes | No | Yes | Yes | No | No |
| Patient isolation | Yes | No | Yes | Yes | No | No |
| Decolonization | Yes | Yes | No | Yes | No | No |
| Variable staffing ratios | Yes | No | Yes | No | Yes | Yes |
| Patient cohorting | Yes | Yes | Yes | No | No | Yes |
| Variable transmissibility | Yes | No | Yes | Yes | Yes | Yes |
| Patients colonized or infected on admission | Yes | No | Yes | Yes | No | Yes |
| Variable # of patient visits | Yes | Yes | Yes | No | Yes | Yes |
| Variable patient lengths of stay | Yes | Yes | Yes | Yes | No | No |
| Bacterial load | No | Yes | No | No | No | No |
| Antibiotic resistance | No | Yes | No | No | No | No |
| Visitors | Yes | No | No | No | Yes | No |

[a] http://www.ross-scientific.com/products.htm

*Variable hand hygiene and hygiene efficacy were not implemented explicitly, but transient HCW colonization times were variable and followed an exponential distribution only increases the relevancy and strength of simulation results, but it can be used to address the concerns of healthcare professionals who are skeptical of the model validity. Ong et al. (2008) implemented a spatially explicit ABM of influenza transmission in a hospital unit, where at any given time, each agent occupies a specific area in the unit. Several types of HCWs are modeled, including physicians, nurses, health attendants, clerks, and cleaners. Ambulant and non-ambulant patients are also modeled. Transmission of influenza is airborne; therefore, additional model considerations must be taken into account because agents can transmit the disease without coming into direct contact with each other. Results directly related to transmission are limited, but the model generates a distribution of contacts between all pairs of agents in the unit that could be incorporated into future modeling work. The model by Meng et al. (2010) incorporates multiple routes of transmission and variable transmission rates between patients, but it produces limited results related to actual dynamics. Temime et al. (2010) describe an agent-based model of pathogenic transmission in a hospital, but the implications of these results are more appropriately addressed in the following section on network models of transmission.

Agent-based models can also be used to address questions related to antibiotic resistance, which is an important issue in disease control. Several models address the implications of antibiotic resistance at the microbiological level and are not to be discussed. However, D'Agata et al. (2007) focused on the effects of antibiotic resistance on transmission and the competition between resistant pathogens in a hospital. Single patient and HCW types are modeled explicitly, with each having eligible states of being susceptible to or colonized with resistant and/or nonresistant pathogens. The authors demonstrate that initializing decolonization treatments on patients quickly and for shorter durations can eliminate both resistant and nonresistant strains from the population. The authors are also able to develop a corresponding differential equation model that facilitates model validation and additional analysis.

### *3.2 Network Models of Transmission*

Network models of transmission provide a different perspective than process-oriented models for analyzing the spread of infectious diseases. They are an abstract representation of the physical interactions that can lead to transmission, in contrast with the direct correlation to reality afforded by process-oriented transmission models. The edges, or connections, between nodes in the network can represent a relationship between patients in a hospital or between individuals in a community. The structure, or distribution, of these connections has a significant effect on transmission dynamics, and different network structures can be designed to represent various scenarios. Whereas process-oriented models can identify the best interventions, network models can provide insight as to where those interventions should be directed, such as targeting individuals for vaccinations or closing schools or hospital wards.

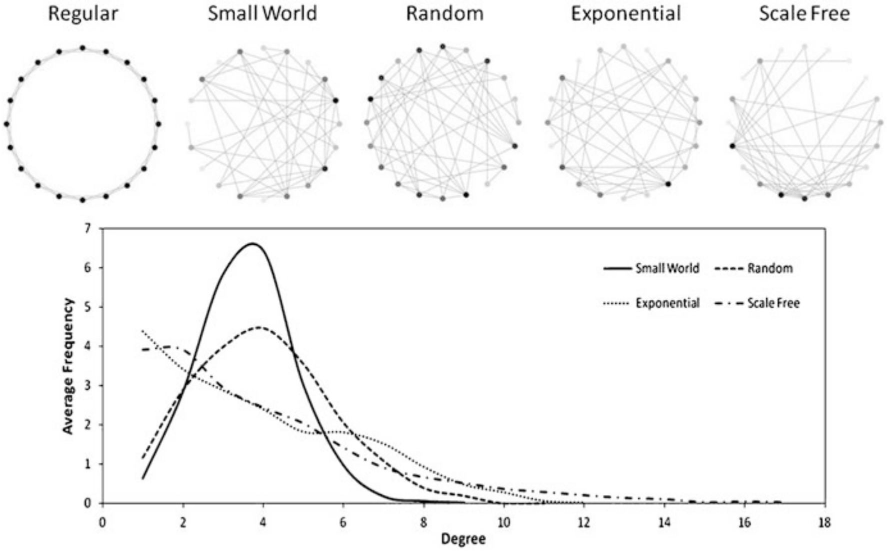


**Fig. 3.3** Sample network instances for regular (degree = 4), small-world (initial degree = 4, rewiring probability $p = 0.5$), random (edge probability = 0.2105), exponential (mean degree = 4), and scale-free (generated using the Barabási–Albert model for preferential attachment) structures. *Darker shaded nodes* have a higher degree relative to the other nodes, and *lighter shaded nodes* have a relatively lower degree. The *bottom plot* shows the mean degree distribution for each network type with 20 nodes and approximately 40 edges, averaged over 100 samples. The regular network degree distribution (not shown in the graph) has a constant degree distribution, with all 20 nodes having a constant degree of 4. As the network structure changes from regular to scale-free, the degree distribution becomes more skewed, with several nodes being highly connected and the remaining nodes having relatively few connections

There are several common structures for interaction networks, which are typically characterized by the frequency distribution of node connections in the network, also known as the degree distribution (Albert and Barabási 2002). Examples of each of the following network types and their corresponding degree distributions are shown in Fig. 3.3. Regular networks have nodes that all have the same degree, and edges can be structured (e.g., nodes are connected to their nearest *k* neighbors) or randomly distributed to other nodes in the network. Random networks are generated by assigning an equal probability to each potential edge in the network. Each edge is then chosen at random based on the given probability, which forms a network in which the degree distribution follows a binomial model. Small-world networks (Watts and Strogatz 1998) interpolate between regular and random networks by rewiring a certain proportion of edges in a structured, regular network. Nodes in this type of network are still highly clustered, but disease could spread more quickly because there are shortcuts to other highly susceptible sections of the network. A random network is a special case of a small-world network that has rewired all of its edges. A special class of networks are exponential networks, whose degree distribution follows an exponential trend. These networks have been found to be the most realistic structure for social interaction networks (Bansal et al. 2007). The last common type of network is a scale-free network. These networks have a power law degree distribution, in that there are a few nodes with a large number of connections and many nodes with a small number of connections. These networks are called scale-free because the average distance between any two nodes increases very slowly as the number of nodes increases. Disease transmission through these types of networks is likely to find the highly connected nodes quickly. However, transmission to the remaining population is likely to take much longer because there are fewer paths to nodes on the periphery of the network.

Mathematical models inherently assume that populations are well mixed, which means each individual has an equal probability of interacting with all other individuals. Interpretations can vary, but, in general, this assumption corresponds most closely to a regular network, whether the connections are structured or random. For certain applications, this configuration could be appropriate if each individual has approximately the same number of social contacts (e.g., child care centers). In other cases, a small-world, exponential, or scale-free network is a more accurate representation because there can be individuals that have connections to several population subgroups. These highly connected individuals often have the greatest effect on transmission, and neglecting to model them explicitly can have a significant effect on the results that are ultimately communicated to healthcare organizations.

The first set of models investigates how direct transmission can occur between individuals in a general population. These models typically consist of a homogeneous population of agents with essentially no individual characteristics other than their infection status. However, heterogeneity enters the model because the degree of each node in the network is not constant. The goal of these models is to characterize how the structure of the network affects the rate and extent of transmission, and to further demonstrate the limitations of homogeneous models.

Bansal et al. (2007) presented strong evidence that homogeneous mixing models, such as the SIR model, do not accurately predict epidemics for realistic contact networks. The authors are able to demonstrate that several empirical contact networks could all be approximated by synthetic networks with exponential degree distributions. Homogeneous mixing models, although reasonably accurate for characterizing transmission dynamics on regular, random networks, do not accurately predict epidemics on the more heterogeneous exponential and scale-free networks.

Christley (2005) focused instead on identifying the most susceptible individuals, or those most likely to become infected in the event of an outbreak, in random and small-world networks. This type of analysis is especially useful because the results could be used in developing strategies for targeting individuals for vaccination, isolation, or quarantine. The authors experiment with various measures of node centrality and determine that the degree of a node proved to be as good an indication of an individual's risk of infection as more complicated measures that would also require more information to compute. Eubank (2005) also proposed several local and global measures of network structure that could have significant implications for transmission of infectious diseases.

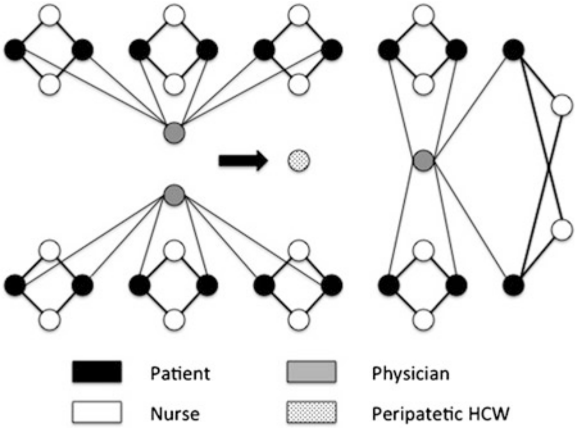


**Fig. 3.4** The network of contacts in the modeled ICU in the Temime model. There are 18 patients and 3 types of HCWs: 2 profiles of HCWs, corresponding to nurses and physicians, are assigned to subgroups of patients, and one peripatetic-type HCW visits every patient once each day

The next set of network models contain multiple types of agents that interact, whether they are explicitly or implicitly represented in the model. These studies are also focused on the structure of the network, but in addition they seek to characterize how interactions between different agent types affect transmission. The degree of heterogeneity in these network models facilitates analysis of the relative effect of each type of HCW. These models can also implement HCW behavior and bring consideration to other potential aspects of transmission such as HCW-to-HCW transmission and patient sharing. These types of interactions are not often considered, but can lead to increased levels of transmission in certain circumstances.

Temime et al. (2009) constructed a model of a hospital ICU with three types of HCWs that visit patients. Two types of HCWs are assigned to specific groups, or cohorts, of patients, whereas the third type visits all of the patients (see Fig. 3.4). The assigned HCWs represent nurses and physicians, and the third type, designated the peripatetic HCW, represents someone who could potentially come into contact with any patient, such as a nursing assistant or respiratory therapist. The model demonstrates the threat posed by the latter type and presented results that a single, noncompliant peripatetic HCW could cause the same level of transmission as if all HCWs were moderately noncompliant (i.e., 19–23% noncompliance). These effects become even more significant when HCW-to-HCW transmission occurs.

Barnes et al. (2012) also designed a network model of patient-to-patient transmission in a hospital. In this model, patients are connected directly if they share an HCW (see Fig. 3.5), which contrasts with the explicitly modeled intermediate HCW nodes in the Temime model. Nurses and physicians in this model are modeled implicitly, in that their states are only stored locally for agents in each respective cohort. By varying the number of patients, nurses, and physicians, networks are generated with different densities, calculated as the ratio of direct connections in the network to the maximum possible number of connections. Simulation experiments show that both nurses and physicians can pose threats to patients in different ways. Nurses visit patients more often, but physicians have the potential to infect patients

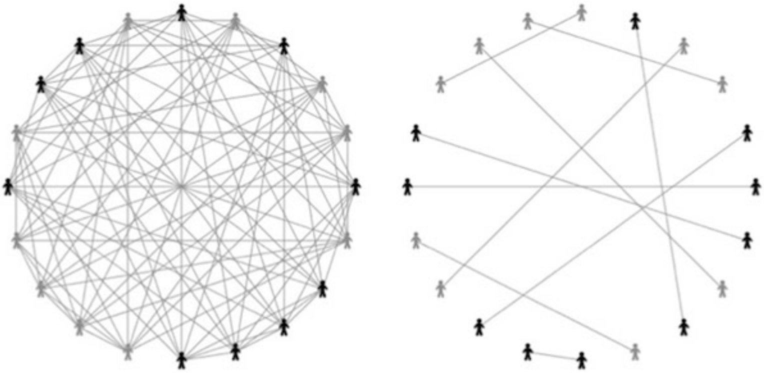


**Fig. 3.5** An example of a dense (*left*) and sparse (*right*) patient network from the Barnes, Golden, and Wasil model. Patients that share a nurse are connected by a link, while patients that share a physician have the same color

in different locations in the network, similar to the threat posed by Temime's peripatetic HCW, albeit to a lesser degree. Barnes et al. also experiment with patient sharing, demonstrating that the sharing of patients between HCWs should be performed in a structured, and not random, manner.

## *3.3 Conclusion*

Overall, ABMs are well suited for testing potential infection control measures and can provide some indication of success before implementing any particular strategy. Among ABMs, process-oriented models have a more established record of contribution, and they have produced relatively widely accepted results concerning the effectiveness of common infection control measures. The foundation has been established for network models of transmission, but there are still many areas to explore, particularly in determining which network measures are the most effective in predicting an outbreak. However, the benefit of modeling the explicit connections between individuals has been demonstrated in the unique results from this type of model. Both static and dynamic network measures may be useful, but the ease of calculation may be critical for practical use.

There are several key results that ABMs have reinforced, and several more that are unique to this methodology. The hand-hygiene compliance of HCWs is critical to preventing outbreaks of infectious diseases. In most cases, however, hand washing is not sufficient to control transmission, and additional measures such as active surveillance, patient isolation, or higher staffing ratios become necessary. Minimizing patient lengths of stay and HCW visits to patients can also reduce the risk of secondary infections, as susceptible patients are exposed less to potential infections and infected patients have fewer opportunities to infect HCWs.

ABMs have also provided insight into the relative threats posed by HCWs that care for patients in a hospital. Nurses and other HCWs that visit patients frequently are at a high risk of becoming at least transiently colonized or infected and can quickly spread disease to other patients in their care. Physicians and other HCWs that may visit many more patients than nurses can potentially create multiple pockets of infection that could lead to an entire unit becoming infected. Maintaining high staff-to-patient ratios and assigning patients to HCWs in a structured way can offset these dangers. Future studies may be able to suggest methods for identifying individuals at a high risk of infection because of their location in a network, and appropriate measures could be taken to prevent that person from becoming infected before isolation or quarantine measures become necessary.

We focused here on general models with implications for potentially many applications, but there are several models that have been applied to empirical networks (Eubank 2005; Meyers et al. 2005). These models have also achieved results that could have a significant impact, for not only their intended scenario but others. They are also good examples of the process involved in modeling real-world scenarios and can provide a good framework for future applications.

## 4 Healthcare Economics and Policy

In addition to modeling individual point-of-care facilities, ABMs can be used to analyze large populations at local, regional, or national levels. The interactions between different facilities are often difficult to observe from the perspective of any individual facility and are often difficult to predict. System behavior is often regulated by policies that are set in place by local, state, or federal governments. Policy changes are not usually enacted one at a time, therefore it is difficult to predict and attribute changes to a single cause when multiple changes are made. Important policy decisions should be informed using the best available data and analysis, which now includes modeling.

### *4.1 Healthcare Economics*

In their paper, entitled "Modeling Healthcare Policy Alternatives," Ringel et al. (2010) address the benefits and limits of using agent-based (microsimulation) models to anticipate the effects of policy changes in health insurance markets. In particular, they address several issues related to the data used to populate the agents in the model and describe the relations between them and the difficulties in defining a choice model for agent decisions. The relevant models (Congressional Budget Office 2007; Girosi et al. 2009; Garret et al. 2008) seek to represent the national population in terms of households and their relationship to employers as well as model their decisions on purchasing health insurance with respect to price, availability, and need. The scope of these models is at the national level and consists of at least three types of agents including households, employers, and insurers.

The naive approach to populating a model would be to assume independence between distributions of attributes. This assumption, however, ignores important correlations which occur in actual populations that might have a significant effect on the validity of the model, especially when dealing with socioeconomic data. Detailed information is available through the US Census and similar sources. However, simple demographic data is often not sufficient. The data sets used by each of the models contained additional information needed to provide a more detailed picture. The Survey of Income and Program Participation (SIPP), which is conducted by the US Census Bureau, was one choice used to construct a base population (Survey of Income and Program Participation SIPP). SIPP includes information on household income, taxes, and participation in government programs. Data sets such as these can also be projected onto or merged with other data sets to adjust for changes in the sample population. Other models used the Current Population Survey (CPS), which is conducted monthly by the US Census to track employment, occupation, earnings, and employee benefits (Current Population Survey CPS). The Medical Expenditure Panel Survey (MEPS), conducted by the agency for healthcare research and quality, provides data on household consumption of healthcare, including information on cost, use of healthcare, and health insurance coverage (Medical Expenditure Panel Survey MEPS). In addition to having a representative population, one must ensure that the relationships between agents are accurately described. In this case, that means an accurate relationship between firms (employers) and households (employees). This data is rarely available from the same source as the previously mentioned data. The CPS data, however, includes information on occupation. This information could be used along with a survey of firm size and industry, as well as geographic data, in order to help recreate a feasible set of firms to employ the households.

Decisions can be modeled either deterministically or probabilistically. When presented with a specific set of inputs, an agent using a deterministic decision model will always make the same decision, whereas probabilistic choice models will assign some probability to each possible choice. When given a history of agent decisions, such rules can be estimated by using a number of techniques. Decision tree learning determines a series of observations that can best recreate the choice results of the observed data. Logistic models, such as the multinomial logit model, can be used to assign probabilities to possible outcomes. The accuracy of such models is limited, however, by the amount and variety of data on previous choices. In addition, such models can break down when presented with conditions outside anything observed in the training data.

With an accurate population and sensible agent behavior, validation begins by first recreating a known set of conditions, before applying any policy changes. The Ringel et al. paper goes on to discuss potential improvements to available data and behavioral parameter estimates. These models, while primarily economic in nature, give detailed accounts and justifications for their choices in defining their agents. These are choices that all researchers will face when attempting to build large-scale models with incomplete population and choice data.

## *4.2 Healthcare Policy*

In addition to economic considerations, healthcare policy decisions, can have a direct or indirect impact on patient outcomes. When presented with a number of policy options, it is important to identify who will be impacted and how they will be affected. MIDAS, the Models of Infectious Disease Agent Study, is a collaborative research initiative with funding through the National Institutes of Health to model the spread of disease using agent-based models. Building on previous MIDAS models, the paper by Lee et al. (2010) uses an agent-based model to analyze the policy question of how to best prioritize, allocate, and ration vaccines during a simulated flu pandemic.

Current recommendations by the Advisory Committee on Immunization Practices (ACIP) in the USA, as stated in the paper, suggest that pregnant women, caregivers for children younger than 6 months, healthcare and emergency medical personnel, everyone between the ages of 6 months and 24 years, and people between the ages of 25 and 64 who have medical conditions associated with higher risk of complications from influenza should be prioritized to receive flu shots. These groups are listed either because they are at higher risk for complications with influenza or because they are more likely to spread influenza if infected, particularly to vulnerable populations. The paper seeks to address how strictly these guidelines should be applied in prioritizing the distribution of vaccines and whether a particular subgroup should be prioritized among the recommended groups.

The model itself is concerned with the Washington DC metropolitan area and has a population of over seven million agents, whose demographic data is drawn from the US Census Bureau's Public Use Microdata files (PUM) (Public-Use Microdata Samples PUMS). PUM files, in addition to including detailed demographic data, track transportation usage and are well suited for modeling small geographic areas. Agents interact with other agents within their household and with either schoolmates or coworkers, depending on their age. In addition, each agent is assigned a generic activity level, which varies with the day of the week. The transmission and mortality rates were calculated from previous flu pandemics. A specified percentage of infected agents require hospitalization and other medical treatments, which served as part of the basis used to calculate the total cost of the pandemic.

A series of trials using the model compared scenarios including relaxing the distribution of the vaccine to include some healthy adults between 25 and 64, prioritizing different at-risk groups, and changing the bounds of the age recommendations. The authors found that while prioritizing those what were more likely to spread the disease did reduce the overall number of infected agents, the population also had higher total morbidity, because more at-risk patients were vulnerable to getting sick. The results as a whole support the use of the ACIP guidelines in selecting who should be vaccinated.

### *4.3 Conclusion*

Important policy issues will continue to arise as legislatures amend existing regulations and prepare for emergency situations. It is important that detailed models be constructed to help understand the far-reaching effects that proposed changes might cause. Enacting policies such as the vaccine prioritization involves making trade-offs between the total number of infected people and the outcomes of those most vulnerable to influenza. Detailed modeling can help to illustrate the implications of the trade-offs involved with policy changes.

## 5 Open Opportunities for Research and Practice

ABM and simulation has been applied successfully to several applications of healthcare operations management, including healthcare delivery, epidemiology, economics, and policy. Building on the foundation established by system dynamics and DES, ABM has provided new insight to these problems by modeling individuals and the interactions between them. This perspective has facilitated analysis at both the individual and system levels, which is not typically possible using other methods. The greatest value of ABMs is that they can be used as virtual environments to evaluate policy alternatives, some of which would be infeasible or unethical to experiment with in practice. This capability can help healthcare organizations make better decisions, which can potentially lead to a higher quality of care for patients and extensive cost savings.

As with the other modeling methodologies, the challenge in constructing ABMs is first generating realistic dynamics for a particular scenario and then analyzing those patterns to identify the best strategies for intervention. Determining the appropriate level of detail for an agent-based model can be a difficult challenge. Extremely detailed models require many parameters, and it is difficult to get accurate values for these parameters from empirical data and the research literature. Parameters are often set at reasonable values, and sensitivity analysis is performed to address any shortcomings associated with poor selection, but this type of analysis can become very time-consuming if model run times are large. Models with few parameters mostly avoid this problem, but are sometimes challenged due to their lack of detail and complexity. One solution to this issue is to use hybrid models, which incorporate system dynamics and agent-based methods. In a hybrid model, one might model a network of agents whose states are affected by both their interaction with other agents and an internal differential equations model. These models can take advantage of the strengths of each methodology, most importantly the lower computational demands of system dynamics and the heterogeneity afforded by ABM. For all levels of fidelity, agent-based models can incorporate and generate large quantities of data. As a consequence, statistical rigor, efficient data analysis techniques, and visualization are all critical to producing insightful results and communicating those findings to healthcare professionals.

Several models focused on only a particular hospital configuration or network instance, which may limit the usefulness of a particular strategy. Certain trends may persist over similar variations, but the strongest impact is likely to be derived from models that suggest robust solutions that have implications for many scenarios. Other models used parameterized configurations, which are more broadly applicable, but may not correspond well to any particular situation. ABMs can also operate on a variety of time scales, and it is important to choose an appropriate scale that is relevant to the problem and compatible with model parameters.

Another challenge with ABMs is related to validation. The level of validation required for a particular agent-based model is primarily driven by its ultimate purpose. Models that are intended to produce accurate quantitative (i.e., predictive) results may require extensive validation, whereas more qualitative (i.e., illustrative) models have less stringent requirements. Validation is difficult for some applications because they lack empirical data as a baseline for comparison. Therefore, it is not always possible to know what simulation outcomes should look like. When empirical data are available, there are several approaches for using them to validate a model (Fagiolo et al. 2007). The best of these approaches primarily involves incorporating empirical data, historical perspective, and subject matter expertise in the selection of the best possible set of model assumptions, parameter settings, and initial conditions. By combining these data sources, agent-based models are more likely to use appropriate values for input parameters and generate reasonable system responses. For any validation approach, it is always necessary to perform sensitivity analyses to identify any parameters that cause nonproportional changes in the system.

Empirical data is typically available for many applications within healthcare operations, especially those related to patient information and resource management. In addition, subject matter expertise is readily available, which should be leveraged to develop models with an appropriate set of agent characteristics and behaviors. However, validation becomes more difficult when the underlying dynamics are not well understood or there is no historical reference, as is generally the case for the transmission of antibiotic-resistant pathogens or the implementation of a novel healthcare policy. For these cases, an agent-based model may have more value in demonstrating relative trends or generating alternative outcomes for comparison, and less value as a predictive model.

Moving forward, new and existing agent-based models should continue to build on the most relevant achievements of other models, including those from other fields that have taken advantage of the methodology. Making these models widely available is a practice that can help to accelerate progress. This evolutionary process is another advantage of ABM, as system dynamics and DES models are more difficult to augment with additional complexity. Each individual model may have limitations, but as the research grows, many of those limitations can eventually be eliminated, thereby increasing the acceptance of these methods by healthcare organizations.

## References

Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1): 47–97

Austin DJ, Anderson RM (1999) Studies of antibiotic resistance within the patient, hospitals and the community using simple mathematical models. Phil Trans Roy Soc Lond B 354(1384): 721–738

Austin DJ, Bonten MJ, Weinstein RA et al (1999) Vancomycin-resistant enterococci in intensive-care hospital settings: Transmission dynamics, persistence, and the impact of infection control programs. Proc Natl Acad Sci USA 96(12):6908–6913

Bansal S, Grenfell BT, Meyers LA (2007) When individual behaviour matters: Homogeneous and network models in epidemiology. J Roy Soc Interface 4(16):879–891

Barnes S, Golden B, Wasil E (2010) MRSA transmission reduction using agent-based modeling and simulation. INFORMS J Comput 22(4):635–646

Barnes S, Golden B, Wasil E (2012) Exploring the effects of network structure and healthcare worker behavior on the transmission of hospital-acquired infections. IIE Tran Healthc Syst Eng 2:259–273

Beggs CB, Shepherd SJ, Kerr KG (2008) Increasing the frequency of hand washing by healthcare workers does not lead to commensurate reductions in staphylococcal infection in a hospital ward. BMC Infect Dis 11:1–11

Bootsma MCJ, Diekmann O, Bonten MJM (2006) Controlling methicillin-resistant Staphylococcus aureus: Quantifying the effects of interventions and rapid diagnostic testing. Proc Natl Acad Sci USA 103(14):5620–5625

Carley KM, Fridsma D, Casman E, Yahja A, Altman N, Chen L-C, Kaminsky B, Nave D (2006) BioWar: Scalable agent-based model of bioattacks. IEEE Trans Syst Man Cybern A 36(2): 252–265

Christley RM (2005) Infection in social networks: Using network analysis to identify high-risk individuals. Am J Epidemiol 162(10):1024–1031

Congressional Budget Office (2007) CBO’s Health insurance simulation model: A technical description. www.cbo.gov/ftpdocs/87xx/doc8712/10-31-HealthInsurModel.pdf. Accessed on August 2011

Cooper BS, Medley GF, Scott GM (1999) Preliminary analysis of the transmission dynamics of nosocomial infections: Stochastic and management effects. J Hosp Infect 43(2):131–147

Cooper BS, Medley GF, Stone SP et al (2004) Methicillin-resistant Staphylococcus aureus in hospitals and the community: Stealth dynamics and control catastrophes. Proc Natl Acad Sci USA 101(27):10223–10228

Cummings D, Burke DS, Epstein JM, Singa RM, Chakravarty S (2004) Toward a containment strategy for smallpox bioterror: An individual-based computational approach. Brookings Inst Pr, Washington, DC, pp 1–55

Current Population Survey (CPS). www.census.gov/cps/. Accessed on August 2011

D’Agata EMC, Magal P, Olivier D et al (2007) Modeling antibiotic resistance in hospitals: The impact of minimizing treatment duration. J Theor Biol 249(3):487–499

Eubank S (2005) Network based models of infectious disease spread. Jpn J Infect Dis 58(6): S9–S13

Fagiolo G, Moneta A, Windrum P (2007) A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems. Comput Econ 30(3):195–226

Fone D, Hollinghurst S, Temple M et al (2003) Systematic review of the use and value of computer simulation modelling in population health and health care delivery. J Publ Health 25(4): 325–335

Garret B, Clemans-Cope L, Bovbjerg R, Masi P (2008) The Urban institute's microsimulation model for reinsurance. www.urban.org/url.cfm?ID=411690. Accessed on August 2011

Girosi F, Cordova A, Eibner C, Gresenz C, Keeler E, Ringel J, Sullivan J, Bertko J, Buntin M, Vardavas R (2009) Overview of the COMPARE microsimulation model. www.rand.org/pubs/working_papers/WR650. Accessed on August 2011

Hotchkiss JR, Strike DG, Simonson DA et al (2005) An agent-based and spatially explicit model of pathogen dissemination in the intensive care unit. Crit Care Med 33(1):168–176

Jones S, Evans R (2008) An agent based simulation tool for scheduling emergency department physicians. AMIA Annu Symp Proc 2008:338–342. Published online 2008. PMCID: PMC2656074

Jun JB, Jacobson SH, Swisher JR (1999) Application of discrete-event simulation in health care clinics: A survey. J Oper Res Soc 50(2):109–123

Kanagarajah A, Lindsay P, Miller A, Parker D (2006) An exploration into the uses of agent based modeling to improve quality of health care. In: Minai A, Braha D, Bar-Yam Y (eds) Proceedings of the 6th international conference on complex systems, Boston, MA

Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc Roy Soc Lond A 115:700–721

Koelling P, Schwandt MJ (2005) Health systems: A dynamics system benefits from system dynamics. In: Kuhl ME, Steiger NM, Armstrong FB, Joines JA (eds) Proceedings of the 2005 Winter Simulation Conference, Orlando, FL, USA, pp 1321–1327

Laskowski M, Mukhi S (2009) Agent-based simulation of emergency departments with patient diversion. Electronic healthcare. Springer, Berlin, pp 25–37. Isbn: 978-3-642-00413-1

Laskowski M, McLeod RD, Friesen MR, Podaima BW, Alfa AS (2009) Models of emergency departments for reducing patient waiting times. PLoS ONE 4(7):e6127. doi:10.1371/journal.pone.0006127

Laskowski M, Demianyki B et al (2010) Uncertainties in RFID tracking systems in an emergency department. Health Care Management (WHCM), 2010 IEEE Workshop

Lee B, Brown S, Korch G, Cooley P, Zimmerman R, Wheaton W, Zimmer S, Grefenstette J, Bailey R, Assi T, Burke D (2010) A computer simulation of vaccine prioritization, allocation, and rationing during the 2009 H1N1 influenza pandemic. Vaccine 28(31):4875–4879

Macal C, North M (2007a) Managing business complexity: Discovering strategic solutions with agent-based modeling and simulation, 1st edn. Oxford University Press, New York

Macal CM, North MJ (2007b) Agent-based modeling and simulation: Desktop ABMS. In: Henderson SG, Biller B, Hsieh M-H, Shortle J, Tew JD, Barton RR (eds) INFORMS Winter Simulation Conference, Washington, DC, USA, pp 95–106

McBryde ES, Pettitt AN, McElwain DLS (2007) A stochastic mathematical model of methicillin resistant Staphylococcus aureus transmission in an intensive care unit: Predicting the impact of interventions. J Theor Biol 245(3):470–481

Medical Expenditure Panel Survey (MEPS). www.meps.ahrq.gov/mepsweb/. Accessed on August 2011

Meng Y, Davies R, Hardy K, Hawkey P (2010) An application of agent-based simulation to the management of hospital-acquired infection. J Simul 4(1):60–67

Meyers LA, Pourbohloul B, Newman MEJ, Skowronski DM, Brunham RC (2005) Network theory and SARS: Predicting outbreak diversity. J Theor Biol 232(1):71–81

Ong B, Chen M, Lee V, Tay J (2008) An individual-based model of influenza in nosocomial environments. In: Bubak M, van Albada GD, Dongarra J, Sloot PMA (eds) Comput Sci, Int Conf Comput Sci 2008, Part I. Lecture Notes in Comp Sci 5101:590–599

Pearce B, Huynh N, Harris S (2010) Modeling interruptions and patient flow in a preoperative hospital environment. Proceedings of the 2010 Winter Simulation Conference Baltimore, MD, USA

Poynton M, Shah V, BeLue R, Mazzotta B, Beil H, Habibullah S (2007) Computer terminal placement and workflow in an emergency department: An agent-based model. Proceedings of the complex systems summer school, Winter Simulation Conference Washington, DC, USA

Public-Use Microdata Samples (PUMS). www.census.gov/main/www/pums.html. Accessed on August 2011

Raboud J, Saskin R, Simor A et al (2003) Modeling transmission of methicillin resistant Staphylococcus aureus among patients admitted to a hospital. Infect Contr Hosp Epidemiol 26(7):607–615

Ringel J, Eibner C, Girosi F, Cordova A, McGlynn E (2010) Modeling health care policy alternatives. Health Serv Res 45:1541–1558

Robotham JV, Jenkins DR, Medley GF (2007) Screening strategies in surveillance and control of methicillin-resistant Staphylococcus aureus (MRSA). Epidemiol Infect 135(2):328–342

Sebille V, Chevret S, Valleron JA (1997) Modeling the spread of resistant nosocomial infections in an intensive-care unit. Infect Contr Hosp Epdemiol 18(2):84–92.

Survey of Income and Program Participation (SIPP). www.census.gov/sipp/. Accessed on August 2011

Temime L, Opatowski L, Pannet Y et al (2009) Peripatetic health-care workers as potential superspreaders. Proc Natl Acad Sci USA 106(43):18420-5

Temime L, Kardas-Sloma L, Opatowski L et al (2010) NosoSim: An agent-based model of nosocomial pathogens circulation in hospitals. Proc Comput Sci 1(1):2245–2252

Wang L (2009) An agent-based simulation for workflow in emergency department. In: Proceedings of the 2009 IEEE systems and information engineering design symposium, University of Virginia, Charlottesville, VA

Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nat 393(6684): 440–442

# Chapter 4
# Optimization in Healthcare Delivery Modeling: Methods and Applications

**Sakine Batun and Mehmet A. Begen**

## 1 Introduction

Healthcare expenditures in many countries account for a significant portion of their gross domestic product (e.g., 17.4% in the USA, 11.4% in Canada, and 9.8% in the UK, according to the publicly available data for 2009 (OPTN 2011)). Besides suffering from high costs, healthcare delivery systems are also experiencing serious capacity-related problems since the world population is rapidly aging (the number of people over age 60 is expected to reach 22% by 2050 (WHO 2012)), and hence, the number of people that will be in need of care is quickly increasing. Due to the high costs involved in healthcare delivery and the forecasted increase in the demand, efficient allocation of healthcare resources has become more crucial.

Decision making in healthcare is a complex and critical process, and an attractive application area for operations research. Managing healthcare involves planning and coordinating several scarce resources (most of which are expensive and highly specialized) and considering multiple stakeholders often with conflicting goals. Moreover, there exists a significant amount of demand and service time uncertainty in the nature of healthcare delivery, which brings additional complexity. Improving the efficiency in healthcare systems has the potential not only to decrease costs but also to facilitate faster access to care. Therefore, decision making in this area plays a critical role both from the economic and the societal perspective.

Optimization is one of the widely used operations research methodologies in modeling and solving healthcare operations management problems. In this chapter, we present the optimization-based studies that consider a decision-making problem in healthcare delivery, such as appointment scheduling, operating room scheduling, capacity planning, workforce scheduling, and some other practical problems. Due to the vast amount of literature, we focus our review on recent studies and refer

S. Batun • M.A. Begen (✉)
Richard Ivey School of Business, University of Western Ontario, ON, Canada
e-mail: sbatun@uwo.ca; mbegen@ivey.uwo.ca

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_4, © Springer Science+Business Media New York 2013

the reader to comprehensive reviews for more detailed information on earlier work. We put special emphasis on appointment scheduling, operating room scheduling, capacity planning in the presence of patient classes with different priorities, and workforce scheduling. For these specific applications, we present detailed examples that illustrate the use of optimization methods, particularly discrete convex analysis, stochastic programming, and approximate dynamic programming.

The remainder of this chapter is organized as follows. In Sect. 2, we review a number of recent studies on appointment scheduling and present different modeling approaches for the single-server appointment scheduling problem. In Sect. 3, we discuss the use of optimization methods in operating room scheduling by providing several examples, and we illustrate the impact of using valid inequalities in increasing the solvability of the problems in this field. In Sect. 4, we present a summary of some recent studies that applied optimization methods to capacity planning problems in healthcare, and we provide a detailed example to demonstrate the use of approximate dynamic programming in modeling and solving the dynamic multipriority patient scheduling. In Sect. 5, we discuss how optimization methods can be used to solve workforce scheduling problems in healthcare and present a detailed example about the use of stochastic programming in modeling an operational-level nurse scheduling problem. In Sect. 6, we briefly discuss a number of studies that utilize optimization methods in some other application areas in healthcare. Finally, in Sect. 7, we conclude the chapter by highlighting some open research problems.

## 2 Appointment Scheduling

Appointment scheduling, which is an important operational problem in many industries (e.g., transportation, manufacturing, service), is encountered in several settings in healthcare delivery systems. Examples include, but are not limited to, scheduling surgeries in an operating room (OR), determining appointment times for a physician at an outpatient clinic, and scheduling arrivals to diagnostic or treatment facilities. Cayirli and Veral (2003) and Gupta and Denton (2008) present detailed reviews of appointment scheduling in various healthcare contexts.

An appointment schedule is composed of a planned start time (i.e., appointment time) for each service at which the required resources will become available. Due to the uncertainty in service times, which is the main challenge in designing optimal appointment schedules, the relevant performance measures (such as patient waiting time, resource idle time, and overtime) are uncertain. Thus, the objective of designing an appointment schedule is often to minimize the expected cost associated with one or a combination of those measures.

Appointment scheduling problems have been studied by several researchers since 1950s, with the earlier work being focused on single-server queueing models. Weiss (1990) considers the problem of determining optimal start times for two surgeries to be processed in a single OR, where the sequence of surgeries is determined a priori and the objective is minimizing resource idle time (of the first surgery) and surgery waiting time (of the second surgery). For this special case (i.e., determining duration allocation only for the first job), he shows that the problem is equivalent to the newsvendor problem, and hence, the solution can be obtained by using a closed-form equation. Weiss (1990) also investigates the impact of surgery sequence, by comparing the minimum expected costs incurred under possible sequences, and concludes that the surgery whose duration follows a distribution with fatter tails has more uncertainty at higher values, and therefore should be scheduled last to achieve lower expected costs.

Wang (1993) studies the single-server appointment scheduling problem where the service times are independent and identical exponentially distributed random variables, and the performance measure is a linear combination of the expected total customer delay time (i.e., the sum of the waiting time and the service time) and the service completion time (i.e., the completion time of the last job). The author shows that the distribution of customer delay times is phase type, which brings computational advantages and makes it possible to find the optimal appointment times by solving a system of nonlinear equations. The numerical results presented by Wang (1993) reveal that optimal interarrival times are dome shaped, i.e., allocated durations exhibit an increasing pattern earlier in the schedule and a decreasing pattern later in the schedule.

Denton and Gupta (2003) and Begen and Queyranne (2011) consider the single-server appointment scheduling problem where the sequence of customers is given and propose efficient solution methods that can be used to solve larger instances. Denton and Gupta (2003) formulate the problem as a two-stage *stochastic linear program* (SLP) and solve it by using a modified L-shaped algorithm where the space of the random job durations is successively partitioned and upper bounds independent of the distribution of job durations are employed. Begen and Queyranne (2011) study the discrete version of the problem; they establish discrete convexity properties of the objective function under mild conditions on the cost parameters and prove the existence of a polynomial time algorithm for the case where job durations are independent integer-valued random variables given by a discrete probability function. They also consider no-shows, emergencies, and a fixed (given) due date on the service completion time of the last patient. We provide the details of both studies in Sect. 2.1.

Erdogan and Denton (2011b) explore the single-server appointment scheduling problem where both the service times and the number of customers to be served are uncertain, and the objective is to minimize the expected cost of overtime and customer waiting time. They consider two different settings: (1) a static problem where an appointment schedule is designed for a fixed number of customers each of whom has a probability of not showing up at the scheduled appointment time and (2) a dynamic problem where the number of customers to be scheduled is uncertain, and the appointment time of each customer is determined when the appointment is requested. The authors formulate the first problem as a two-stage SLP and solve it by using the L-shaped algorithm. In line with the results from earlier studies (Wang 1993; Denton and Gupta 2003), their numerical results also indicate that the optimal interarrival times are dome shaped when the difference between per unit time costs of overtime and customer waiting time is small. When the probability of no-shows increases, the allocated durations decrease to avoid the high idling that may result from longer interarrival times. When the per unit time cost of overtime becomes significantly higher than that of customer waiting time, double bookings are observed in the optimal appointment schedule. The authors formulate the dynamic appointment scheduling problem as a multistage SLP where the decision stages are defined by customer appointment requests. At each stage, the requested appointment is scheduled by determining the corresponding interarrival time. The authors exploit the structural properties of their model and employ the nested decomposition algorithm to solve instances generated based on real data. Their numerical results reveal that the presence of uncertainty in the number of customers to be scheduled results in an increase and a decrease of the interarrival times earlier and later in the schedule, respectively.

## 2.1 Single-Server Appointment Scheduling

In this section, we present three different modeling approaches from the literature for the single-server appointment scheduling problem that arises in many contexts in healthcare. Since the appointment scheduling problem is also encountered in several other settings, we use the generic terms *job* and *server* to refer to health services and healthcare providers/resources, respectively, in the remainder of this section.

### 2.1.1 Discrete Time Modeling Approach

Begen and Queyranne (2011) consider the problem of designing an optimal appointment schedule for a given sequence of jobs on a single server, where the job durations are uncertain and the system is subject to underage and overage costs (i.e., costs associated with the resource underutilization and overutilization). Due to the random processing times, jobs may finish earlier or later than they are expected to as illustrated in Fig. 4.1. If a job finishes earlier than the next appointment time, the system incurs underage costs because the server stays idle until the arrival of the next job. On the other hand, if a job is completed later than the next appointment time, the system incurs overage costs due to the waiting time of the next job and the overtime that may be experienced at the end of the schedule. The cost of overtime can be treated as the overage cost for the last job. The trade-off between the server idle time, job waiting time, and overtime can be balanced by using an appointment schedule that minimizes the expected total cost of underage and overage.

Begen and Queyranne (2011) study the single-server appointment scheduling problem, which is referred to as ASP, under the assumption that job durations are integer-valued discrete random variables, and they propose a solution framework for the problem based on the discrete convexity properties of the objective function. We now present the details of their approach.

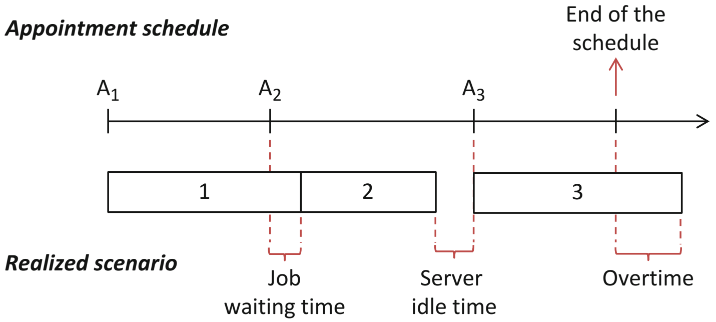


**Fig. 4.1** An appointment schedule and a realized scenario of the random processing times

There are $n$ jobs (indexed from 1 to $n$) to be processed in the order of increasing index. The integer-valued processing times for these jobs are given by a discrete joint distribution. It is assumed that the start time for the first job is zero, which is its appointment time. To compute the underage and overage cost for the $n$th job, a dummy job with a processing time of zero is introduced as the $(n+1)$st job, and its appointment time is the total time available for the $n$ real jobs.

The vector of random processing times is represented by $\mathbf{p} = \{p_1, p_2, \ldots, p_n, 0\}$, where $p_i$ is the random processing time of job $i$. $p_i \in [\underline{p}_i, \overline{p}_i]$, and $\overline{p}_{\max} = \max_i\{\overline{p}_i\}$.

Coefficients $u_i$ and $o_i$ are the per unit time underage and overage costs associated with the completion of job $i$.

The decision variables are the appointment times, represented in vector notation as $\mathbf{A} = \{A_1, A_2, \ldots, A_n, A_{n+1}\}$ with $A_1 = 0$. Note that the allocated duration for job $i$ is given by $A_{i+1} - A_i$ for $i = 1, \ldots, n$. $\mathbf{S} = \{S_1, S_2, \ldots, S_n, S_{n+1}\}$ and $\mathbf{C} = \{C_1, C_2, \ldots, C_n, C_{n+1}\}$ are job start times and completion times, which are the additional variables used to define the objective function. $S_1 = 0$ and $C_1 = S_1 + p_1 = p_1$ as the first job starts on time. The start time of the other jobs depends on their appointment time and the completion time of the immediately preceding jobs: $S_i = \max\{A_i, C_{i-1}\}$, and $C_i = S_i + p_i$ for $i = 2, \ldots, n$. The cost of appointment vector $\mathbf{A}$ given processing duration vector $\mathbf{p}$ is

$$F(\mathbf{A}|\mathbf{p}) = \sum_{i=1}^{n} o_i (C_i - A_{i+1})^+ + u_i (A_{i+1} - C_i)^+, \tag{4.1}$$

where $(x)^+ = \max(0, x)$.

The objective function to be minimized is the expected cost of the appointment vector:

$$F(\mathbf{A}) = \mathrm{E}_{\mathbf{p}}[F(\mathbf{A}|\mathbf{p})]. \tag{4.2}$$

Begen and Queyranne (2011) prove the following important properties of the objective function and the optimal appointment vector:

1. *Continuity*: [Begen and Queyranne (2011), Lemma 4.2] Functions $F(.|p)$ and $F(.)$ are continuous.

2. *Existence of an optimal vector*: [Begen and Queyranne (2011), Lemma 4.3] Let the compact set $\mathscr{K}$ be the cartesian product of the intervals $[\underline{A}_i, \overline{A}_i]$ (i.e., $\mathscr{K} = \prod_{i=1}^{n}[\underline{A}_i, \overline{A}_i]$), where $\underline{A}_1 = \overline{A}_1 = 0$, $\underline{A}_i = \sum_{j<i} \underline{p}_j$, and $\overline{A}_i = \sum_{j<i} \overline{p}_j$. There exists an appointment vector $\mathbf{A}^* \in \mathscr{K}$ such that $F(\mathbf{A}^*) \leq F(\mathbf{A})$ for any appointment vector $\mathbf{A}$.
3. *Nondecreasing appointment times*: [Begen and Queyranne (2011), Lemma 4.4] There exists an optimal appointment vector $\mathbf{A}^* \in \mathscr{K}$ with nondecreasing components, i.e., $A_i^* \leq A_{i+1}^*$ for all $i = 1, \ldots, n$.
4. *Appointment vector integrality*: [Begen and Queyranne (2011), Theorem 5.1] If the processing times are integer-valued random variables, then there exists an optimal appointment vector which is integer. (The proof is available in the appendix.)
5. *Computation of expected cost for a given appointment schedule*: [Begen and Queyranne (2011), Theorem 7.2] If the processing durations are stochastically independent and $\mathbf{A}$ is an integer appointment vector, then $F(\mathbf{A})$ may be computed in $\mathscr{O}(n^2 \overline{p}_{\max}^2)$ time. (The sketch of the proof is available in the appendix.)
6. *L-convexity*: [Begen and Queyranne (2011), Theorem 6.1] If there exist real numbers $\alpha_i$ $(1 \leq i \leq n)$ such that $0 \leq \alpha_i \leq o_i$ and $u_i + \alpha_i$ is nonincreasing in $i$, i.e., $u_i + \alpha_i \geq u_{i+1} + \alpha_{i+1}$ for all $i = 1, \ldots, n-1$ (which is referred to as the $\alpha$-monotonicity property (Begen and Queyranne 2011)), then $F(\mathbf{A})$ is L-convex.[1]
7. *Polynomial time algorithm*: [Begen and Queyranne (2011), Theorem 7.1] If the cost vectors $(\mathbf{u}, \mathbf{o})$ are $\alpha$-monotone and the processing durations are integer, then $F$ can be minimized using polynomial time and a polynomial number of expected cost evaluations.

Note that L-convexity of the objective function and the existence of a polynomial algorithm requires mild and intuitive conditions on the cost parameters. Since $\alpha_i$ is bounded by $o_i$, it captures the overage cost for job $i$ to some extent. Therefore, the $\alpha$-monotonicity property implies that the costs associated with the jobs planned to processed earlier in the schedule are higher. This is a reasonable condition since the impact of completion time of job $i$ on the overall schedule would be higher when there are more jobs to be processed after job $i$. Another way to look at this is to consider the case where $u_i$ is nonincreasing in $i$, which satisfies the $\alpha$-monotonicity property regardless of the overage cost structure. The nonincreasing structure of $u_i$ in $i$ is an intuitive condition as the underutilization of the resource earlier in the day has a higher impact on the schedule. Although the server will certainly be idle between $A_1 = 0$ and $A_2$ if job 1 is a no-show case, it is possible (due to the accumulated overtime) not to observe any idle time after $A_n$ even if job $n$ fails to show up.

[1] A function $f : \mathbb{Z}^q \to \mathbb{R} \cup \{\infty\}$ is L-convex iff $f(z) + f(y) \geq f(z \vee y) + f(z \wedge y)$ for all $z, y \in \mathbb{Z}^q$ and $\exists r \in \mathbb{R} : f(z+1) = f(z) + r \quad \forall z \in \mathbb{Z}^q$ where $z \vee y = (\max(z_i, y_i) : 0* \leq i \leq q) \in \mathbb{Z}^q, z \wedge y = (\min(z_i, y_i) : 0 \leq i \leq q) \in \mathbb{Z}^q$ (Murota 2003).

The modeling framework of Begen and Queyranne (2011) discussed here is highly flexible and can be used to formulate the ASP under various settings. For example, if the customer waiting is not important and the focus is only on minimizing underutilization and overutilization of the server, then the values of parameters can be set as $o_i = 0$ for $i \neq n$ and $u_i = u$ for all $i$. Begen and Queyranne (2011) present important extensions of their model to handle a given due date for the total processing time (i.e., for the end of the schedule), customer no-shows, and emergencies.

If there is a due date $D$ (such that $0 < D < \sum_{i=1}^{n} \overline{p}_i$) for the total processing time, then $A_{n+1}$ is constrained to be equal to $D$. If $D$ is integer and the processing times are integer variables, appointment vector integrality and the existence of a polynomial time algorithm remain valid [Begen and Queyranne (2011), Corollaries 8.2 and 8.5].

Customer no-shows can be included in the model by modifying the distributions of processing times based on the probability of no-shows. Let $noshow_i$ be the probability that customer $i$ will not show up. Then, the probability distribution of $p_i$ should be updated such that $p_i$ will be zero with probability $noshow_i$, and it will be $k$ ($\underline{p}_i \leq k \leq \overline{p}_i$) with probability $(1 - noshow_i)P\{p_i = k\}$.

Emergency jobs arrive randomly during the day, and they need to be processed as soon as possible. Begen and Queyranne (2011) incorporate the presence of emergency jobs into their model by considering a non-preemptive approach where the emergency jobs are processed immediately after the completion of the current job. The authors also assume that no emergency jobs arrive during the period when the server is idle, which fits well to the environments where the arrival rate of emergency jobs is low and/or the server idle time is significantly small compared to the total processing time of the jobs. Since the emergency jobs that arrive during the processing of job $i$ are processed immediately after job $i$ is completed, there is no idle time between job $i$ and the emergency jobs. Therefore, updating the probability distribution of job $i$'s duration by considering the addition of emergency jobs as an increase in job $i$'s duration extends the model for the case with emergency jobs.

Let $m^i_{\max}$ and $m^e_{\max}$ be the maximum number of emergency jobs that may arrive during the processing of job $i$ and the processing of an emergency job that arrive during job $i$, respectively. Then, the maximum number of emergency jobs to be processed after job $i$ is given by $M = m^i_{\max} + m^i_{\max} \times m^e_{\max}$. Let $\tilde{p}_i = p_i + P^i$ be the updated processing time of job $i$ where $P^i$ is the total processing time of emergent jobs to be processed after job $i$. Since the distribution of $p_i$ is already known, the distribution of $\tilde{p}_i$ can be obtained once that of $P^i$ is found. Let $m^i$ $(0 \leq m^i \leq M)$ be the number of emergency jobs to be processed after job $i$ and $P^i_k = \sum_{j=1}^{k} p^e$ $(1 \leq k \leq M)$ be the total processing time of $k$ emergency jobs where $p^e$ is a discrete random variable representing the processing time of a single emergency job. $P^i_1 = p^e$, and hence, their distributions are the same. The distributions of $P^i_k$ $(1 < k \leq M)$ can be found by using $P^i_k = P^i_{k-1} + p^e$ starting from $k = 2$. And finally, the distribution of $P^i$ $(1 \leq i \leq n)$ can be obtained as follows:

$$P\{P^i = 0\} = P\{m^i = 0\} + \sum_{k=1}^{M} P\{m^i = k\}P\{P^i_k = 0\}, \tag{4.3}$$

$$P\{P^i = j\} = \sum_{k=1}^{M} P\{m^i = k\}P\{P^i_k = j\} \quad \text{for} \quad j = 1,2,\ldots,M\overline{p}_{\max}. \tag{4.4}$$

### 2.1.2 A Sampling-Based Approach

Begen et al. (2012) study the ASP under the assumption that job duration probability distributions are not known, which is commonly experienced in practice. They assume that there is an (unknown) underlying (true) joint discrete distribution for job durations, and only samples are available, e.g., daily historical observations of surgery durations. Each sample is a vector of durations where each component corresponds to a job duration, and these vectors are independent. Unlike the common assumption in the literature, in Begen et al. (2012), the authors do not require the job duration probability distributions to be independent, i.e., job durations (the components of each duration vector) can be correlated. The authors use the convexity and subdifferential[2] of the objective function of the ASP developed in Begen (2010) to determine the number of independent samples required to obtain a provably near-optimal solution with high probability. The bound on the samples is polynomial in the number $n$ of jobs, accuracy level $\varepsilon$, confidence level $1-\delta$, and (underage and overage) cost coefficients $\mathbf{u}$ and $\mathbf{o}$, and it does not depend on the underlying distribution. The proposed approach guarantees that if the number of samples $N = N(\varepsilon,\delta,\mathbf{u},\mathbf{o})$ is greater than the obtained bound, then the cost of the sampling-based optimal schedule is with high probability (at least $(1-\delta)$) no more than $(1+\varepsilon)$ times the cost of an optimal schedule if the true distribution were known. Their approach is nonparametric and provides a sampling solution to the ASP, which is a multivariable non-separable integer stochastic program. Furthermore, as in Begen and Queyranne (2011), it can handle a given due date $D$ on the completion time of the last job.

We now present a brief overview of the approach used in Begen et al. (2012). They define $\mathbf{p}^k = (p^k_1, p^k_2, \ldots, p^k_n)$ as the $k$th observation in the $N$ samples, $\nu = \min_i\{u_i, o_i\}$, $o_{\max} = \max_i\{o_i\}$, and $u_{\max} = \max_i\{u_i\}$. The symbol " $\widehat{}$ " is used to denote quantities obtained from samples. $\widehat{\mathbf{p}} = \widehat{\mathbf{p}}(N)$ is defined as the empirical joint probability distribution obtained from $N$ independent observations of $\mathbf{p}$, i.e., $\text{Prob}\{\widehat{\mathbf{p}} = \mathbf{p}^k\} = \frac{1}{N}$ for $1 \le k \le N$. As in Begen and Queyranne (2011), a true optimal appointment vector is denoted with $\mathbf{A}^*$, i.e., $\mathbf{A}^*$ is a minimizer of $F_{\mathbf{p}}(\mathbf{A}) = \mathrm{E}_{\mathbf{p}}\left(F(\mathbf{A}|\mathbf{p})\right)$. Similarly, let $\widehat{\mathbf{A}} = \widehat{\mathbf{A}}(N)$ be a minimizer of $F_{\widehat{\mathbf{p}}}(\mathbf{A}) = \mathrm{E}_{\widehat{\mathbf{p}}}\left(F(\mathbf{A}|\widehat{\mathbf{p}})\right)$. Then the sampling ASP is defined as minimizing $F_{\widehat{\mathbf{p}}}(\mathbf{A})$, i.e., finding an $\widehat{\mathbf{A}}(N)$ for a given sample size $N$. The first step in the analysis is to prove that sampling ASP can be

[2] Subdifferential of a convex function $f$ is the set of all subgradients and denoted with $\partial f$.

solved in polynomial time for a given $N$. Then, with an application of Hoeffding's inequality (Hoeffding 1963), Begen et al. (2012) establish a link between the probabilities of a given event with respect to $\mathbf{p}$ and $\widehat{\mathbf{p}}$ as a function of sample size $N$ for a given accuracy level $\varepsilon'$ (absolute difference of the probabilities with respect to $\mathbf{p}$ and $\widehat{\mathbf{p}}$) and a confidence level $1-\delta'$. After that, the authors provide a similar result for a family of events $\mathscr{F}$. The next step is to use $\partial F_{\mathbf{p}}(\widehat{\mathbf{A}})$ characterization (which contains many events and probabilities) to show the existence of a subgradient $g \in \partial F_{\mathbf{p}}(\widehat{\mathbf{A}})$ such that each $k$ component of the subgradient $g_k$ satisfies $|g_k| < \varepsilon' K'$ with probability at least $1-|\mathscr{F}|\delta'$ where $|\mathscr{F}| = |\mathscr{F}|(n)$ and $K' = K'(n, \mathbf{u}, \mathbf{o})$ are some constants. Then Begen et al. (2012) shows that if there exists $g \in \partial F_{\mathbf{p}}(\widehat{\mathbf{A}})$ with $|g_k| < \varepsilon \nu / 3(n+1)n$ for all $1 \leq k \leq n+1$, then $F_{\mathbf{p}}(\widehat{\mathbf{A}}) \leq (1+\varepsilon) F_{\mathbf{p}}(\mathbf{A}^*)$. The last part of the analysis consists of a series of results establishing lower bounds on the objective function, using an application of Jensen's inequality and a new version of a multidimensional bounding lemma of Levi et al. (2007). All these results are put together in Theorem 14 of Begen et al. (2012) to give the bound on the number of samples. Before presenting the final result, we provide the new multidimensional bounding lemma below. We need a definition first. Let $f : \mathbb{R}^m \mapsto \mathbb{R}$ be convex. A point $y$ is an $\alpha$-point if there exists a subgradient $g \in \partial f(y)$ such that $||g||_1 \leq \alpha$ [Levi et al. (2007), Definition 3.3].

**Lemma 1 (Begen et al. (2012), Lemma 13).** *Let $f : \mathbb{R}^m \mapsto \mathbb{R}$ be convex, finite with a global minimizer $y^*$. Assume that there exists $\bar{f}$ such that $f \geq \bar{f} = \lambda ||y - \widetilde{y}||_1$ for some $\lambda > 0$ and $\widetilde{y} \in \mathbb{R}^m$. If $\widehat{y}$ is an $\alpha$-point for $\alpha = \lambda\varepsilon/3$, then $f(\widehat{y}) \leq (1+\varepsilon) f(y^*)$, where $\varepsilon \in [0,1]$ (The proof is available in the appendix).*

**Theorem 1 (Begen et al. (2012), Theorem 14).** *Let $0 < \varepsilon \leq 1$ (accuracy level) and $0 < 1-\delta < 1$ (confidence level) be given. If*

$$N > \left(4.5(1/\varepsilon)^2 \left(n^2(n+1)(14 o_{\max} + 6 u_{\max})/\nu\right)^2 \ln(2(5n^2+5)/\delta)\right),$$

*then $F_{\mathbf{p}}(\widehat{\mathbf{A}}) \leq (1+\varepsilon) F_{\mathbf{p}}(\mathbf{A}^*)$ with probability at least $1-\delta$.*

In Begen et al. (2012), the authors provide a sampling-based framework to solve the ASP (provably near optimal) with correlated job durations which does not require any distribution information on job durations.

### 2.1.3 Stochastic Programming Approach

Stochastic programming, which is a branch of mathematical programming that involves optimization problems with random parameters, is another methodology commonly used to formulate and solve appointment scheduling problems. Denton and Gupta (2003) consider the single-server ASP where the sequence of customers is fixed, the job durations are continuously distributed random variables, and the objective is minimizing the total expected cost of customer waiting, server idling, and overtime with respect to a due date for the end of processing (which is referred to as the session length). The authors formulate the problem as a two-stage SLP where the appointment times are determined before the resolution of uncertainty in job durations. They derive upper bounds that are independent of the distribution of job durations and solve the problem by using these bounds in a modified L-shaped algorithm that is based on successively partitioning the space of the random job durations. We now present their model, briefly describe their solution approach, and discuss the important insights drawn from their numerical results. The following notation is used in the formulation:

### Problem Parameters

$n$: Number of jobs to be scheduled
$d$: Session length
$\mathbf{c}^w$: Vector of cost coefficients for customer waiting
$\mathbf{c}^s$: Vector of cost coefficients for server idling
$c_l$: Cost coefficient for tardiness
$\mathbf{Z}(\omega)$: Vector of random job durations where $\omega$ represents the scenario index

### First-Stage Decision Variables

$\mathbf{x}$: Vector of allocated durations for the first $n-1$ jobs

### Second-Stage Decision Variables

$\mathbf{w}(\omega)$: Vector of waiting times for the jobs where $w_i(\omega)$ is the waiting time of job $i$
$\mathbf{s}(\omega)$: Vector of server idle times between consecutive jobs where $s_i(\omega)$ is the idle time between jobs $i-1$ and $i$
$l(\omega)$: Tardiness with respect to $d$
$g(\omega)$: Earliness with respect to $d$

Note that $\mathbf{x}$ is the vector of *first-stage* decision variables, and the *second-stage* decision variables are defined to describe the objective function. It is assumed that the appointment time for the first job is zero. The appointment time for job $i$ $(1 < i \leq n)$ is given by $\sum_{j=1}^{i-1} x_j$, which is the sum of durations allocated to its predecessors. The first job starts on time, and hence, $s_1(\omega) = w_1(\omega) = 0$ under each scenario. For each scenario, $s_i(\omega)$ and $w_i(\omega)$ for other jobs (i.e., $1 < i \leq n$) can be written in a recursive way as follows:

$$w_i(\omega) = (w_{i-1}(\omega) + Z_{i-1}(\omega) - x_{i-1})^+, \tag{4.5}$$

$$s_i(\omega) = (-w_{i-1}(\omega) - Z_{i-1}(\omega) + x_{i-1})^+. \tag{4.6}$$

The tardiness and earliness with respect to the session length are given by

$$l(\omega) = (w_n(\omega) + Z_n(\omega) + \sum_{i=1}^{n-1} x_i - d)^+, \tag{4.7}$$

$$g(\omega) = (-w_n(\omega) - Z_n(\omega) - \sum_{i=1}^{n-1} x_i + d)^+. \tag{4.8}$$

Considering these equalities that determine the values of $\mathbf{w}(\omega), \mathbf{s}(\omega), l(\omega)$, and $g(\omega)$ for a given $(\mathbf{x}, \mathbf{Z}(\omega))$, the problem can be formulated as the following two-stage SLP:

$$\min \quad E_{\mathbf{Z}} \left[ \sum_{i=2}^{n} c_i^w w_i(\omega) + \sum_{i=2}^{n} c_i^s s_i(\omega) + c_l l(\omega) \right] \tag{4.9}$$

s.t.

$$w_2(\omega) - s_2(\omega) = Z_1(\omega) - x_1 \qquad \forall \omega, \tag{4.10}$$

$$- w_{i-1}(\omega) + w_i(\omega) - s_i(\omega) = Z_{i-1}(\omega) - x_{i-1} \qquad \forall i > 2, \omega, \tag{4.11}$$

$$- w_n(\omega) + l(\omega) - g(\omega) = Z_n(\omega) - d + \sum_{i=1}^{n-1} x_i \qquad \forall \omega, \tag{4.12}$$

$$x_i \geq 0 \qquad \forall i < n, \tag{4.13}$$

$$w_i(\omega), s_i(\omega) \geq 0 \qquad \forall i > 1, \omega, \tag{4.14}$$

$$l(\omega), g(\omega) \geq 0 \qquad \forall \omega. \tag{4.15}$$

Note that $s_1(\omega)$ and $w_1(\omega)$ are not included in the formulation because they are equal to zero. The vector of first-stage decision variables $(\mathbf{x})$ is determined before the resolution of uncertainty and hence is scenario independent. On the other hand, the second-stage decisions $(\mathbf{w}(\omega), \mathbf{s}(\omega), l(\omega)$, and $g(\omega))$ are made after the realization of the random scenario, and they are scenario dependent. The above formulation can be written in a compact form as follows:

$$\min_{\mathbf{x} \geq 0} \quad \{\mathscr{Q}(x)\}, \tag{4.16}$$

where

$$\mathscr{Q}(x) = E_{\mathbf{Z}}[Q(\mathbf{x}, \mathbf{Z}(\omega))], \tag{4.17}$$

$$Q(\mathbf{x}, \mathbf{Z}(\omega)) = \min_{\mathbf{y}(\omega) \geq 0} \{\mathbf{cy}(\omega) | \mathbf{Tx} + \mathbf{Wy} = \mathbf{h}(\omega)\}, \tag{4.18}$$

$$\mathbf{c} = \begin{bmatrix} c^w \\ c_l \\ c^s \\ 0 \end{bmatrix}, \mathbf{y}(\omega) = \begin{bmatrix} \mathbf{w}(\omega) \\ l(\omega) \\ \mathbf{s}(\omega) \\ g(\omega) \end{bmatrix}, \mathbf{h}(\omega) = \begin{bmatrix} Z_1(\omega) \\ \vdots \\ Z_n(\omega) - d \end{bmatrix},$$

$$\mathbf{T} = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ -1 & \cdots & -1 \end{bmatrix}, \mathbf{W} = \left[ \mathbf{W}' | -\mathbf{I} \right], \tag{4.19}$$

where **I** is the identity matrix and

$$\mathbf{W}' = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}. \tag{4.20}$$

$\mathcal{Q}(x)$ and the coefficient matrices **T** and **W** are called the *recourse function* and *technology* and *recourse* matrices, respectively, in the stochastic programming literature.

Given a first-stage solution, the second-stage problem (4.18), which is also known as the *subproblem*, can be solved independently for every scenario when the number of scenarios (i.e., possible job duration realizations) is finite. Owing to this property, stochastic programs can be efficiently solved using decomposition methods (Higle and Sen 1991; Ruszczyński 1986; Van Slyke and Wets 1969). One of the commonly used decomposition methods for two-stage SLPs is the L-shaped algorithm (Van Slyke and Wets 1969), which is an extension of Benders decomposition (Benders 1962) for stochastic programs. It provides a framework where a two-stage SLP is decomposed into a master problem (composed of the first-stage variables and constraints) and scenario-dependent subproblems, and the master problem is solved iteratively until the optimal solution is reached. At each iteration, after solving the master problem, second-stage subproblems are solved given the solution of the master problem. Dual variables of the subproblems are used to generate feasibility and optimality cuts, which are added to the master problem to ensure the second-stage feasibility and to approximate the recourse function, respectively, throughout the iterations. The optimality cuts include an auxiliary variable, $\theta$, which is used to outer linearize the recourse function via optimality cuts that are added at each iteration.

To solve (4.9)–(4.15) when job durations are continuously distributed, Denton and Gupta (2003) propose an approach where the space of random job durations is partitioned and a scenario is constructed from each partition based on the expectation of the job durations on that partition. The resulting discrete problem is then solved by using the L-shaped algorithm. If the gap between the lower and upper bounds associated with the solution returned by the L-shaped algorithm is larger than a desired tolerance level, the current partition is refined, and a new and larger size discrete problem is constructed to be solved by the L-shaped algorithm. The refinement of the partition at each iteration is made in a way that the expectation of the approximate distribution of job durations converges to that of the true distribution. These steps continue in an iterative manner until a solution with an acceptable level of optimality gap (i.e., the gap between the corresponding lower and upper bounds) is found. Denton and Gupta (2003) derive upper bounds which are independent of the job duration distributions. These upper bounds are then used to measure the gap from the lower bound returned (as the $\theta$ variable) by the L-shaped algorithm.

The numerical results of Denton and Gupta (2003) reveal that the optimal vector of allocated durations to jobs exhibits a dome-shaped structure if the job durations are independent and identically distributed, and the waiting and idling costs are uniform. While the dome-shaped pattern is particularly remarkable when the ratio of idling to waiting cost is high, the allocated durations are observed to be more uniform when this ratio is low.

## 3 Operating Room Scheduling

Appointment scheduling and OR scheduling, which is also known as surgery scheduling, are closely related problems (Chap. 5). OR scheduling may involve decisions other than the surgery start times including, but not limited to, the allocation of available OR time among surgical specialties, allocation of surgeries to ORs, sequence of surgeries in ORs, and cancelation/rescheduling of surgeries. Moreover, multiple types of resources such as ORs, surgeons, and surgical equipment may be considered in OR scheduling problems.

In many surgical environments, OR scheduling decisions are sequentially made in three stages (Santibáñez et al. 2007). At the first level, available OR time is allocated among surgical specialties by considering several factors such as the demand forecast for each specialty, the level of supporting resources, and the OR time. At the second level, a block schedule is generated by assigning OR time slots, which are also referred to as blocks of OR time, to surgeons or surgical groups each day. Finally, at the third level, surgeries are scheduled in ORs on a daily basis. At this level, the number of surgeries, their sequence, and planned start times (appointment times) are determined. Both deterministic and stochastic optimization models have been widely used to investigate OR scheduling problems. Extensive reviews of the related studies can be found in Cardoen et al. (2010), Erdogan and Denton (2011a), Guerriero and Guido (2011), Gupta (2007), May et al. (2011), as well as chap. 5 of this book.

Blake and Donald (2002) study the problem of generating a weekly schedule for a multi-OR surgical suite at Toronto's Mount Sinai Hospital. Available OR time is typically represented as a collection of blocks of OR time, and a block is characterized by its length. The authors propose a *mixed-integer program* (MIP) to distribute the available OR time among the surgical wards based on ward-specific target times. The objective is to minimize the undersupply of OR time subject to constraints associated with the lower and upper bounds on the number of blocks that could be assigned. Santibáñez et al. (2007) explore the related problem of scheduling OR blocks in a system of hospitals. To design a weekly schedule, they formulate the problem as a MIP which determines the assignment of ORs to surgical specialties and the number of surgeries from each surgical group in each OR. The constraints in their model ensure that the required level of resources (surgeons, OR time, and staffed beds) does not exceed the existing capacity. Hospital specific lower and upper bounds on the number of OR blocks assigned to each surgical specialty and the number of surgeries from each surgical group are also represented as a set of constraints. The authors consider two different objectives: minimizing the maximum number of required beds during the planning horizon and maximizing the throughput of patients. As observed from their numerical results, the first objective is helpful in reducing the variability in the bed utilization, whereas the second objective can be used to stabilize the wait lists. Constructing a cyclic schedule of surgeries through the use of an optimization-based approach has also been considered in several other studies (Adan et al. 2009; Beliën and Demeulemeester 2007; Testi et al. 2007; van Oostrum et al. 2008).

Velásquez and Melo (2006) study the problem of scheduling multiple ORs where each surgery has a preferred start time and the objective is to meet these preferences as much as possible. They formulate the problem as a set packing problem by discretizing the planning horizon and considering all possible combinations of resources for each time unit. They use column generation and constraint branching to solve practical instances of the problem.

Jebali et al. (2006) consider a multi-OR scheduling problem in the presence of surgical equipment-related eligibility constraints. They propose a two-step hierarchical solution approach for the problem. In the first step, surgeries are assigned to ORs using a MIP that minimizes the total cost of overtime, OR idle time, and patient waiting time. In the second step, the total overtime is reduced further by employing another MIP that determines the surgery sequence in ORs. Fei et al. (2010) propose a similar hierarchical approach to solve a multi-OR scheduling problem over a weekly planning horizon where the objective is to minimize the total cost of OR idle time and overtime. They first assign a date for each surgery by solving a set-partitioning problem through the use of a column-generation-based heuristic, then solve the daily scheduling problem by using a hybrid genetic algorithm.

Denton et al. (2007) investigate the joint impact of surgery sequencing and start time decisions in the presence of uncertainty in surgery durations by extending the single-server appointment scheduling problem (Denton and Gupta 2003). They consider several different surgery sequencing rules including shortest expected duration first, smallest variance of durations first, and smallest coefficient of variation first. Their computational results indicate that the performance of OR schedules is sensitive to both sequencing and start time decisions, and sequencing surgeries in the order of increasing variance of durations performs better than the other two sequencing rules in most of the test instances. The impact of these two types of decisions is also studied by Cayirli et al. (2006). They generate the efficient frontier by simulating a single-server appointment system under several sequencing and appointment rules to explore the trade-off between patient-related and doctor-related measures. Their results also indicate that both sequencing and start time decisions are important in designing well-performing surgery schedules.

Denton et al. (2010) explore the problem of allocating surgeries to ORs on a given surgical day where the surgery durations are uncertain and the objective is to minimize the total fixed cost of opening ORs and the expected overtime cost. They formulate the problem as a two-stage *stochastic mixed-integer program* (SMIP) where the number of ORs to open and the surgery-to-OR allocation decisions are made before the day starts, i.e., before the resolution of uncertainty in surgery durations. The authors also present a robust formulation, which is easier to solve compared to the two-stage SMIP, where the objective is to minimize the maximum cost associated with the uncertainty described by lower and upper bounds on surgery durations. To solve the realistic-sized instances of the problem, the authors develop valid inequalities that reduce symmetry and use lower and upper bounds on the optimal number of ORs to open each day. Moreover, they test the performance of a heuristic rule which allocates the surgeries to ORs in a successive manner where the surgery with the largest expected duration is assigned to the first available OR in each step. Their numerical results show that the value of the stochastic solution is particularly high when the per unit time cost of overtime is high, the heuristic performs notably well when the per unit cost of overtime is low, and the solution obtained by solving the robust formulation performs approximately as well as the heuristic solution.

Batun et al. (2011) extend the stochastic multi-OR scheduling problem introduced by Denton et al. (2010) by including surgeons, besides ORs, as a set of resources in the model. The authors propose a stochastic programming-based approach for modeling and solving the problem, and they estimate the value of different resource usage schemes used in practice. We present the details of their approach and briefly discuss their results in Sect. 3.1.

## *3.1 Scheduling Multiple Operating Rooms Under Uncertainty*

Batun et al. (2011) formulate and solve the multi-OR scheduling problem with multiple surgeons as a two-stage SMIP where the objective is to minimize the total fixed cost of opening ORs and the expected overtime and surgeon idling cost. Decisions made before the resolution of uncertainty include the number of ORs to open, the allocation of surgeries to ORs, the sequence of surgeries in each OR, and the times at which surgeons start their first surgery of the day. Pre-incision, incision, and post-incision phases of the surgery and the resource setup times are explicitly considered, which makes it possible to accurately model different resource usage schemes such as OR pooling and parallel surgery processing. OR pooling is defined as treating ORs as a common shared resource among surgeons and allowing surgeries of different surgeons to be scheduled in the same OR. Although surgeons are key members of the surgical team in academic medical centers, pre-incision and post-incision phases may also be performed by other members of the team such as surgical fellows. Therefore, surgeons need to be in the OR only during the incision phase. As a result, multiple surgeries by a single surgeon may occur simultaneously across multiple ORs, which is referred to as parallel surgery processing. The authors solve the problem by employing the L-shaped algorithm both in an iterative framework and in a branch-and-cut framework. To increase the solvability of practical instances, they utilize a number of valid inequalities: symmetry elimination constraints which are first introduced by Denton et al. (2010) in the OR scheduling context, induced constraints to ensure second-stage feasibility, and a set of widely applicable valid inequalities which are based on Jensen's inequality (Jensen 1906). They perform a series of computational experiments to estimate the value of the stochastic solution, to quantify the potential benefit of OR pooling, and to illustrate the impact of parallel surgery processing. Their results indicate that capturing uncertainty and pooling ORs may achieve significant cost reductions especially when the per unit time surgeon idling cost is high, and the impact of parallel surgery processing becomes higher as the resource setup times and the parallelizable portion of surgeries (which is comprised of the pre-incision and post-incision phases) increase.

We now present the two-stage SMIP, formulated by Batun et al. (2011) and the valid inequalities used to increase the problem solvability.

## Model Formulation

### *Indices*

$i, j$: Surgery indices
$k$: Surgeon index
$q, r$: OR indices
$\omega$: Scenario index
$i_k$: Index of the first surgery of surgeon $k$

### *Parameters*

$L$: Session length for each OR (i.e., regular working time)
$c^f$: Daily fixed cost of opening an OR
$c^o$: Per minute overtime cost of an OR
$c^S$: Per minute idle time cost of a surgeon
$s^S$: Surgeon turnover time between two consecutive surgeries
$s^R$: OR turnover time between two consecutive surgeries
$n$: Total number of surgeries to be scheduled
$n_R$: Total number of available ORs
$n_S$: Total number of surgeons

$b_{ijk}$: Binary parameter denoting whether surgery $i$ immediately precedes surgery $j$ in surgeon $k$'s surgery listing

$pre_i(\omega)$: Pre-incision duration of surgery $i$ under scenario $\omega$

$p_i(\omega)$: Incision duration of surgery $i$ under scenario $\omega$

$post_i(\omega)$: Post-incision duration of surgery $i$ under scenario $\omega$

$\xi(\omega)$: Random vector composed of pre-incision, incision, and post-incision durations of surgeries; $\xi(\omega) = (\mathrm{pre}_1(\omega),\ldots,\mathrm{pre}_n(\omega), p_1(\omega),\ldots, p_n(\omega), \mathrm{post}_1(\omega), \ldots, \mathrm{post}_n(\omega))$, and the finite support of $\xi(\omega)$ is $\Xi$ where $\Xi \in \mathbb{R}^{3n}_{+}$

### First-Stage Decision Variables

$x_r$: Binary decision variable denoting whether OR $r$ is opened or not

$y_{ir}$: Binary decision variable denoting whether surgery $i$ is allocated to OR $r$ or not

$z_{ijr}$: Binary decision variable denoting whether surgery $i$ precedes surgery $j$ in OR $r$ or not (defined for $(i,j,r) : i \neq j$)

$t_k$: Start time for surgeon $k$

### Second-Stage Decision Variables

$C_{ir}(\omega)$: Completion time for surgery $i$ in OR $r$ under scenario $\omega$

$I_{ij}(\omega)$: Surgeon idle time between surgeries $i$ and $j$ under scenario $\omega$ (defined for $(i,j) : \sum_{k=1}^{nS} b_{ijk} = 1$)

$I_k(\omega)$: Idle time of surgeon $k$ before his/her first surgery under scenario $\omega$

$O_r(\omega)$: Overtime in OR $r$, with respect to session length $L$ under scenario $\omega$

Considering the above notation, and letting the vector form of the first-stage decision variables be represented by $x, y, z$, and $t$, the two-stage SMIP formulated by Batun et al. (2011) is as follows:

$$\min \sum_{r=1}^{n_R} c^f x_r + \mathscr{Q}(x,y,z,t) \tag{4.21}$$

s.t.

$$y_{ir} \leq x_r \forall i, r, \tag{4.22}$$

$$\sum_{r=1}^{n_R} y_{ir} = 1 \forall i, \tag{4.23}$$

$$z_{ijr} + z_{jir} \leq y_{ir} \forall i, j > i, r, \tag{4.24}$$

$$z_{ijr} + z_{jir} \leq y_{jr} \forall i, j > i, r, \tag{4.25}$$

$$z_{ijr} + z_{jir} \geq y_{ir} + y_{jr} - 1 \forall i, j > i, r, \tag{4.26}$$

$$t_k \leq L \quad \forall k, \tag{4.27}$$

$$x_r, y_{ir}, z_{ijr} \in \{0,1\} \forall i, j \neq i, r, \tag{4.28}$$

$$t_k \geq 0 \forall k, \tag{4.29}$$

where

$$\mathscr{Q}(x,y,z,t) = E_{\xi}\left[Q(x,y,z,t,\xi(\omega))\right] \tag{4.30}$$

and

$$Q(x,y,z,t,\xi(\omega)) = \min \sum_{r=1}^{n_R} c^o O_r(\omega) + \sum_{(i,j):\sum_{k=1}^{n_S} b_{ijk}=1} c^S I_{ij}(\omega) + \sum_{k=1}^{n_S} c^S I_k(\omega) \tag{4.31}$$

s.t.

$$C_{ir}(\omega) \le M y_{ir} \forall i,r, \tag{4.32}$$

$$C_{jr}(\omega) \ge C_{ir}(\omega) + s^R + \text{pre}_j(\omega) + p_j(\omega) + \text{post}_j(\omega) - M(1 - z_{ijr})$$

$$\forall i, j \ne i, r, \tag{4.33}$$

$$\sum_{r=1}^{n_R} C_{i_k r}(\omega) = t_k + I_k(\omega) + \text{pre}_{i_k}(\omega) + p_{i_k}(\omega) + \text{post}_{i_k}(\omega) \forall k, \tag{4.34}$$

$$\sum_{r=1}^{n_R} C_{ir}(\omega) \ge t_k + \text{pre}_i(\omega) + p_i(\omega) + \text{post}_i(\omega) \forall (i,k) : \sum_{j=1}^{n} b_{jik} = 1, \tag{4.35}$$

$$\sum_{r=1}^{n_R} C_{jr}(\omega) = \sum_{r=1}^{n_R} C_{ir}(\omega) - \text{post}_i(\omega) + s^S + p_j(\omega) + \text{post}_j(\omega) + I_{ij}(\omega)$$

$$\forall (i,j) : \sum_{k=1}^{n_S} b_{ijk} = 1, \tag{4.36}$$

$$O_r(\omega) \ge C_{ir}(\omega) - L \forall i,r, \tag{4.37}$$

$$C_{ir}(\omega), I_k(\omega), I_{ij}(\omega), O_r(\omega) \ge 0 \forall i,j,r,k. \tag{4.38}$$

The objective function (4.21) is the sum of the first-stage cost (the fixed cost of opening ORs) and the expected second-stage cost (the sum of expected overtime costs and surgeon idle time costs) over all scenarios. Constraints (4.22) and (4.23) determine the set of ORs to be opened and the allocation of the surgeries to ORs. The sequence of surgeries in each OR is described by constraints (4.24)–(4.26), which ensure that a precedence relation exists between two surgeries if and only if they are both assigned to the same OR. Constraint (4.27) represents the operational requirement that each surgeon should start performing his/her first surgery of the day before the daily session ends.

The second-stage problem for a given first-stage solution $(x,y,z,t)$ and a scenario $\omega$ is formulated by (4.31)–(4.38). Constraints (4.32)–(4.33) describe the surgery completion times based on the assignment and sequencing decisions, where the $M$

parameter represents an upper bound on the surgery completion times. Constraints (4.34)–(4.36) provide the relation between surgery completion times and surgeon idle times by considering the surgery sequence in surgeons' surgery listing. Constraint (4.37) defines the overtime in each OR.

The authors initially tried to solve (4.21)–(4.38) by using the L-shaped algorithm, which fails to solve even moderate-sized instances in a reasonable amount of time. Therefore, they enhance the algorithm by utilizing the valid inequalities discussed in the remainder of this section.

### Symmetry-Breaking Constraints

The above model is that of a surgical environment with identical ORs, and hence, there is complete symmetry with respect to ORs. Solving highly symmetric *integer programs* (IPs) by using a standard solution technique requires too much computational effort due to the existence of many alternative solutions. Symmetry-breaking constraints have been widely used to overcome this difficulty (Margot 2002; Ostrowski et al. 2011; Sherali and Smith 2001). Batun et al. (2011) add the following first-stage constraints, which are introduced by Denton et al. (2010) in the context of OR scheduling, to (4.21)–(4.38) to increase the problem solvability by reducing symmetry:

- An arbitrary ordering of ORs is introduced, and open ORs are enforced to be the smaller-indexed ones:

$$x_r \geq x_{r+1} \qquad \forall r < n_R. \tag{4.39}$$

- Open ORs are differentiated from each other by imposing a lexicographic order in terms of the indices of surgeries in each OR:

$$\sum_{r=1}^{i} y_{ir} = 1 \quad \forall i \leq \min\{n, n_R\}. \tag{4.40}$$

- Symmetry is reduced further by ensuring that the $i$th surgery is allocated to one of the first $r$ ORs if the first $i-1$ surgeries are allocated to the first $r-1$ ORs:

$$\sum_{q=r}^{\min\{i,n_R\}} y_{iq} \leq \sum_{j=r-1}^{i-1} y_{j,r-1} \quad \forall (i,r) : i \geq r > 1. \tag{4.41}$$

### Induced Constraints

Since the L-shaped algorithm is a decomposition-based algorithm which solves the master problem (first-stage problem) and the subproblems (second-stage problem) sequentially in an iterative framework, feasibility of the second stage is not guaranteed for every first-stage solution generated throughout the iterations. In other words, the master problem may yield an infeasible solution. To overcome this issue, feasibility cuts are added, which may be time-consuming if infeasibility with respect to second stage is frequently encountered during the algorithm. Therefore, instead of generating feasibility cuts at many iterations of the L-shaped algorithm, Batun et al. (2011) add the following *induced constraints* to the master problem a priori to ensure second-stage feasibility, i.e., to induce relatively complete recourse.

**Proposition 1 (Batun et al. (2011), Proposition 1).** *A first-stage solution $(x,y,z,t)$ is feasible for first- and second-stage problems if it satisfies (4.22)–(4.29), (4.39)–(4.41), and*

$$u_j \geq u_i + d - nd\left(1 - \sum_{r=1}^{n_R} z_{ijr}\right) \quad \forall i, j \neq i, \tag{4.42}$$

$$u_j \geq u_i + d \quad \forall (i,j) : \sum_{k=1}^{n_S} b_{ijk} = 1, \tag{4.43}$$

*where $u_i$'s are nonnegative auxiliary first-stage decision variables and $d$ is a positive finite scalar.*

Imposing a lower bound, $d$, on the difference of completion times of surgeries which are allocated to the same OR or performed by the same surgeon, constraints (4.42) and (4.43) ensure that $z$ yields an acyclic surgery sequence within each OR and with respect to each surgeon across the ORs, respectively. As a result, any feasible first-stage solution that also satisfies (4.42) and (4.43) is feasible for the second-stage problem under each scenario. Therefore, adding (4.42) and (4.43) to the problem eliminates the computational burden of generating feasibility cuts, and the problem can be solved by an L-shaped algorithm that generates optimality cuts only.

### Jensen's Based Valid Inequalities

As described in Sect. 2.1.3, L-shaped optimality cuts include the first-stage variables and $\theta$, which is an auxiliary variable that links the first and second stages by approximating the recourse function $\mathscr{Q}(x,y,z,t)$ through outer linearization. In a minimization problem, $\theta$ acts as a lower bound on $\mathscr{Q}(x,y,z,t)$ throughout the iterations, progressively becomes a better approximation, and finally becomes equal to $\mathscr{Q}(x,y,z,t)$, which is indeed the stopping criterion for the algorithm.

In order to increase the convergence, the authors provide a stronger formulation of the first-stage problem by bounding $\theta$ through the use of a set of valid inequalities based on the following proposition:

**Proposition 2 (Based on Batun et al. (2011), Propositions 2 and 3).** *For a feasible first-stage solution* $(x,y,z,t)$ *of our problem, let* $Q(x,y,z,t,\bar{\xi}(\omega))$ *be the corresponding second-stage cost under the mean value scenario. Then,*

$$\theta \geq Q(x,y,z,t,\bar{\xi}(\omega)) \tag{4.44}$$

*is a valid inequality, and hence can be added to the first-stage problem.*

The valid inequality described by (4.44) is applicable to two-stage stochastic programs whose recourse function is convex in the random components of the problem parameters (i.e., in $\xi(\omega)$). By using auxiliary parameters and decision variables to describe the second-stage problem under the mean value scenario, Batun et al. (2011) utilize (4.44) as a first-stage constraint, which results in a stronger master problem formulation that significantly speeds up the convergence of the L-shaped algorithm.

## 4 Capacity Planning

Due to the need to achieve high service levels, the existence of expensive and specialized resources, and the uncertainty in demand, capacity planning is an important and challenging problem in healthcare. It takes place at different decision levels ranging from strategic to operational and involves decisions concerning the use of resources such as hospital beds, ORs, diagnostic or treatment equipment, and workforce. Studies considering the use of some of these resources are reviewed in other sections of this chapter. In this section, we focus our attention on the studies related to the bed capacity planning and capacity planning in the presence of different groups of patients. For more extensive reviews, we refer the reader to Green (2004) and Smith-Daniels et al. (1988).

Depending on the level of decision making, bed capacity planning problems may be considered in a single- or a multihospital setting. Akcali et al. (2006) formulate an aggregate level bed capacity planning problem in a hospital over a finite horizon as a nonlinear IP which determines the optimal timing and magnitude of changes in bed capacity. The objective of their model is to minimize the sum of the expected patient waiting cost, the cost of changing bed capacity, and the cost of operating at the planned levels of capacity over the considered horizon. Besides representing the initial conditions (i.e., the number of beds at the beginning of the horizon) and the flow balance, the constraints of the model also impose bounds on the expected delay for a patient and the cost of capacity change in any period. In practice, bed capacity is changed in batches rather than an arbitrary integer value. Considering this practical restriction, the authors convert their formulation to a nonlinear binary IP for which they develop a network representation. They solve the problem by finding the shortest path between the superficial source node and the superficial sink node in polynomial time using Dijkstra's algorithm.

Duncan and Noble (1979) study the allocation of the available bed capacity to different specialties in a multihospital setting. They formulate the problem as a maximum flow network model where the number of beds allocated to a specialty in a hospital is defined as the flow on the arc connecting the corresponding pair of nodes. The constraints of the model ensure that the available bed capacity at each hospital is not exceeded, the required capacity for each specialty is supplied, and some other logical and resource-related requirements are met. The authors solve the problem by using a two-step procedure. They first determine the set of feasible allocation schemes (i.e., the set of arcs to be used) with respect to the logical constraints, and then calculate the maximum flow for every feasible scheme to select the best among them.

Ben Abdelaziz and Masmoudi (2012) consider an aggregate level multi-objective bed capacity planning problem in a multihospital setting where the hospitals are located in different regional areas in a country. The specialties in the hospitals deliver primary, secondary, and tertiary care, and the random demand for these different levels of care must be met by the hospital receiving the demand, one of the hospitals in the same regional area, and one of the hospitals in the country, respectively. The decision to be made is how much to expand the existing staffed bed capacity in each specialty at each hospital in a way that the sum of the capacity expansion cost, the expected cost of transferring unmet demand (for secondary and tertiary care) to private hospitals, and the salary of the new physicians and nurses is minimized. In other words, the authors treat different objectives by combining them into a single objective function. They formulate the problem as a two-stage stochastic IP and solve the corresponding extensive form in their particular implementation. Multi-objective optimization-based approaches are also considered in other studies that explore bed capacity planning problems at different levels (Li et al. 2009; Stummer et al. 2004).

Sandikci et al. (2011) explore the trade-offs inherent in managing the inpatient bed capacity in an urban teaching hospital whose mission makes it necessary to consider some special case-mix-related requirements. Due to these requirements, not only the demand but also the utility (which is defined as a combined measure of complexity of care and cost of care) associated with each patient type should be considered when managing the bed capacity. There are two possible strategies that can be adopted: (1) pooling the bed capacity and (2) subdividing the bed capacity into specialty wings. While the former strategy is better in terms of achieving high bed occupancy levels, the latter one provides an improved control of case-mix and increased efficiency of care. The authors formulate the problem of designing wings of specialties as a mathematical program that maximizes the total expected utility generated by the occupied beds where the expected number of occupied beds and the utility from an occupied bed are wing dependent. The decisions in their model include how many wings to form, how many beds to allocate to each wing, and which specialties to assign to each wing. Moreover, the average length-of-stay (LOS) for each wing is endogenously determined by using a function that decreases as the demand increases and/or the number of specialties assigned to the wing decreases. For a fixed sequence of patient types, the wings can be formed by placing cuts in the sequence, and the optimal placement of the cuts can be efficiently found through the use of a *dynamic program* (DP). The preliminary results provided by the authors indicate that the utility increase relative to one-wing solution (i.e., pooled capacity of beds) becomes very significant as the daily demand and/or the impact of narrowed focus of wings on the LOS increases. The results also reveal that the optimal capacity allocated to lower-utility patients tends to decrease with the increasing demand.

Another stream of research in this vein focuses on the scheduling problems concerning the use of scarce resources that receive demand from multiple sources. Most of the related studies consider stochastic DPs to formulate and solve these problems. We refer the reader to Begen (2011) and Patrick and Begen (2011) for an overview of stochastic DPs and their applications in healthcare. Gerchak et al. (1996) explore the problem of allocating daily surgical capacity to elective and emergent cases under uncertainty in elective surgery durations and emergent surgery demand. At the beginning of each day, the scheduler needs to decide how many of the elective surgery requests should be accepted (i.e., how many of them should be postponed and added to the demand of the next day). The authors formulate the problem as a stochastic DP that maximizes the expected profit, which is comprised of the revenue generated by performing the elective surgeries, penalty incurred due to postponing the booking of the elective surgeries, and the cost of overtime experienced due to exceeding the daily capacity. They prove that the optimal policy is monotone (i.e., the optimal number of elective surgery requests to accept increases with the number of requests), but not necessarily of a control limit type (i.e., a threshold level that bounds the number of elective surgery requests to be accepted does not exist).

Green et al. (2006) study the scheduling of inpatients and outpatients on a diagnostic facility over a finite number of identical service slots during a day, where the demand from inpatients arises randomly and the outpatients are booked patients with an associated probability of no-show. Besides inpatients and outpatients, there are also emergent patients that may arrive during the day, and they need to be served as soon as possible, i.e., in the next service slot. If there is no emergent patient, and there are both inpatients and outpatients waiting to be served at the beginning of a service slot, then the scheduler needs to decide which type of patient should be served next. The authors formulate the problem as a finite-horizon DP that maximizes the profit, which is composed of the revenue generated by serving the patients, patient waiting cost, and end-of-day penalty cost of not being able to serve some of the patients. While the first two of these metrics are greater for an outpatient, the end-of-day penalty cost of not seeing an inpatient dominates that of an outpatient, which results in a trade-off between serving an inpatient and an outpatient. Green et al. (2006) analyze structural properties of their model and prove that the optimal policy is of a control limit type.

Gupta and Wang (2008) consider the problem of scheduling patients in a primary care clinic where the patients are grouped based on the type of their appointment request: regular patients (i.e., the patients with a scheduled appointment) and same-day patients (i.e., the patients who request a same-day appointment). See Patrick

(2012), Robinson and Chen (2009) and the references therein for research related to same-day patients (open access). All same-day patients need to be accommodated either by serving at the clinic (using regular or overtime capacity) or by diverting to other clinics. Patients' preferences, in terms of the physician and the appointment time, are considered when booking the future appointments. Scheduling too many appointments in advance may create capacity problems in meeting the same-day demand, whereas scheduling too few future appointments may result in unused capacity. Considering this trade-off, the problem is to decide whether or not to accept a future appointment request given the patient's preferences and the state of the system. The authors formulate the problem as a discrete-time finite-horizon *Markov decision process* (MDP) that maximizes the revenue. They explore the properties of their model and prove that the optimal policy exhibits a control limit structure when the clinic is served by a single physician. For the case where the clinic is served by multiple physicians, they propose heuristic methods that perform quite well across a number of instances.

The problem of booking appointments in a primary care clinic in the presence of different patient groups and patient preferences is also studied by Wang and Gupta (2011). They propose a more realistic framework where patients do not have complete knowledge about the system state and patient preferences are adaptively updated each time an appointment request is made. The authors consider the problem for each working day separately. An appointment request from a patient is described as an unordered set of preferred physician and time slot combinations on the considered working day. Once a request is made, either an appointment is booked considering the patient's preferences or the request is denied and the patient is asked to try another set of combinations or another day. The patients are modeled as different revenue classes based on the type of their appointments (which may be a scheduled or a same-day appointment) and the physician they would like to see (i.e., their preferred care provider (PCP)). The objective is maximizing the expected revenue. Therefore, a request may be denied (even if there are available combinations among the set of requested ones) if reserving some appointment slots for future arrivals results in a higher expected revenue. The problem is formulated as a stochastic dynamic program where at most one appointment request can take place during a decision epoch, and the same-day demand is realized after the last decision epoch and before the working day starts. The state of the system is described by a matrix whose entries correspond to the remaining capacity for each physician-time slot combination. Due to the curse of dimensionality, it is not possible to solve the problem for a reasonable number of decision epochs, physicians, and time slots. Therefore, the authors analyze the properties of optimal booking decisions and propose two heuristic methods which are based on a metric that captures the popularity of each physician-time slot combination. For each combination, the metric is computed by dividing the expected number of times that the combination will be in the preferred set of combinations by the remaining capacity for that combination. It is observed from the results of their extensive numerical study that the solutions returned by heuristic methods are significantly better than the ones given by the policy that books the first available PCP-time slot combination for each arriving patient.

Wang and Gupta (2011) also illustrate the impact of using correct estimates for patient preferences by comparing the results of two experiments executed based on correct estimates and naive estimates (which assumes that every physician-time slot combination is acceptable to each patient), respectively. Their results reveal that using correct estimates is likely to bring more benefit for the clinics where the physicians' workload is imbalanced. They also present the results of experiments carried out under different levels of patient no-show probabilities and service time variability, and they conclude from their results that using the proposed heuristic methods is reasonable when no-show probabilities are not too high and the service time variability is not more than the variability observed in the empirical studies.

When investigating the problem of booking appointments where the patients are modeled as different revenue classes and the patient preferences are taken into consideration, both Gupta and Wang (2008) and Wang and Gupta (2011) assume that the available time slots and the corresponding capacities are given (i.e., the available physician time on each working day has already been divided into time slots of desired/required lengths). The problem of determining the clinic capacity and constructing the time slots has been studied by a number of researchers, and the review of relevant studies can be found in Gupta and Wang (2012).

Scheduling in the presence of multiple patient groups with different priorities is also the focus of several other optimization-based studies (e.g., Erdelyi and Topaloglu (2010), Gocgun et al. (2011), Kolisch and Sickinger (2008), Min and Yih (2010)). We provide the details of the problem studied by Patrick et al. (2008) and their modeling and approximate dynamic-programming-based solution approach in Sect. 4.1.

## 4.1 Dynamic Multipriority Patient Scheduling

In this section, we introduce the problem of scheduling patients in multiple priority classes on a diagnostic resource studied by Patrick et al. (2008). A diagnostic resource (such as a computer tomography (CT) scanner or a magnetic resonance imaging (MRI) scan) serves patients in different priority classes that are formed based on the urgency of the service, which is reflected in the priority-specific wait time targets. Emergency patients and inpatients may be either higher- or lower-priority patients depending on how soon they need to be served, which typically ranges from "immediate" to "24 h." Outpatients, on the other hand, are typically lower-priority patients who need to be served within a larger time window ranging from days to weeks. Every day, the available capacity over the considered booking horizon should be allocated between the priority classes in such a way that the wait time targets are met to the greatest possible extent. Booking lower-priority patients too soon may result in insufficient capacity for later-arriving higher-priority patients. On the other hand, booking them too far into the future may create idle capacity.

Patrick et al. (2008) formulate this problem as a discounted infinite-horizon MDP and solve it using LP-based *approximate dynamic programming* (ADP) methods. We now present the details of their approach.

### Model Formulation

Daily decision epochs are considered over an $N$-day rolling horizon. Due to the dynamic structure of the booking horizon, day $n$ at the current decision epoch becomes day $n-1$ in the next decision epoch. Therefore, at the beginning of each day, there are no scheduled appointments on the $N$th day as implied by the length of the booking horizon. Although it is finite, the booking horizon is rolling and hence evolving day by day, which leads to an infinite-horizon problem where the capacity allocation decisions are made every day.

The state of the system is represented by $\mathbf{s} = (\mathbf{x}, \mathbf{y}) = (x_1, x_2, \ldots, x_N; y_1, y_2, \ldots, y_I)$, where $x_n$ and $y_i$ denote the number of patients already booked on day $n$ and the number of priority $i$ patients waiting to be booked, respectively. There are $I$ patient priorities, and 1 is the top priority. Based on the state definition, the state space can be written as $S = \{(\mathbf{x}, \mathbf{y}) | x_n \leq C_1, 1 \leq n \leq N; 0 \leq y \leq Q_i, 1 \leq i \leq I; (\mathbf{x}, \mathbf{y}) \in \mathbb{Z}_N \times \mathbb{Z}_I\}$, where $C_1$ is the daily capacity of the diagnostic resource and $Q_i$ is the maximum number of priority $i$ patients that can arrive in a day. At each decision epoch, the assignment of the available appointment slots to the waiting patients is to be determined. It is assumed that postponing the scheduling to the next day, or diverting the patients to a surge capacity (overtime or outsourcing), is also viable where the number of patients diverted per day is limited by $C_2$. Considering these available actions, and letting $a_{in}$ and $z_i$ be the number of priority $i$ patients to be booked on day $n$ and to be diverted to the surge capacity, the vector of possible actions is denoted by $(\mathbf{a}, \mathbf{z}) = \{a_{in}, z_i\}$. An action is feasible if it meets the requirements (4.45)–(4.48) given below:

- The daily capacity of the diagnostic resource is not exceeded:

$$x_n + \sum_{i=1}^{I} a_{in} \leq C_1 \quad \forall n \in 1, \ldots, N. \tag{4.45}$$

- The upper limit on the number of diversions is not exceeded:

$$\sum_{i=1}^{I} z_i \leq C_2. \tag{4.46}$$

- The total number of bookings and the diversions for priority class $i$ does not exceed the number of waiting priority $i$ patients:

$$\sum_{n=1}^{N} a_{in} + z_i \leq y_i \quad \forall i \in 1, \ldots, I. \tag{4.47}$$

- The number of bookings and diversions is positive integers:

$$(\mathbf{a},\mathbf{z}) \in \mathbb{Z}_{IN} \times \mathbb{Z}_I. \tag{4.48}$$

Upon the implementation of action $(\mathbf{a},\mathbf{z})$, the system transitions from state $(x_1,x_2,\ldots,x_N;y_1,y_2,\ldots,y_I)$ to state $(x_2+\sum_{i=1}^I a_{i2},\ldots,x_N+\sum_{i=1}^I a_{iN},0;y'_1+y_1-\sum_{n=1}^N a_{1n}-z_1,\ldots,y'_I+y_I-\sum_{n=1}^N a_{In}-z_I)$ with probability $p(\mathbf{y}')=\prod_{i=1}^I p(y'_i)$ where $\mathbf{y}'$ denotes the number of patient arrivals in a day. As implied by the definition of $p(\mathbf{y}')$, the arrivals of patients in different priority classes are independent.

The cost associated with action $(\mathbf{a},\mathbf{z})$ is $c(\mathbf{a},\mathbf{z})=\sum_{i,n} b(i,n)a_{in}+\sum_{i=1}^I d(i)z_i+\sum_{i=1}^I f(i)(y_i-\sum_{n=1}^N a_{in}-z_i)$, where $b(i,n)$, $d(i)$, and $f(i)$ are the unit costs of booking a priority $i$ patient on day $n$, diverting a priority $i$ patient, and delaying the booking of a priority $i$ patient for one day, respectively. Cost coefficients $b(i,n)$ and $f(i)$ are intuitively decreasing in $i$, and the structure of $d(i)$ depends on the type of the available surge capacity. If the surge capacity is in the form of overtime, then it is reasonable to assume that $d(i)$ is constant in $i$. If diversion means rejecting the patient and sending him/her to another healthcare delivery provider, then $d(i)$ should be decreasing in $i$.

The priority-specific wait time target, which is denoted by $T(i)$, is increasing in $i$. $b(i,n)$ should be increasing in $n$, and it should be positive only if booking the appointment of a priority $i$ patient on day $n$ does not meet the wait time target, i.e., $n>T(i)$. Moreover, it is reasonable to set $b(i,n)$ in a way that the cost of delaying a patient $k$ days and then booking his/her appointment on a day within $T(i)$ is equal to the cost of booking him/her $k$ days late initially. Based on these assumptions, $b(i,n)$ can be chosen as $\sum_{k=1}^{n-T(i)} \gamma^{k-1} f(i)\ \forall n>T(i)$ and as 0 otherwise, where $\gamma$ is the daily discount factor.

Considering the components described above, the multipriority patient scheduling problem can be formulated as an infinite-horizon MDP with the following optimality equations for all $(\mathbf{x},\mathbf{y})\in S$:

$$\begin{aligned} v(\mathbf{x},\mathbf{y}) = \min_{(\mathbf{a},\mathbf{z})\in A_{\mathbf{x},\mathbf{y}}} \Bigg\{ c(\mathbf{a},\mathbf{z}) + \gamma \sum_{\mathbf{y}'\in D} p(\mathbf{y}')v\Bigg( & x_2+\sum_{i=1}^I a_{i2},\ldots,x_N+\sum_{i=1}^I a_{iN},0; \\ & y'_1+y_1-\sum_{n=1}^N a_{1n}-z_1,\ldots,y'_I+y_I-\sum_{n=1}^N a_{In}-z_I\Bigg)\Bigg\}, \end{aligned} \tag{4.49}$$

where $v(\mathbf{x},\mathbf{y})$ is the value function, $A_{\mathbf{x},\mathbf{y}}$ is the set of feasible actions defined by (4.45)–(4.48), and $D$ is the set of all possible patient arrival streams.

### Solution Approach

The size of the state space, given by $(C_1)^N \prod_{i=1}^I Q_i$, is very large for realistic values of $C_1, N, I$, and $Q_i$. For a problem instance where $N=60$, $C_1=30$, $I=3$, and

$Q_1 = Q_2 = Q_3 = 10$, the size of the state space is $30^{60} \times 10^3$. Therefore, standard solution methods such as the value iteration or the policy iteration fail to solve (4.49) for practical instances.

ADP-based approaches, which have been developed to overcome *the curse of dimensionality* in solving MDPs, typically consider a specific class of value functions and find an approximate optimal solution within that class. To find an approximate solution to (4.49), Patrick et al. (2008) employ an LP-based analytical method (Adelman 2004; de Farias and Van Roy 2003; Schweitzer and Seidmann 1985) with the following steps:

1. The model described by the optimality equations (4.49) is reformulated as

$$\max_{\mathbf{v}} \sum_{(\mathbf{x},\mathbf{y})\in S} \alpha(\mathbf{x},\mathbf{y})v(\mathbf{x},\mathbf{y}) \tag{4.50}$$

s.t.

$$c(\mathbf{a},\mathbf{z}) + \gamma \sum_{\mathbf{y}'\in D} \left[ p(\mathbf{y}')v\left( x_2 + \sum_{i=1}^{I} a_{i2}, \ldots, x_N + \sum_{i=1}^{I} a_{iN}, 0; \right.\right.$$
$$\left.\left. y'_1 + y_1 - \sum_{n=1}^{N} a_{1n} - z_1, \ldots, y'_I + y_I - \sum_{n=1}^{N} a_{In} - z_I \right)\right] \geq v(\mathbf{x},\mathbf{y})$$
$$\forall (\mathbf{a},\mathbf{z}) \in A_{\mathbf{x},\mathbf{y}} \quad \text{and} \quad (\mathbf{x},\mathbf{y}) \in S, \tag{4.51}$$

where $\alpha$ is strictly positive. Although not necessary for the validity of the conversion, it is assumed that $\alpha$ is the probability distribution over the initial state of the system. Note that (4.50)–(4.51) is a large-scale LP with a variable for each state, and a constraint for each state-action pair.

2. The value function is approximated by using the following linear combination of basis functions:

$$v(\mathbf{x},\mathbf{y}) = W_0 + \sum_{n=1}^{N} V_n x_n + \sum_{i=1}^{I} W_i y_i, \tag{4.52}$$

where $V_n$ and $W_i$ are nonnegative and $W_0$ is unrestricted in sign. $V_n$ and $W_i$ can be interpreted as the marginal infinite-horizon discounted costs of an occupied appointment slot on day $n$ and a priority $i$ patient waiting to be booked, respectively.

3. An *approximate linear program* (ALP) is created by substituting (4.52) into (4.50) and (4.51). This reformulation reduces the number of variables from $(C_1)^N \prod_{i=1}^{I} Q_i$ to $N + I + 1$.
4. Column generation is employed to solve the dual of the ALP, which has a reasonable number of constraints and a large number of variables. Then, the best linear value function approximation (denoted by $v_{\text{ALP}}$) is obtained by determining the coefficients ($\mathbf{V}$ and $\mathbf{W}$) based on the shadow prices in the optimal solution.

5. The *approximate optimal policy* (AOP) is extracted from the ALP solution by substituting $v = v_{\mathrm{ALP}}$ into the optimality Equations (4.49) and then solving the resulting equations for $(\mathbf{a},\mathbf{z}) \in A_{\mathbf{x},\mathbf{y}}$ where $(\mathbf{x},\mathbf{y}) \in S$ to obtain the best decision for every possible state. This corresponds to solving the following IP:

$$\min_{(\mathbf{a},\mathbf{z})\in A_{\mathbf{x},\mathbf{y}}} \left\{ \sum_{i=1}^{I}\sum_{n=1}^{N} b(i,a)a_{in} + \sum_{i=1}^{I}\left( d(i)z_i + f(i)\left( y_i - \sum_{n=1}^{N} a_{in} - z_i \right)\right) \right.$$
$$+\gamma \sum_{\mathbf{y}\in D} p(\mathbf{y}) \left[ W_0 + \sum_{n=1}^{N-1} V_n \left( x_{n+1} + \sum_{i=1}^{I} a_{i,n+1} \right) \right.$$
$$\left.\left. + \sum_{i=1}^{I} W_i \left( y'_i + y_i - \sum_{n=1}^{N} a_{in} - z_i \right)\right]\right\}$$
$$= \min_{(\mathbf{a},\mathbf{z})\in A_{\mathbf{x},\mathbf{y}}} \left\{ \sum_{i=1}^{I}\sum_{n=1}^{N} (b(i,n) + \gamma V_{n-1} - f_i - \gamma W_i) a_{in} \right.$$
$$\left. + \sum_{i=1}^{I} (d(i) - f(i) - \gamma W_i) z_i \right\} + \text{a constant term.} \tag{4.53}$$

Renaming $b(i,n) + \gamma V_{n-1} - f_i - \gamma W_i$ as $A_{in}$, and $d(i) - f(i) - \gamma W_i$ as $Z_i$, the simplified form of (4.53) can be written as

$$\min_{(\mathbf{a},\mathbf{z})\in A_{\mathbf{x},\mathbf{y}}} \left\{ \sum_{i=1}^{I}\sum_{n=1}^{N} A_{in} a_{in} + \sum_{i=1}^{I} Z_i z_i \right\} + \text{a constant term.} \tag{4.54}$$

The coefficients $A_{in}$ and $Z_i$ together determine the approximate optimal action for priority class $i$. In the AOP, priority $i$ patients are booked to days for which $A_{in} < 0$ and overtime is only used for priority classes for which $Z_i < 0$.

Under reasonable conditions on the problem parameters, it is proved by Patrick et al. (2008) that $v_{\mathrm{ALP}}$ and the corresponding AOP can be obtained directly without using the techniques described in Steps 4 and 5 of the above solution method. The structure of the AOP where the considered surge capacity is overtime (i.e., diversion cost $d(i)$ is constant in $i$) is characterized by the following theorem. (We refer the reader to Patrick et al. (2008) for the details about the structure of the AOP under decreasing diversion costs, and for the structure of the $v_{\mathrm{ALP}}$ under both types of diversion costs.)

**Theorem 2 (Patrick et al. (2008), Theorem 3).** *Suppose that the diversion costs $d(i) = d$ for all $i$, the booking cost is nondecreasing in $n$ and nonincreasing in $i$ with $b(i,n) = 0$ for all $i$ and $n \le T(i)$. Suppose further that*

$$b(i,n) + \gamma^{n-T(1)} d > b(i,T(i)) + \gamma^{T(i)-T(1)} d \quad \forall i \quad \text{and} \quad \forall n > T(i), \tag{4.55}$$

$$\sum_{i=1}^{I} \frac{\gamma^{T(i)-n}}{1-\gamma} I_{T(i)>n} \lambda_i + \sum_{m=n}^{N} \sum_{\mathbf{x},\mathbf{y}} \gamma^{[m-n]^+} \alpha(\mathbf{x},\mathbf{y}) x_n < \frac{C_1}{1-\gamma} \quad \forall n \geq T(1), \tag{4.56}$$

$$0 < \sum_{i=1}^{I} \frac{\gamma^{T(i)-T(1)}}{1-\gamma} \lambda_i + \sum_{n=1}^{N} \sum_{\mathbf{x},\mathbf{y}} \gamma^{[n-T(1)]^+} \alpha(\mathbf{x},\mathbf{y}) x_n - T(1)C_1 - \frac{\gamma C_1}{1-\gamma} < \frac{C_2}{1-\gamma}, \tag{4.57}$$

*where $I_{T(i)>n}$ is an indicator variable which is equal to one if $T(i) > n$ and zero otherwise, $\lambda_i$ is the arrival rate for demand from priority class i, $C_1$ is equal to the base capacity, $C_2$ is the surge capacity (i.e., overtime), and $\gamma$ is the discount rate.*

*Moreover, for each priority class i, let $(LB(i)$ and $UB(i))$ be the interval described by*

$$LB(i) = \min\{n|A_{in} < 0\} = \min\{n|f(i) > (\gamma^{[n-T(1)-1]^+ +1} - \gamma^{T(i)-T(1)+1})d\}, \tag{4.58}$$

$$UB(i) = T(i). \tag{4.59}$$

*Then, the AOP may be implemented as follows:*

*(i) Book patients in order of priority class.*
*(ii) Book as much priority 1 demand as possible into the interval $(1, T(1))$ starting with day 1 and working up to day $T(1)$.*
*(iii) For each successive priority class, book incoming demand into any available slots in the interval $\{1 \cup (LB(i), UB(i))\}$ starting with day 1, then day $UB(i)$, and working down to day $LB(i)$.*
*(iv) If there is any remaining demand, use overtime for a given priority class only if*

$$f(i) > (1 - \gamma^{T(i)-T(1)+1})d \tag{4.60}$$

*is satisfied (i.e., only if $Z_i < 0$) and giving precedence to higher-priority demand should the overtime capacity constraint be an issue.*
*(v) For all remaining demand, delay booking.*

If the conditions given in Theorem 2 hold, then the AOP is described by only $2I + 1$ numbers (which are the $LB(i)$ and $UB(i)$ values for each priority class $i$, and the lowest-priority class for which the overtime is used) that can be obtained by using (4.58)–(4.60) (i.e., without solving the IP given by (4.53)).

As implied by its name, the optimality of the AOP is not guaranteed. Therefore, investigating the performance of the approximation is of great practical value. By using a simulation model, Patrick et al. (2008) compare the AOP to a strict booking policy where a priority $i$ patient is booked on day $n$ only if the number of available appointment slots on that day is more than a priority-specific threshold level. Their extensive numerical results reveal that the AOP performs notably well under many different problem settings characterized by the size of the healthcare delivery institution and the composition of the demand stream.

## 5 Workforce Scheduling

Workforce scheduling in healthcare involves determining the staffing needs of healthcare facilities both at the aggregate and the operational levels, and designing the staff schedules accordingly. The majority of the related literature focuses on nurse scheduling problems (e.g., Kao and Queyranne (1985), Azaiez and Al Sharif (2005), Beliën and Demeulemeester (2008), Maenhout and Vanhoucke (2010)) which typically include decisions such as determining the number of nurses to be employed, assigning nurses to a set of shifts in such a way that feasibility requirements are met and an appropriate staffing level is achieved, and assigning patients to nurses at the beginning of shifts. We introduce some common nurse scheduling problems by providing examples from the literature. We refer the reader to Burke et al. (2004), Cheang et al. (2003), and Sitompul and Randhawa (1990) for extensive reviews of the related studies.

Lavieri and Puterman (2009) propose a *linear program* (LP) to model a workforce planning problem at the provincial level (in the province of British Columbia, Canada). The decisions in their model include the number of students to admit to different types of nursing programs, the number of nurses and managers to recruit from outside of the province, and the number of nurses to promote to managerial positions in each year over a 20-year planning horizon. The constraints of the model represent several requirements, some of which are as follows: the number of students should be within the bounds imposed by the size of the nursing programs, the number of nurses should be enough to meet the population's need, the ratio of nurses to managers should be greater than a prespecified minimum level, the number of students admitted to different types of programs should be balanced, the number of nurses to recruit should not exceed the number of available nurses allowed to work in the province, and the number of years of experience should be taken into account when promoting the nurses. The objective of the model is to determine the workforce size in each year that minimizes the total cost of training (educating students and promoting nurses), recruitment, and annual salaries over the considered planning horizon. The model carefully captures the dynamics of the system by explicitly considering the aging of students and nurses; the students progress in their degree program or graduate unless they quit, and the nurses become more experienced unless they use a parental leave or retire. Implementation of the proposed model reveals that a feasible solution for the workforce scheduling of British Columbia could be obtained only by having a large initial recruitment or by changing nurse to population ratios (i.e., by degrading the quality of service).

Purnomo and Bard (2007) construct cyclic biweekly schedules for nurses by using a MIP which minimizes a weighted sum of the number of uncovered shifts (i.e., the shifts covered by outside nurses) and the violation of soft constraints that represent the preferences of nurses. The hard constraints of the model describe a set of legal and institutional requirements associated with the maximum amount of time a nurse is allowed to work during a day/week, and minimum amount of break between the consecutive shifts of a nurse. The authors develop a branch-and-price algorithm to solve large instances of the problem and enhance their algorithm by using two different branching rules (subproblem variable branching and master problem variable branching) and an effective rounding heuristic. We observe from their numerical results that the proposed algorithm can solve large problem instances (i.e., instances with up to 200 nurses) within a reasonable amount of time. The reported results also indicate that subproblem variable branching is more suitable for small and medium size problems, whereas master problem variable branching performs better for large size problems.

Punnakitikashem et al. (2008) study an operational-level nurse scheduling problem which considers the patient-to-nurse assignment decisions at the beginning of a shift. They explicitly consider the uncertainty in the amount of care required by each patient and formulate the problem as a two-stage SMIP where the objective is to minimize the expected excess workload on nurses (i.e., the difference between the workload and the available time for care). We present the details of their approach and briefly discuss their results in Sect. 5.1.

Scheduling physicians in an emergency department shares significant similarities with the problem of designing a shift schedule for nurses, and is the focus of a number of studies including Beaulieu et al. (2000), Brunner et al. (2009), Carter and Lapierre (2001), Rousseau et al. (2002), Vassilacopoulos (1985).

Resident scheduling is another commonly studied problem in the context of healthcare workforce planning. A recent study by Cohn et al. (2009) explores the problem of scheduling on-call shifts of residents in a multihospital setting over a 365-day planning horizon. A desirable schedule must ensure that each hospital has an appropriate level of residents each night in order to meet patient needs, each resident completes a specified number of calls at each hospital during the year in order to meet his/her educational requirements, and each resident has an acceptable amount of rest in order to avoid fatigue that may result in poor quality service. Moreover, residents' preferences on their schedule should be taken into consideration as much as possible. The authors formulate the problem as an IP that seeks a feasible solution with respect to these requirements, and then expand their model to consider different objectives that are important to the residents. The number of vacation requests denied, number of night calls that do not match a resident's daytime location, and number of Friday/Saturday calls are some of the considered metrics. The authors successfully applied their model to schedule the chief residents in the psychiatry program at Boston University School of Medicine (BUSM), and the proposed schedule became effective shortly after the completion of the project.

## *5.1 Nurse Assignment Problem*

In this section, we describe the nurse assignment problem studied by Punnakitikashem et al. (2008), present the SMIP formulation introduced by the authors, and briefly summarize the insights revealed by their numerical results.

Prior to the beginning of a shift, each patient needs to be assigned to a nurse, and the assignment decisions are typically made by a *charge nurse*. Due to the regulations, there may be some restrictions on the assignment decisions (e.g., assigning a particular patient to a specific type of nurse may not be feasible). It is assumed that the charge nurse determines which nurses can be assigned to which patients before the assignment problem is solved.

It is very rare in practice that a patient is reassigned to another nurse during the shift. Therefore, it is assumed that there is no reassignment during the shift. However, the assignment decisions are dynamically updated due to the admission of new patients and the discharge of some of the existing patients. Admission and discharge events are captured explicitly by taking the probability of their occurrence into account when modeling the amount of care required by the patients. As a result of this approach, the set of considered patients include the potential patients as well as the existing ones, and hence the authors are able to consider a fixed number of patients in their model.

There are two types of care provided by nurses: *direct care*, which is the amount of time spent with patients, and *indirect care*, which is the amount of time spent on other tasks. The amount of required care during the shift is modeled by considering the shift as a collection of several smaller time periods. The required indirect care in a time period can be provided any time during the shift, whereas the required direct care should be provided within that time period.

As the workload of a nurse increases, her/his patients receive less care. Therefore, excess workload on nurses has a negative impact on the quality of care. To reflect this, the assignment cost is modeled as a nondecreasing piecewise linear convex function of the amount of assigned workload.

We now present the notation used in Punnakitikashem et al. (2008).

***Sets and Indices***

$\omega$: Scenario index where $\Omega$ represents the set of random scenarios

$\tau, t$: Time period indices where $T$ represents the set of time periods in the shift

$n$: Nurse index where $N$ represents the set of all nurses

$p$: Patient index where $P$ represents the set of all patients

$N(p)$: Set of nurses that can be assigned to patient $p$

$P(n)$: Set of patients that can be assigned to nurse $n$

$i$: Piece/interval index such that $i \leq k$ where $k$ is the number of pieces/intervals in the considered piecewise cost function

***Parameters***

$m_{\tau ni}, m_{\tau n(i+1)}$: Lowest and highest levels of workload that describe the $i$th interval i.e., the interval for which the associated cost is given by the $i$th piece of the cost function)

$\alpha_{\tau ni}$: Marginal cost associated with the $i$th piece of the cost function, (because the considered cost function is nondecreasing and convex, $0 = m_{\tau n1} < \cdots < m_{\tau nk}$ and $0 \leq \alpha_{\tau n1} < \cdots < \alpha_{\tau nk}$. For notational convenience, $m_{\tau n(k+1)}$ is $\infty$.)

$\phi(\omega)$: Probability that scenario $\omega$ occurs
$d_{tp}(\omega)$: Amount of direct care required by patient $p$ in time period $t$
$g_{tp}(\omega)$: Amount of indirect care required by patient $p$ at the beginning of time period $t$ until the end of the shift
$\xi(\omega)$: Random vector composed of $d_{tp}(\omega)$ and $g_{tp}(\omega)$ parameters for all patients and time periods

***First-Stage Decision Variables***

$X_{pn}$: Binary decision variable denoting patient $p \in P$ is assigned to nurse $n \in N(p)$

***Second-Stage Decision Variables***

$A_{\tau ni}(\omega)$: Amount of workload assigned to nurse $n$ between time durations $m_{\tau ni}$ and $m_{\tau n(i+1)}$ in scenario $\omega$
$G_{t\tau n}(\omega)$: Total indirect care to be performed during or after time period $t$ and is performed in time period $\tau$ by nurse $n$ in scenario $\omega$

Considering the above notation, and letting $X$ be the vector form of the first-stage decision variables, two-stage SMIP formulated by Punnakitikashem et al. (2008) is as follows:

$$\min \quad \mathscr{Q}(X) \tag{4.61}$$

s.t.

$$\sum_{n \in N(p)} X_{pn} = 1 \forall p \in P, \tag{4.62}$$

$$X_{pn} \in \{0,1\} \forall p \in P(n), n \in N, \tag{4.63}$$

where the recourse function $\mathscr{Q}(X)$ is defined as

$$\mathscr{Q}(X) = E_{\xi}\left[Q(X,\xi(\omega))\right] \tag{4.64}$$

and

$$Q(X,\xi(\omega)) = \min \quad \sum_{n \in N} \sum_{\tau \in T} \sum_{i=1}^{k} \alpha_{\tau ni} A_{\tau ni}(\omega) \tag{4.65}$$

s.t.

$$\sum_{p \in P(n)} g_{tp}(\omega) X_{pn} = \sum_{\tau=t}^{|T|} G_{t\tau n}(\omega) \quad \forall t \in T, n \in N, \tag{4.66}$$

$$\sum_{p \in P(n)} d_{tp}(\omega) X_{pn} + \sum_{t=1}^{\tau} G_{t\tau n}(\omega) = \sum_{i=1}^{k} A_{\tau ni}(\omega) \quad \forall \tau \in T, n \in N, \tag{4.67}$$

$$A_{\tau ni}(\omega) \leq m_{\tau n(i+1)} - m_{\tau ni} \quad \forall \tau \in T, 1 \leq i \leq k, n \in N, \tag{4.68}$$

$$G_{t\tau n}(\omega) \geq 0 \quad \forall t, \tau \in T, t \leq \tau, n \in N, \tag{4.69}$$

$$A_{\tau ni}(\omega) \geq 0 \quad \forall \tau \in T, 1 \leq i \leq k, n \in N. \tag{4.70}$$

The objective function (4.61) is the total expected cost of assignment. Due to the structure of the objective function, minimizing the total expected cost is equivalent to minimizing the expected excess workload on nurses. Constraint (4.62) ensures that each patient is assigned to a nurse.

For a given first-stage solution $X$ and a random scenario $\omega$, the second-stage problem is described by (4.65)–(4.70). Constraint (4.66) determines the total indirect care performed by each nurse in each time period. Constraint (4.67) describes the total workload of each nurse in each period as the sum of workload resulting from direct and indirect care to be performed, and constraint (4.68) computes the marginal workload variables associated with each interval.

Punnakitikashem et al. (2008) use the L-shaped algorithm to solve (4.61)–(4.70) and increased the problem solvability by using a polynomial time procedure to solve the second-stage problem, deriving lower and upper bounds on the number of patients that can be assigned to a nurse and adding valid inequalities that reduce symmetry with respect to nurses. The authors execute experiments by using instances generated based on real data provided from Baylor Regional Medical Center in Grapevine, Texas. Because the considered instances are not solvable within a reasonable amount of time, the authors use the algorithm to obtain a good solution within a time limit of 30 min rather than finding the optimal solution. They compare the expected excess workload achieved by the solution returned by the L-shaped algorithm to that of the solutions obtained by different alternative approaches, which are solving the mean value problem, using a heuristic that balances workloads based on the expected required total care, and random assignment. Their numerical results show that the L-shaped algorithm substantially reduces the expected excess workload over the other methods. We also observe from their results that the expected value of perfect information is high when the set of nurses scheduled for the shift do not have sufficient time and skills to care for the patients, and the value of the stochastic solution is particularly high when there is significant variance in the required care by the patients.

## 6 Other Applications

Besides the application areas discussed in earlier sections, optimization-based approaches have been used to address several other problems in healthcare including healthcare facility location, organ allocation and transplantation, vaccine design and disease screening (e.g., Ayvaci et al. (2012), Demirci et al. (2011), Özaltın et al. (2011), Ingolfsson et al. (2008), Jacobs et al. (1996), Kurt et al. (2011), Kong et al. (2010), Maxwell et al. (2010), Segev et al. (2005), Zhang et al. (2010)). In this section, we briefly describe some of the problems considered in recent studies, and discuss the methods employed by the authors.

Zhang et al. (2010) study location and capacity allocation decisions associated with preventive healthcare facilities over a network of geographic population zones, where the objective is to maximize the accessibility for potential patients. Ease of access to the facilities has a positive influence on the participation in preventive healthcare programs, which play a crucial role in prevention and early detection of diseases. The authors formulate the problem as a bilevel nonlinear optimization model, where the facility location and capacity allocation (allocation of a number of total available servers to facilities) decisions are made at the upper level. Given the upper level decisions, the nonlinear lower level problem describes the nature that clients choose the facility with the minimum expected total time of providing the care, which is comprised of the travel time to the facility, and the waiting and service time at the facility. The model captures the assumption that the participation rate at each population zone decreases with the increasing expected total time, and the requirement that a facility can be opened only if the number of its potential clients is more than a prespecified minimum workload (which is necessary to ensure the quality of the service). The problem is difficult to solve due to the existence of binary variables at the upper level and the nonlinearity at the lower level. Therefore, the authors develop a heuristic method to solve practical instances of the problem. Within the proposed solution framework, the upper level problem is solved with a tabu search procedure, and the lower level problem is solved with the gradient projection method (which is an exact solution algorithm for convex optimization problems) or an approximation algorithm depending on the number of servers allocated to each facility. For their numerical study, the authors consider the problem of designing a network of mammography centers in Montreal, Quebec. It is observed from their results that while centralizing the capacity at high-density zones to raise the participation is a favorable strategy when the available number of servers is medium/large, increasing the spatial coverage becomes more important, and hence, decentralizing the capacity to different zones becomes a better strategy, when the available number of servers is limited.

Segev et al. (2005) propose an optimization-based approach to improve the number and quality of living-donor kidney transplants. Living-donor kidney transplantation for a donor–recipient pair is a viable option only when the donor is willing, healthy, and compatible with the recipient. An incompatible donor–recipient pair may still benefit from transplantation if there exists another pair in the same situation such that the pairs become compatible when the donors are exchanged, which is known as the kidney paired donation (KPD). Segev et al. (2005) compare the outcomes associated with two different KPD matching schemes, which are the first-accept matching and optimized matching, by applying these schemes to the simulated pools of incompatible donor–recipient pairs. In the first-accept matching scheme, the first pair on the list is picked, and the pool is searched to find a matching pair. Once such a pair is found, both pairs are removed from the list since they form an acceptable match. Matching continues in the same manner, until no acceptable match is detected. In the optimized matching scheme, all acceptable matches in the pool are identified, all feasible solutions (i.e., combination of acceptable matches), are compared and the one with the best outcome is picked.

The numerical results presented by the authors indicate that the optimized matching scheme performs significantly better than the first-accept matching scheme, and the performance gap becomes more prominent with the increasing pool size. It is observed from their results that utilizing the optimized matching scheme at the national level could achieve matching 47.7% of the incompatible pairs in the USA, and this would require only 2.9% of the pairs to travel. Based on these results, the authors conclude that establishing a national level KPD program and optimizing the matching decisions are extremely important. (It is worth mentioning that a national KPD pilot program in the USA began in 2010 (OPTN 2010).)

Kurt et al. (2011) explore transplant timing decisions in a prearranged kidney exchange that involves an arbitrary number of incompatible patient–donor pairs for whom the only feasible cross-exchange of kidneys is a cyclic exchange. The patients' health statuses probabilistically change according to a discrete-time finite-state Markov chain, and at each decision epoch, each patient decides between offering to exchange and waiting. The process continues until an exchange occurs (which only happens if all patients offer to exchange) or a patient dies (which leaves no feasible solution for the problem). Each patient receives a terminal reward, if an exchange occurs, receives an intermediate reward and moves to the next decision epoch otherwise. The timing of the exchange depends on the choices of all patients, and hence, a patient cannot simply optimize his/her own objective independent of the choices of the other patients. The authors formulate the problem as a nonzero sum stochastic game and use a MIP to find the equilibrium that maximizes the sum of the patients' total expected payoffs. Their numerical results reveal that designing the cross-exchange of kidneys without considering the timing decisions may result in suboptimal solutions in terms of social welfare, and matching patients at similar stages of disease results in better solutions.

Ayvaci et al. (2012) investigate the impact of budgetary restrictions on the breast cancer diagnostic decisions. After a patient is screened with mammography, radiologists assess her risk of cancer and then choose one of the following three options considering the risk level: (1) routine follow-up mammography, (2) short-term follow-up mammography, and (3) biopsy (i.e., immediate diagnostic action). Sending the patient to biopsy when her risk level is too low may result in a false positive, whereas being too late in taking the immediate diagnostic action may hinder early detection of cancer, which is the main trade-off to be balanced. The health state of the patient probabilistically evolves, and is described by a risk score, or one of the two absorbing states: death and post-cancer. The authors formulate the problem as a finite-horizon discrete-time constrained MDP that maximizes the total quality adjusted life years (QALYs) of the patient under budget-related constraints. They derive the equivalent LP formulation of the constrained MDP model and then convert it to a MIP by adding binary variables that restrict the set of feasible solutions to deterministic policies. The authors solve the problem under different budget levels in order to estimate the incremental cost-effectiveness ratios. Their comparison to the actual clinical practice indicates that optimal policies may achieve the same level of QALYs under significantly lower budget levels. Their numerical results also reveal that eliminating short-term follow-ups becomes favorable as the budget level decreases. It is also observed from their results that an increase in the budget has higher impact (in terms of gained QALYs) on younger patients.

Özaltın et al. (2011) propose a multistage SMIP that integrates the two important decisions regarding flu shot design: the composition of the vaccine (i.e., which strains to include in the vaccine) and the timing of composition decisions. These decisions are made before the flu season starts, under uncertainty associated with the flu season and the vaccine production process. The vaccine is comprised of a number of strains, which are chosen among a set of candidate strains recommended by the World Health Organization based on the available surveillance data and epidemiological analysis. The composition of a vaccine determines its effectiveness; when there is a good match between the selected strains and the ones that emerged in the flu season, the vaccine is highly effective. Timing of composition decisions is crucial as it has a direct impact on the production start times for strains, and hence on the flu shot availability. Making the composition decisions too early (which may result in a low vaccine effectiveness due to insufficient surveillance data) or too late (which may result in delays or shortages in flu shot supply) is not desirable; the trade-off between the effectiveness and the availability should be balanced. The authors formulate the problem as a multistage SMIP where the composition decisions are made in all stages except for the last stage, and the resolution of uncertainty occurs in the last stage. The objective of their model is to maximize the prevented cost of flu cases minus the vaccine shortage cost. The authors solve the problem by using a branch-and-price algorithm where the pricing problems are formulated and solved as a DP. Their results show that the value of integrating timing and composition decisions is significant, and provide some important policy insights.

## 7 Conclusion

In this chapter, we surveyed many recent successful optimization-based studies of healthcare operations management. Optimization has been widely used by several researchers to model healthcare problems such as appointment scheduling, operating room scheduling, capacity planning, and workforce scheduling. We pointed out many of the related studies and presented detailed modeling examples in Sects. 2–5. In Sect. 6, we reviewed some recent research in other areas including healthcare facility location, organ allocation and transplantation, disease screening, and vaccine design.

Due to the complexity of the problems in healthcare, a common modeling approach has been to consider different levels of decisions, resources, stakeholders, and other components/aspects separately. Building and solving integrated models, which would provide more realistic representations of the problems and enable better optimization, is an open research challenge. Some examples from surgery scheduling would be simultaneous determination of the surgery sequence in an operating room and the start times for surgeries, simultaneous consideration of advance and real-time decisions (i.e., day-of-surgery rescheduling activities), and concurrent planning of different resources such as operating rooms and recovery rooms. An example from organ allocation and transplantation would be building integrated models that consider both the societal and the patient perspective. Incorporating the transplant timing decisions of patients into the models used to obtain organ allocation policies would be of great practical value. There are several other important practical research questions waiting to be explored, and hence, the use of optimization methods to solve healthcare operations management problems continues to be a major topic of investigation.

## Appendix

**Property 4 Appointment vector integrality:** [Begen and Queyranne (2011), Theorem 5.1] If the processing times are integer-valued random variables, then there exists an optimal appointment vector which is integer.

*Sketch of Proof.* Let $\mathscr{A}$ denote the set of optimal appointment vectors in $\mathscr{K}$. $\mathscr{A}$ is nonempty, bounded, and closed by Lemma 4.3 of Begen and Queyranne (2011), and $F$ is continuous by Lemma 4.2 of Begen and Queyranne (2011).

For $\mathbf{A} \in \mathscr{A}$, let $I(\mathbf{A})$ be a function of $\mathbf{A}$ described as

$$I(\mathbf{A}) = \begin{cases} \min\{A_j : j \in \{2,\dots,n+1\} \quad \text{and} \quad A_j \notin \mathbb{Z}\} & \text{if } \mathbf{A} \notin \mathbb{Z}^{n+1}, \\ n\overline{p}_{\max} + 1 & \text{if } \mathbf{A} \in \mathbb{Z}^{n+1}. \end{cases}$$

$I(.)$ is upper semicontinuous on $\mathscr{A}$ (see the proof of Theorem 5.1 of Begen and Queyranne (2011) for details). Therefore, there exists an appointment schedule $\mathbf{A}^* \in \mathscr{A}$ that maximizes $I(.)$.

Assume that $\mathbf{A}^* \notin \mathbb{Z}^{n+1}$. Let $A^*_f$ be the first noninteger component of $\mathbf{A}^*$ and $J$ be the set of all jobs $j \geq f$ such that $A^*_j$ and $A^*_f$ have the same fractional part (i.e., $A^*_j - A^*_f$ is integer). Let $\varphi(x)$ denote the distance of $x$ to the nearest integer (i.e., $\varphi(x) = \min\{x - \lfloor x \rfloor, \lceil x \rceil - x\}$). Let $\Delta_1$, $\Delta_2$, and $\Delta$ be positive scalars described by

$$\Delta_1 = \frac{1}{4}\min\{\varphi(A^*_j - A^*_k) : j \in J, k \notin J\} > 0,$$

$$\Delta_2 = \frac{1}{4}\min\{\varphi(A^*_j - A^*_k) : j \notin J, k \notin J, A^*_j - A^*_k \notin \mathbb{Z}\} > 0,$$

$$0 < \Delta < \frac{1}{4}\min\{\Delta_1, \Delta_2\}.$$

Let $\mathbf{A}'$ and $\mathbf{A}''$ be two new appointment schedules constructed from $\mathbf{A}$ by subtracting and adding $\Delta$ to the noninteger components (i.e., $A'_j = A^*_j - \Delta$ and $A''_j = A^*_j + \Delta$ if $j \in J$, $A'_j = A^*_j$ and $A''_j = A^*_j$ otherwise). As proven by Lemmas

5.1–5.8 in Begen and Queyranne (2011), $\Delta$ is small enough that using $\mathbf{A}'$ or $\mathbf{A}''$ instead of $\mathbf{A}$ does not create any event change, i.e., if a job is early, late, or on time in $\mathbf{A}$, then it is also early, late, or on time, respectively, in $\mathbf{A}'$ and $\mathbf{A}''$.

Because there is no event change when moving from $\mathbf{A}'$ to $\mathbf{A}''$, $F(.|p=r)$ changes linearly between $\mathbf{A}'$ and $\mathbf{A}''$ for every realization $\mathbf{r}$ of $\mathbf{p}$. Hence, $F(.)$ also changes linearly between $\mathbf{A}'$ and $\mathbf{A}''$.

By optimality, $F(\mathbf{A}^*) \leq F(\mathbf{A}')$ and $F(\mathbf{A}^*) \leq F(\mathbf{A}'')$. Because $F(.)$ changes linearly between $\mathbf{A}'$ and $\mathbf{A}''$, we must have $F(\mathbf{A}^*) = F(\mathbf{A}') = F(\mathbf{A}'')$.

Note that $\mathbf{A}'' \geq \mathbf{A}^* \geq \underline{\mathbf{A}}$ and $\mathbf{A}'' \leq \overline{\mathbf{A}}$ (because $A''_j = A^*_j + \Delta < \lceil A^*_j \rceil$ for every $i$). So, $\mathbf{A}'' \in \mathscr{K}$ and therefore $\mathbf{A}'' \in \mathscr{A}$. But $I(\mathbf{A}^*) < I(\mathbf{A}'')$ (because $I(\mathbf{A}^*) = A^*_f < A^*_f + \Delta = A''_f = I(\mathbf{A}''))$, which contradicts with the definition of $\mathbf{A}^*$ (i.e., the maximizer of $I(.)$). Therefore, there exists an integer appointment vector $\mathbf{A}^* \in \mathscr{A}$.

**Property 5. Computation of expected cost for a given appointment schedule:** [Begen and Queyranne (2011), Theorem 7.2] If the processing durations are stochastically independent and $\mathbf{A}$ is an integer appointment vector, then $F(\mathbf{A})$ may be computed in $\mathscr{O}(n^2\overline{p}^2_{\max})$ time.

*Proof.* As the first job starts on time, $S_1 = 0$ (since $A_1 = 0$) and $C_1 = p_1$. Therefore, the distribution of $C_1$ is that of $p_1$. For job $i$, such that $1 < i \leq n$, we have $S_i = \max\{A_i, C_{i-1}\}$. So, the probability distribution of $S_i$ is

$$P\{S_i = k\} = \begin{cases} 0 & \text{if } k < A_i, \\ P\{C_i \leq k\} & \text{if } k = A_i, \\ P\{C_i = k\} & \text{if } k > A_i, \end{cases} \tag{4.71}$$

where $k = 1, \ldots, n\overline{p}_{\max}$. $S_i$ is completely described by $A_1, A_2, \ldots, A_{i-1}$ and $p_1, p_2, \ldots, p_{i-1}$, and hence does not depend on $p_i$. Owing to the independence of $S_i$ and $p_i$, we have

$$P\{C_i = k\} = P\{S_i = k - p_i\} = \sum_{j=1}^{\overline{p}_i} P\{S_i = k - j\} P\{p_i = j\}. \tag{4.72}$$

$P\{C_{i-1} \leq k\} = P\{C_{i-1} = k\} + P\{C_{i-1} \leq k-1\}$, and for each $i-1$, it can be computed in $\mathscr{O}((i-1)\overline{p}_{\max})$ time. Thus, $P\{C_i = k\}$ can be computed once we have the distribution of $S_i$. For each job $i$ and value $k$, computing $P\{S_i = k\}$ (by (4.71)) and $P\{C_i = k\}$ (by (4.72)) requires a constant number of operations and $\mathscr{O}(\overline{p}_i + 1)$ operations, respectively. Therefore, obtaining the start time and completion time distributions for job $i$ requires $\mathscr{O}(n\overline{p}_{\max}\overline{p}_{\max})$ operations. The distribution of $T_i$ and $E_i$, and their expected values, can then be computed in $\mathscr{O}(n\overline{p}_{\max})$ time. Therefore, the expected cost $F(\mathbf{A})$ can be computed in $\mathscr{O}(n^2\overline{p}^2_{\max})$ time.

**Lemma 1 (Begen et al. (2012), Lemma 13).** *Let $f : \mathbb{R}^m \mapsto \mathbb{R}$ be convex, finite with a global minimizer $y^*$. Assume that there exists $\bar{f}$ such that $f \geq \bar{f} = \lambda||y - \widetilde{y}||_1$ for some $\lambda > 0$ and $\widetilde{y} \in \mathbb{R}^m$. If $\widehat{y}$ is an $\alpha$-point for $\alpha = \lambda\varepsilon/3$, then $f(\widehat{y}) \leq (1+\varepsilon)f(y^*)$, where $\varepsilon \in [0,1]$.*

*Proof.* Let $L = f(y^*)/\lambda$. Consider the norm $l_1$ ball $B = B(\widetilde{y}, L)$, then $y^* \in B(\widetilde{y}, L) = \{y : \lambda||y^* - \widetilde{y}|| \leq f(y^*)\}$. The subgradient inequality at $\widehat{y}$ combined with Cauchy–Schwartz inequality yields $f(\hat{y}) - f(y^*) \leq \alpha||\widehat{y} - y^*||_1$ (since Cauchy–Schwarz inequality also holds for $l_1$ norm). We also have

$$||\hat{y} - y^*||_1 \leq ||\widehat{y} - \widetilde{y}||_1 + ||\widetilde{y} - y^*||_1 \leq f(\hat{y})/\lambda + L = f(\hat{y})/\lambda + f(y^*)/\lambda.$$

So we obtain $f(\widetilde{y}) - f(y^*) \leq \alpha(f(\widehat{y})/\lambda + f(y^*)/\lambda)$ and hence $f(\widehat{y}) \leq f(y^*)(\lambda + \alpha)/(\lambda - \alpha)$. If we choose $\alpha \leq \lambda\varepsilon/3$, the result follows.

## References

Adan I, Bekkers J, Dellaert N, Vissers J, Yu X (2009) Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. Health Care Manag Sci 12(2):129–141

Adelman D (2004) A price-directed approach to stochastic inventory/routing. Oper Res 52(4): 499–514

Akcali E, Côté MJ, Lin C (2006) A network flow approach to optimizing hospital bed capacity decisions. Health Care Manag Sci 9(4):391–404

Ayvaci MUS, Alagoz O, Burnside ES (2012) The effect of budgetary restrictions on breast cancer diagnostic decisions. Manuf Serv Oper Manag MSOM Fall 2012 14:600–617. doi:10.1287/msom.1110.0371

Azaiez MN, Al Sharif SS (2005) A 0–1 goal programming model for nurse scheduling. Comp Oper Res 32(3):491–507

Batun S, Denton BT, Huschka TR, Schaefer AJ (2011) Operating room pooling and parallel surgery processing under uncertainty. INFORMS J Comput 23(2):220–237

Beaulieu H, Ferland JA, Gendron B, Michelon P (2000) Dynamic bid prices in revenue management. Health Care Manag Sci 3(3):193–200

Begen MA (2010) Appointment scheduling with discrete random durations and applications. Ph.D. Thesis, University of British Columbia, Vancouver, BC, 2010. http://hdl.handle.net/2429/23332

Begen MA (2011) Stochastic dynamic programming models and applications. In: Cochran JJ, Cox LA, Keskinocak P, Kharoufeh J, Smith JC (eds) Wiley Encyclopedia of operations research and management science. http://ca.wiley.com/WileyCDA/Section/id-380199.html

Begen MA, Queyranne M (2011) Appointment scheduling with discrete random durations. Math Oper Res 36(2):240–257

Begen MA, Levi R, Queyranne M (2012) A sampling-based approach to appointment scheduling. Oper Res 60(3):675–681

Beliën J, Demeulemeester E (2007) Building cyclic master surgery schedules with leveled resulting bed occupancy. Eur J Oper Res 176 (2): 1185–1204

Beliën J, Demeulemeester E (2008) A branch-and-price approach for integrating nurse and surgery scheduling. Eur J Oper Res 189(3):652–668

Ben Abdelaziz F, Masmoudi M (2012). A multiobjective stochastic program for hospital bed planning. J Oper Res Soc 63:530–538. doi:10.1057/jors.2011.39

Benders JF (1962). Partitioning procedures for solving mixed variables programming problems. Numer Math 4(1):238–252

Blake JT, Donald J (2002) Mount sinai hospital uses integer programming to allocate operating room time. Interfaces 32(2):63–73

Brunner JO, Bard JF, Kolisch R (2009) Flexible shift scheduling of physicians. Health Care Manag Sci 12(3):285–305

Burke EK, De Causmaecker P, Vanden Berghe G, Van Landeghem H (2004) The state of the art of nurse rostering. J Scheduling 7(6):441–499

Cardoen B, Demeulemeester E, Beliën J (2010) Operating room planning and scheduling: A literature review. Eur J Oper Res 201(3):921–932

Carter MW, Lapierre SD (2001) Scheduling emergency room physicians. Health Care Manag Sci 4(4):347–360

Cayirli T, Veral E (2003) Outpatient scheduling in health care: A review of literature. Prod Oper Manag 12(4):519–549

Cayirli T, Veral E, Rosen H (2006) Designing appointment scheduling systems for ambulatory care services. Health Care Manag Sci 9(1):47–58

Cheang B, Li H, Li A, Rodrigues B (2003) Nurse rostering problems - a bibliographic survey. Eur J Oper Res 151(3):447–460

Cohn A, Root S, Kymissis C, Esses J, Westmoreland N (2009) Scheduling medical residents at boston university school of medicine. Interfaces 39(1):186–195

de Farias DP, Van Roy B (2003) The linear programming approach to approximate dynamic programming. Oper Res 51(6):850–865

Demirci MC, Schaefer AJ, Romeijn HE, Roberts MS (2011) An exact method for balancing efficiency and equity in the liver allocation hierarchy. INFORMS J Comput 24:260–275. doi:10.1287/ijoc.1110.0445

Denton BT, Gupta D (2003) Sequential bounding approach for optimal appointment scheduling. IIE Trans 35(11):1003–1016

Denton BT, Viapiano J, Vogl A (2007) Optimization of surgery sequencing and scheduling decisions under uncertainty. Health Care Manag Sci 10(1):13–24

Denton BT, Miller AJ, Balasubramanian HJ, Huschka TR (2010) Optimal allocation of surgery blocks to operating rooms under uncertainty. Oper Res 58(4):802–816

Duncan IB, Noble BM (1979) The allocation of specialties to hospitals in a health district. J Oper Res Soc 30(11):953–964

Erdelyi A, Topaloglu H (2010) Approximate dynamic programming for dynamic capacity allocation with multiple priority levels. IIE Trans 43(2):129–142

Erdogan SA, Denton BT (2011a) Surgery planning and scheduling. In: Cochran JJ, Cox LA, Keskinocak P, Kharoufeh J, Smith JC (eds) Wiley Encyclopedia of operations research and management science. http://ca.wiley.com/WileyCDA/Section/id-380199.html

Erdogan SA, Denton BT (2011b) Dynamic appointment scheduling of a stochastic server with uncertain demand. INFORMS J Comput doi:10.1287/ijoc.1110.0482

Fei H, Meskens N, Chu C (2010) A planning and scheduling problem for an operating theatre using an open scheduling strategy. Comp Ind Eng 58(2):221–230

Gerchak Y, Gupta D, Henig M (1996) Reservation planning for elective surgery under uncertain demand for emergency surgery. Manag Sci 42(3):321–334

Gocgun Y, Bresnahan BW, Ghate A, Gunn ML (2011) A Markov decision process approach to multi-category patient scheduling in a diagnostic facility. Artif Intell Med 53(2):73–81

Green L (2004) Capacity planning and management in hospitals. In: Operations research and health care. Kluwer Academic, Boston pp 15–41 http://download.springer.com/static/pdf/35/bfm%253A978-1-4020-8066-1%252F1.pdf?auth66=1354561552_2bd21358bc672a54c03e5b8676bb7107&ext=.pdf

Green LV, Savin S, Wang B (2006) Managing patient service in a diagnostic medical facility. Oper Res 54(1):11–25

Guerriero F, Guido R (2011) Operational research in the management of the operating theatre: A survey. Health Care Manag Sci 14(1):89–114

Gupta D (2007) Surgical suites' operations management. Prod Oper Manag 16(6):689–700

Gupta D, Denton BT (2008) Appointment scheduling in health care: Challenges and opportunities. IIE Trans 40(9):800–819

Gupta D, Wang L (2008) Revenue management for a primary-care clinic in the presence of patient choice. Oper Res 56(3):576–592

Gupta D, Wang W (2012) Patient appointments in ambulatory care. In: Hall R (ed) Handbook of healthcare system scheduling. International series in operations research & management science, vol 168. Springer Science, pp 65–104. http://download.springer.com/static/pdf/171/bfm%253A978-1-4614-1734-7%252F1.pdf?auth66=1354560464_15c1d07ff12a9238cf2660eba8357239&ext=.pdf

Higle JL, Sen S (1991) Stochastic decomposition: An algorithm for two-stage linear programs with recourse. Math Oper Res 16(3):650–669

Hoeffding W (1963) Probability inequalities for sums of bounded random variables. J Am Stat Assoc 58(301):13–30

Ingolfsson A, Budge S, Erkut E (2008) Optimal ambulance location with random delays and travel times. Health Care Manag Sci 11(3):262–274

Jacobs DA, Silan MN, Clemson BA (1996) An analysis of alternative locations and service areas of american red cross blood facilities. Interfaces 26(3):40–50

Jebali A, Alouane ABH, Ladet P (2006) Operating rooms scheduling. Int J Prod Econ 99(1/2): 52–62

Jensen JL (1906) Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Math 30(1):175–193

Kao EPC, Queyranne M (1985) Budgeting costs of nursing in a hospital. Manag Sci 31(5):608–621

Kolisch R, Sickinger S (2008) Providing radiology health care services to stochastic demand of different customer classes. OR Spectrum 30(2):375–395

Kong N, Schaefer AJ, Hunsaker BK, Roberts MS (2010) Maximizing the efficiency of the u. s. liver allocation system through region design. Manag Sci 56(12):2111–2122

Kurt M, Roberts MS, Schaefer AJ, Unver MU (2011) Valuing prearranged paired kidney exchanges: A stochastic game approach. Working paper

Lavieri MS, Puterman ML (2009) Optimizing nursing human resource planning in British Columbia. Health Care Manag Sci 12(2):119–128

Levi R, Roundy RO, Shmoys DB (2007) Provably near-optimal sampling-based policies for stochastic inventory control models. Math Oper Res 32(4):821–838

Li X, Beullens P, Jones D, Tamiz M (2009) An integrated queuing andmulti-objective bed allocationmodel with application to a hospital in China. J Oper Res Soc 60(3):330–338

Maenhout B, Vanhoucke M (2010) Branching strategies in a branch-and-price approach for a multiple objective nurse scheduling problem. J Scheduling 13(1):77–93

Margot F (2002) Pruning by isomorphism in branch-and-cut. Math Program 94(1):71–90

Maxwell MS, Restrepo M, Henderson SG, Topaloglu H (2010) Approximate dynamic programming for ambulance redeployment. INFORMS J Comput 22(2):266–281

May JH, Spangler WE, Strum DP, Vargas LG (2011) The surgical scheduling problem: Current research and future opportunities. Prod Oper Manag 20(3):392–405

Min D, Yih Y (2010) An elective surgery scheduling problem considering patient priority. Comp Oper Res 37(6):1091–1099

Murota K (2003) Discrete convex analysis: Monographs on discrete mathematics and applications 10. Society for Industrial and Applied Mathematics, Philadelphia

Organ Procurement and Transplantation Network (OPTN) (2010) Kidney paired donation pilot program. http://optn.transplant.hrsa.gov/resources/KPDPP. Accessed 27 Apr 2012

Organisation for Economic Co-operation and Development (OPTN) (2011) Oecd health data 2011 - frequently requested data. http://www.oecd.org/dataoecd/52/42/49188719.xls. Accessed 30 Apr 2012

Ostrowski J, Linderoth J, Rossi F, Smriglio S (2011) Orbital branching. Math Program 126(1): 147–178

Özaltın OY, Prokopyev OA, Schaefer AJ, Roberts MS (2011) Optimizing the societal benefits of the annual influenza vaccine: A stochastic programming approach. Oper Res 59(5):1131–1143

Patrick J (2012) A markov decision model for determining optimal outpatient scheduling. Health Care Manag Sci 15(2):91–102

Patrick J, Begen MA (2011) Markov decision processes and its applications in healthcare. In: Yih Y (ed) Handbook of healthcare delivery systems. CRC, Boca Raton

Patrick J, Puterman ML, Queyranne M (2008) Dynamic multipriority patient scheduling for a diagnostic resource. Oper Res 56(6):1507–1525

Punnakitikashem P, Rosenberger JM, Behan DB (2008) Stochastic programming for nurse assignment. Comput Optim Appl 40(3):321–349

Purnomo HW, Bard JF (2007) Cyclic preference scheduling for nurses using branch and price. Nav Res Logist 54(2):200–220

Robinson LW, Chen RR (2009) A comparison of traditional and open-access policies for appointment scheduling. Manuf Serv Oper Manag 12:330–346. doi:10.1287/msom.1090.0270

Rousseau LM, Pesant G, Gendreau M (2002) A general approach to the physician rostering problem. Ann Oper Res 115(1–4):193–205

Ruszczyński A (1986) A regularized decomposition method for minimizing a sum of polyhedral functions. Math Program 35(3):309–333

Sandikci B, Best T, Eisenstein D, Meltzer D (2011) Managing limited inpatient bed capacity. Working paper

Santibáñez P, Begen M, Atkins D (2007) Surgical block scheduling in a system of hospitals: An application to resource and wait list management in a BC health authority. Health Care Manag Sci 10(3):269–282

Schweitzer PJ, Seidmann A (1985) Generalized polynomial approximations in markovian decision processes. J Math Anal Appl 110(2):568–582

Segev DL, Gentry SE, Warren DS, Reeb B, Montgomery RA (2005) Kidney paired donation and optimizing the use of live donor organs. J Am Med Assoc 293(15):1883–1890

Sherali HD, Smith JC (2001) Improving discrete model representations via symmetry considerations. Manag Sci 47(10):1396–1407

Sitompul D, Randhawa SU (1990) Nurse scheduling models: A state-of-the-art review. J Soc Health 2(1):62–72

Smith-Daniels VL, Schweikhart SB, Smith-Daniels DE (1988) Capacity management in health care services: Review and future research directions. Decis Sci 19(4):889–919

Stummer C, Doerner K, Focke A, Heidenberger K (2004) Determining location and size of medical departments in a hospital network: A multiobjective decision support approach. Health Care Manag Sci 7(1):63–71

Testi A, Tanfani E, Torre G (2007) A three-phase approach for operating theatre schedules. Health Care Manag Sci 10(2):163–172

van Oostrum JM, Van Houdenhoven M, Hurink JL, Hans EW, Wullink G, Kazemier G (2008) A master surgical scheduling approach for cyclic scheduling in operating room departments. OR Spectrum 30(2):355–374

Van Slyke R, Wets RJ (1969) L-shaped linear programs with applications to optimal control and stochastic programming. SIAM J Appl Math 17(4):638–663

Vassilacopoulos G (1985) Allocating doctors to shifts in an accident and emergency department. J Oper Res Soc 36(6):517–523

Velásquez R, Melo MT (2006) A set packing approach for scheduling elective surgical procedures. In: Operations Research Proceedings, vol 2005. Springer, Germany pp 425–430. http://www.amazon.com/Operations-Research-Proceedings-2005-International/dp/3540325379

Wang PP (1993) Static and dynamic scheduling of customer arrivals to a single-server system. Nav Res Logist 40(3):345–360

Wang W, Gupta D (2011) Adaptive appointment systems with patient preferences. Manuf Serv Oper Manag 13(3):373–389

Weiss EN (1990) Models for determining estimated start times and case orderings in hospital operating rooms. IIE Trans 22(2):143–150

World Health Organization (WHO) (2012) 10 facts on ageing and the life course. http://www.who.int/features/factfiles/ageing/en/index.html. Accessed 30 Apr 2012

Zhang Y, Berman O, Marcotte P, Verter V (2010) A bilevel model for preventive healthcare facility network design with congestion. IIE Trans 42(12):865–880

# Chapter 5
# Operating Room Planning and Scheduling

**Erik Demeulemeester, Jeroen Beliën, Brecht Cardoen, and Michael Samudra**

## 1 Introduction

Healthcare has a heavy financial burden for governments within the European Union as well as oversees. While growing economies and newly emerging technologies could lead us to believe that supporting our respective national healthcare systems might get less expensive over time, data show the contrary is true. For example, within the USA, the NHE (National Health Expenditure) as a share of the Gross Domestic Product (GDP) increased from 15.6% in 2004 up to 16.2% in 2009 [112]. The data suggest that increasing health expenditures are an ongoing trend with an estimated annual growth rate of 6.3%. The NHE share of the GDP in the USA is thereby being projected to hit 19.6% by the year 2019 [31].

E. Demeulemeester (✉) • M. Samudra
Faculty of Business and Economics, Department of Decision Sciences and Information Management, Research Center for Operations Management, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium
e-mail: Erik.Demeulemeester@econ.kuleuven.be; samudra@kuleuven.be

J. Beliën
Faculty of Business and Economics, Department of Decision Sciences and Information Management, Research Center for Operations Management, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

Hogeschool-Universiteit Brussel, Center for Informatics, Modeling and Simulation, Stormstraat 2, 1000 Brussel, Belgium
e-mail: Jeroen.Belien@hubrussel.be; Jeroen.Belien@econ.kuleuven.be

B. Cardoen
Faculty of Business and Economics, Department of Decision Sciences and Information Management, Research Center for Operations Management, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

Vlerick Leuven Gent Management School, Operations and Technology Management Center, Reep 1, 9000 Ghent, Belgium
e-mail: Brecht.Cardoen@vlerick.com; Brecht.Cardoen@econ.kuleuven.be

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_5, © Springer Science+Business Media New York 2013

Similarly on the European continent, even though differences exist across member states, healthcare spending as a share of the GDP is increasing, where countries that are hit hard by the global recession are most affected. For example, in Ireland, the percentage of GDP devoted to healthcare increased from 7.6% in 2004 to 9.5% in 2009. In the UK during the same time interval, a rise from 8% to 9.8% was experienced [112].

A large amount, estimated at 31% of the spending on healthcare, pertains to hospitals [32], which are consequently being pressured to reduce costs. Hospitals are expensive from the patient perspective as well. For instance, Milliman Medical Index (MMI) estimates that, for a typical family of four, 48% of the family healthcare spending involves hospital costs [104].

Within the hospital, special attention is given to ORs as they represent the largest costs and provide the largest revenues [1]. It comes as no surprise that the body of literature dealing with topics related to OR efficiency and profitability is steadily increasing. Out of the many aspects, we focus our attention on planning and scheduling problems and do not include topics related to business process reengineering, the impact of introducing new technologies, or facility design.

Our work is the natural continuation of the literature review carried out by Cardoen et al. [29]. In this chapter we complement the work by adding the recent body of literature (2007–2010) and including a more in-depth analysis of the trends being followed by the research community. Trends are investigated on the interval starting from 2000 and ending at 2010, while detailed descriptions are only provided for the most recent contributions being published after 2006 and which were not included in [29].

In the past 60 years, a large body of literature on the management of ORs has evolved. Magerlein and Martin [95] distinguish between advance scheduling and allocation scheduling as they provide us with a review on surgical demand scheduling. Advance scheduling is the process of fixing a surgery date for a patient, whereas allocation scheduling determines the OR and the starting time of the procedure on the specific day of surgery. Blake and Carter [17] elaborate on this taxonomy in their literature review and add the domain of external resource scheduling, which they define as the process of identifying and reserving all resources external to the surgical suite necessary to ensure appropriate care for a patient before and after surgery. Przasnyski [125] structures the literature on OR scheduling based on general areas of concern, such as cost containment or scheduling of specific resources.

The more recent review by Gupta and Denton [66] focuses on appointment scheduling from a general perspective and describes three commonly encountered healthcare scheduling environments, namely primary care, specialty care, and elective surgery (non-emergencies). In primary care usually a primary care physician, such as a general practitioner or a family physician, acts as the principal point of consultation. Specialty care is focused on a specific and often complex diagnosis and treatment, whereas elective surgery is focused on a specific procedure. They discuss various factors, which affect the performance of the appointment system, including arrival and service time variability, patient and provider preferences, and performance measures such as patient waiting time, OR idle time, and overtime. According to these factors, the existing literature is classified into three groups. The difference between Gupta and Denton's review and our review is that our focus is limited to scheduling problems, which arise in close relationship to the OR. Consequently, we include elective surgery scheduling and exclude considerations related to primary and specialty care.

In the literature review of Guerriero and Guido [63], a selected number of articles are categorized according to the commonly used three hierarchical decision levels: strategic, tactical, and operational. Strategic decisions involve defining both number and types of surgeries to be performed, and hence affect the OR function in the long term. The tactical level usually involves the construction of a cyclic schedule, which assigns time blocks to surgeons or surgeon groups. The final, operational level does not influence the number and type of performed procedures, but deals mostly with daily staffing and surgery scheduling decisions. The three hierarchical levels give the OR planning problem structure. Nonetheless, in our literature review we chose not to use the three hierarchical levels but rather to define descriptive fields. The justification for our decision can be found in Sect. 2. Other reviews, in which OR management is covered as a part of global healthcare services, can be found in [21, 124, 135, 165].

We searched the databases *Pubmed* and *Web of Knowledge*[1] for relevant manuscripts, which are written in English and appeared in 2000 or afterwards. Search phrases included combinations of the following words: operating, surgery, case, room, theatre(er), scheduling, planning, and sequencing. We searched in both titles and abstracts and in addition checked the complete reference list of any found article. As the search process happened in an unbiased way, we believe to have arrived at a set of articles, which objectively represents most of the articles in the field. At the end of the search procedure, a set of 181 articles was identified of which 136 [2, 3, 5–16, 18–20, 22–25, 27, 28, 33–41, 43, 45, 46, 48–62, 64, 65, 67–84, 86–94, 96–100, 105–111, 113–115, 117–123, 126–132, 136–148, 150–157, 161–164, 166–168] were found to be technically oriented. We define an article as "technical" if it contains an algorithmic description of a method directly related to patient scheduling. Some articles missing this algorithmic component instead provide managerial insights. Those types of articles are classified as "managerial" and excluded from the classification process itself. However, these are mentioned in the text where they seem appropriate. The quantitative descriptions provided in later sections, which try to give insights into the dynamics of the trends followed by the research community, are exclusively based on the technical contributions. The distribution of included articles according to their year of publication is shown in Fig. 5.1.

The structure we use is meant to balance between simplicity and expressiveness. We provide a simplified, but in our belief for the majority of the readers, sufficiently accurate way to identify and select articles they are interested in. In the detailed tables, all researched manuscripts are listed and categorized. Pooling these tables

[1] Includes: Web of Science, Current Contents Connect, and Inspec.

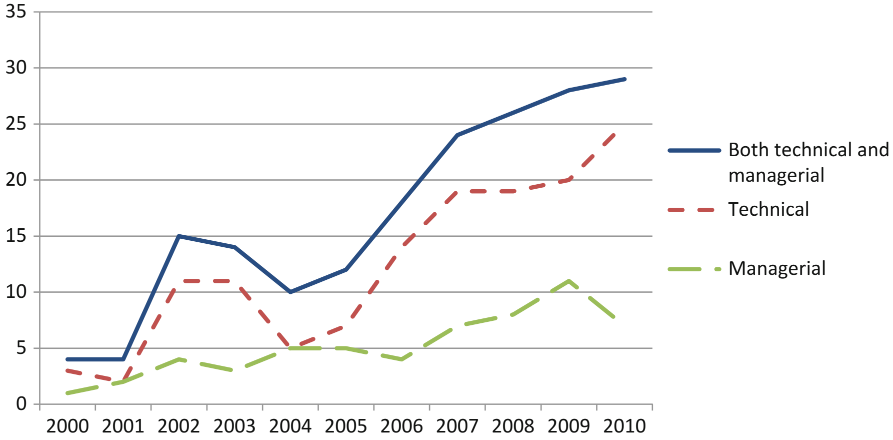


**Fig. 5.1** The distribution of included articles according to their year of publication. Figures appearing in the text are based on technical contributions only

over the several fields should enable the reader to reconstruct the content of specific papers. They furthermore act as a reference tool to obtain the subset of papers that correspond to a certain characteristic. To search for literature in a more direct way the database containing the detailed classification of each analyzed article is made available at www.econ.kuleuven.be/healthcare/review2011 in the form of an Excel [Microsoft, Redmond, WA] spreadsheet.

## 2 Organization of the Review

As in [63], many authors differentiate between strategic (long term), tactical (medium term), and operational (short term) approaches, and situate their planning or scheduling problem accordingly. With respect to the operational level, a further distinction can be made between offline (i.e., before schedule execution) and online (i.e., during schedule execution) approaches. The boundaries between these major categories can vary considerably for different settings and hence are often perceived as vague and interrelated [134]. Furthermore, this categorization seems to lack an adequate level of detail. Other taxonomies, for instance, are structured and categorized on a specific characteristic of the paper, such as the use of the solution or evaluation technique. However, when a researcher is interested in finding papers on OR utilization, a taxonomy based on solution technique does not seem very helpful. Therefore, we propose a literature review that is structured using descriptive fields. As in Cardoen et al. [29] each field analyzes the manuscripts from a different perspective, which can be either problem or technically oriented. In particular, we distinguish between six fields:

- Patient characteristics (Sect. 3): reviewing the literature according to the elective (inpatient or outpatient) or nonelective (urgency or emergency) status of the patient.
- Performance measures (Sect. 4): discussion of the performance criteria such as utilization, waiting time, preferences, throughput, financial value, makespan, and patient deferral.
- Decision delineation (Sect. 5): indicating what type of decision has to be made (date, time, room, or capacity) and whether this decision applies to a medical discipline, a surgeon, or a patient (type).
- Uncertainty (Sect. 6): indicating to what extent researchers incorporate arrival or duration uncertainty (stochastic versus deterministic approaches).
- Research methodology (Sect. 7): providing information on the type of analysis that is performed and the solution or evaluation technique that is applied.
- Applicability of research (Sect. 8): information on the testing (data) of research and its implementation in practice.

Each section clarifies the terminology if needed and includes a brief discussion based on a selection of appropriate manuscripts. Furthermore, a detailed table is included in which all relevant manuscripts are listed and categorized. Plots are provided to point out some of the trends followed by the research community. It should be noted that, if not stated otherwise, the percentages are calculated in relation to the total amount of technical papers. Also note that some categories are not interpretable for some methods and even though rare, some articles contain more than one single method. As a consequence, the sum of mutually exclusive categories does not necessarily add up to 100%.

## 3 Patient Characteristics

Two major patient classes are considered in the literature, namely elective patients and nonelective patients. The former class represents patients for whom the surgery can be planned in advance, whereas the latter class groups patients for whom a surgery is unexpected and hence needs to be planned on short notice. A nonelective surgery is considered an emergency if it has to be performed immediately and an urgency if it can be postponed for a short time (i.e., days).

As shown in Fig. 5.2 and Table 5.1, the literature on elective patient scheduling is vast compared to the nonelective counterpart. Although many researchers do not indicate what type of elective patients they are considering, some distinguish between *inpatients* and *outpatients*. Inpatients refer to hospitalized patients who have to stay overnight, whereas outpatients typically enter and leave the hospital on the same day. As pointed out by Medpac [103], in reality an ongoing shift of services from the inpatient to the outpatient setting is present, which is reflected in a higher growth rate of the latter. Moreover, according to the MMI, outpatient costs increase with a yearly value of 10%, of which 90% are attributable to increasing prices

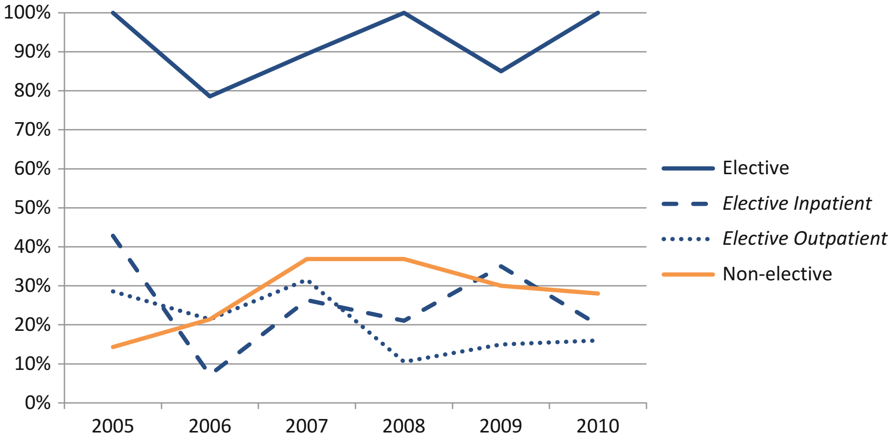


**Fig. 5.2** The majority of contributions relate to the elective patient. Surprising is that, contrary to what might be expected, the share of outpatient-related articles is not increasing. (As some articles deal with both elective and nonelective patients, the sum of the two will add up to a value higher than 100%.)

of existing and more expensive emerging services [104]. Despite the increasing importance of outpatient care in general, the share of outpatient-related academic articles as shown in Fig. 5.2 remains constant in time.

**Table 5.1** It is not always specified what type of patients is considered and, especially for the elective patient case, not always clear whether an inpatient or outpatient setting is used

| | |
|---|---|
| Elective | |
| Inpatient | [2, 8, 13, 14, 16, 23, 25, 36, 48, 62, 67, 72, 74, 84, 94, 108, 109, 118, 123, 136, 137, 139, 140, 143, 144, 148, 154, 156, 167] |
| Outpatient | [8, 15, 23, 25, 27, 28, 40, 43, 46, 48, 52, 61, 67, 74–76, 84, 108, 109, 123, 126, 129, 139, 140, 148, 154, 157, 167] |
| Not specified | [3, 5–7, 9, 11, 22, 24, 34, 35, 37–39, 45, 49, 50, 54–60, 65, 68–71, 73, 77, 80–83, 86–92, 96, 99, 100, 105–107, 111, 113–115, 117, 119–122, 127, 128, 130, 132, 138, 141, 142, 145, 146, 150–153, 155, 161–164, 166, 168] |
| Nonelective | |
| Urgent | [22, 54, 65, 67, 96, 109, 110, 123, 168] |
| Emergent | [9, 25, 60, 67, 71, 72, 82, 87–90, 92, 100, 106, 108, 113, 122, 123, 139, 141, 142, 150, 154, 163, 167] |
| Not specified | [83, 84, 114, 151, 162] |

Scientific contributions on outpatients often (22 out of 37 articles) investigate ORs in an integrated way, i.e., a preoperative or postoperative unit is taken into account. For example, Huschka et al. [76] use both an intake and a recovery area as part of a simulation model of an outpatient procedure center. Several daily scheduling and sequencing heuristics are applied and tested on their impact on patient waiting time and the amount of OR overtime. The authors found that defining the order of surgeries has less influence on patient waiting time and OR overtime than the arrival time schedules. Additionally, the importance of a proper daily surgery mix is stressed.

Other methods focus on assigning patients to days and do not define the exact procedure starting times. Lamiri et al. [88] consider several stochastic optimization methods to plan elective surgeries in case OR capacity is shared by both elective and emergency patients and present a solution method combining Monte Carlo sampling and mixed integer programming. Several heuristics were also tested, from which the most efficient one proved to be tabu search. In their problem setting, it is surprising that the quality of heuristic solutions degrades as the variability on the amount of emergency arrivals decreases, i.e., the stochastic problem is easier than the deterministic one. Planning for stochastic emergency arrivals poses a problem, because it leaves an uncertain amount of capacity left for elective patients. Ferrand et al. [60] investigate whether eliminating this source of uncertainty by channeling emergency arrivals to dedicated emergency ORs is beneficial. This requires however that a constant number of ORs be reserved for emergencies and therefore leaves less free capacity for elective patients. Nevertheless, based on their results using a simulation model, they find elective patients benefit from this, whereas emergency arrivals have an increased waiting time.

A scenario where a hospital dedicates all of its ORs to emergency services is the case of a disaster. As a consequence, all elective surgeries are canceled while resources are redirected to provide quick care to non-electives. This type of nonelective patient is an urgency, as quick but not necessarily immediate care is required. Nouaouri et al. [110] sequence a large number of patients resulting from a disaster, with the objective of maximizing patient throughput. Their approach identifies patients which cannot be served by the given hospital and therefore have to be transported to another one.

Zonderland et al. [168] refers to patients who have to be treated within 1–2 weeks as *semi-urgencies* and uses queuing theory to analyze the trade-off between reserving too much OR time for their arrival (which results in unused OR time) and excessive cancelations of elective surgeries. Additional complexity results from the fact that canceled elective patients turn into semi-urgent patients, which consequently need to be served within the following 2 weeks. An insight gained by the authors is that focusing only on the average behavior of the system can result in undesired system outcomes, i.e., shortage of capacity, which ultimately leads to the cancelation of many elective patients.

## 4 Performance Measures

Different performance measures emphasize different priorities and will favor the interests of some stakeholders over others. A hospital administrator could, for example, be interested in achieving high utilization levels and low costs. Medical

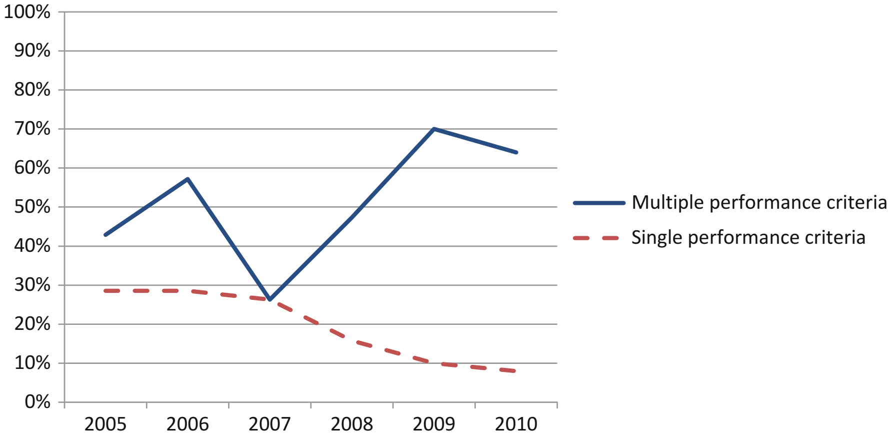


**Fig. 5.3** It is increasingly popular to use multiple instead of single performance criteria. The criteria are not interpretable for all articles, their sum is thus lower than 100%

staff, on the other hand, might care less about cost factors and rather aim to achieve low overtime. The patient as the client of the hospital might care little about the above factors and only desires high quality service and short waiting times. Many authors in the scientific community try to find a compromise between the interests of different stakeholders and simultaneously enforce several kinds of performance criteria. As a matter of fact, as shown in Fig. 5.3, a gradually decreasing number of authors restrict themselves to only one single performance measure.

The major performance measures we distinguish can be found in Table 5.2: utilization, waiting time, leveling, preference, throughput, financial measures, makespan, and patient deferral. As shown in Fig. 5.4, patient waiting time is a frequently used performance measure, which is understandable as one of the major problems in general healthcare consists of long waiting lists but also extensive waits on the day of surgery. Wachtel and Dexter [159, 160] investigated increases in waiting time on the day of surgery, for both surgeon and patient, caused by tardiness from scheduled start times. They concluded that the total duration of preceding cases is an important predictor of tardiness, i.e., the tardiness per case grew larger as the day progressed. If case durations are systematically underestimated, tardiness results. The estimated amount of underutilized (overutilized) time, which was to be expected at the end of the workday, caused average tardiness to increase (decrease). A reduction of tardiness can be achieved by modifying the OR schedule to incorporate corrections for both lateness of first cases of the day and case duration bias.

Figure 5.4 shows that patient throughput is rarely used as a performance measure, and patient or surgeon preferences are increasingly gaining popularity. Preference most often covers some qualitative aspect, whereas throughput is associated with volume. Noteworthy is the fact that both in general healthcare [149] and in the

**Table 5.2** The main performance criteria are: utilization, waiting time, leveling, preference (e.g., priority scoring), financial (e.g., minimization of surgery costs), makespan (completion time), patient deferral, and other (e.g., number of required porter teams)

| | |
|---|---|
| Utilization | |
| Underutilization/undertime | |
| OR | [2, 19, 20, 36, 50, 55–59, 68, 73, 78–80, 90, 91, 100, 111, 113, 115, 140, 142, 148, 152, 156, 161, 164, 167, 168] |
| Ward | [156] |
| ICU | [2, 36, 156] |
| PACU | [2, 36] |
| Overutilization/overtime | |
| OR | [2, 11, 22, 25, 33, 36–41, 50, 51, 54–60, 64, 65, 69, 73, 76–80, 87–91, 96, 98, 100, 105, 106, 111, 113, 115, 119, 121–123, 127, 128, 132, 137, 140–142, 145, 148, 150, 151, 156, 161, 163, 164] |
| Ward | [54, 156] |
| ICU | [2, 36, 115, 156] |
| PACU | [2, 27, 28, 36] |
| General | [3, 8, 9, 12, 22, 23, 25, 33, 34, 48, 50, 54, 60, 61, 69, 71, 72, 91, 111, 122, 132, 139, 142, 145, 148, 150, 151, 163] |
| Waiting time | |
| Patient | [3, 9, 25, 33, 36–38, 40, 54, 60, 61, 64, 65, 67, 76, 78, 79, 81, 91, 93, 106, 107, 109, 111, 118, 120–122, 132, 136, 139, 141–144, 148, 150, 154, 157, 163, 167] |
| Surgeon | [11, 37, 38, 91, 157] |
| Leveling | |
| OR | [15, 98, 99, 111] |
| Ward | [13, 14, 16, 68, 94, 130, 140, 153] |
| PACU | [15, 27, 28, 75, 96, 97, 131, 139, 152] |
| Patient volume | [111, 140, 142] |
| Preference | [16, 18, 27, 28, 34, 35, 52, 62, 83, 88, 94, 105–107, 114, 115, 118, 138, 140, 143–146, 155, 162, 166] |
| Throughput | [3, 8, 9, 12, 71, 100, 110, 126, 130, 132, 142, 145, 154] |
| Financial | [11, 18, 24, 39, 43, 45, 46, 48, 49, 65, 84, 93, 100, 108, 138] |
| Makespan | [5–7, 35, 57–59, 73–75, 86, 96, 123, 129, 137, 147] |
| Patient deferral | [22, 25, 36, 54, 71, 72, 81, 82, 92, 120–122, 132, 139, 145, 168] |
| Other | [2, 9, 10, 12, 33, 35, 36, 61, 68, 77, 83, 87, 89, 90, 93, 96, 99, 100, 117, 121, 123, 127, 128, 136, 141, 142, 153, 155] |

operations research literature value or quality-based approaches seem to be getting increasingly important. For example, the preferences of cataract surgery patients of one surgeon are investigated by Dexter et al. [42]. The surgeon's patients place a high value on receiving care on the day chosen by them, at a single site, during a single visit, and in the morning. Preferences can also be embodied in patient priorities, i.e., preference to perform surgery first on patients who have a more acute condition. Priorities are most often defined at the level of a patient group. Testi et al. [144, 146] define a model where the position of a patient on a waiting list is defined

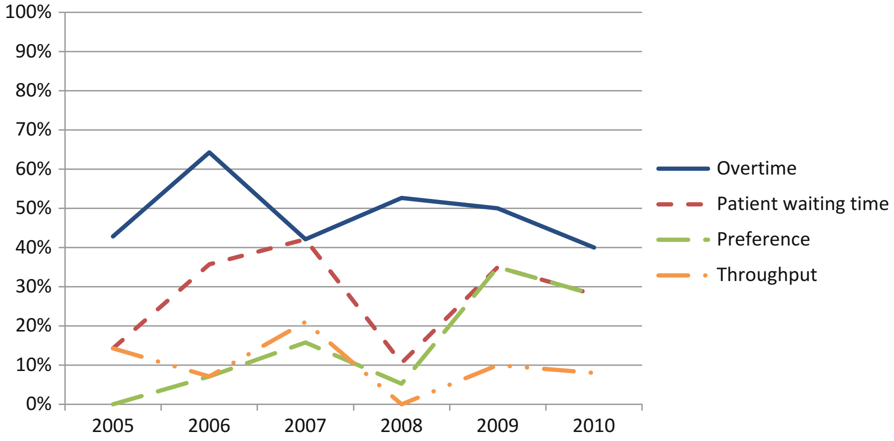


**Fig. 5.4** Some selected performance criteria. Overtime, despite used slightly less in 2010, it is still the most frequently used performance measure. From 2008 onward, preference-related measures seem to become increasingly popular

by a priority scoring algorithm, which considers both patient urgency (progression of disease, pain or dysfunction and disability) and time spent on the surgical waiting list. It is easy to see that priority scoring minimizes the total weighted waiting time of all patients, i.e., using an algorithm where patient priorities are uniform; we will minimize the average patient waiting time.

Including patient priorities drives OR scheduling in a more patient-oriented direction. Min and Yih [105] go a step further and explicitly incorporate an additional payment factor, defined as the cost of overtime. In their model, if many high priority patients are on a waiting list, ORs will be kept open longer in order to avoid high postponement costs, i.e., the surgery postponement costs are balanced versus OR overtime costs. The authors establish that patient prioritization is only useful if the difference between the cost coefficients associated to different priority classes is high, as otherwise a similar schedule can be obtained by using average postponement costs. Additionally, the relative cost ratio between the cost of patient postponement and OR overtime should not be low, as a low ratio would imply high overtime costs and therefore prioritization would only marginally affect the surgery schedule.

Purely financial objectives are rarely used in the literature. In Stanciu and Vargas [138], *protection levels*, the amount of OR time reserved in a partitioned fashion for each individual patient class, are used to determine which patients to accept and which to postpone during the planning period under study. A patient class is a combination of the patient reimbursement level and the type of surgery. A patient class enjoys higher priority if its expected revenue per unit surgery time is higher. The goal is to maximize expected revenues incurred by the surgical unit. Patients, given their class, are accepted or postponed to a later date when the protection level for their class can accommodate them. The central question becomes how many requests to accept from low revenue patients, and how much capacity to reserve for future high revenue patients. Financial considerations are also expressed by Wachtel and Dexter [158], who argue that if additional OR capacity is expanded, it should be assigned to those subspecialties that have the greatest contribution margin per OR hour (revenue minus variable cost), that have the potential for growth, and that have minimal need for a limited resource such as ICU beds.

Figure 5.4 also reveals the fact that minimizing overtime is a popular performance measure. This is understandable as overtime results both in the dissatisfaction of the surgical staff and high costs for the hospital (as higher wages typically apply for the time beyond the normal working hours). Reducing overtime is consequently highly beneficial and often desirable to practitioners. Dexter and Macario [47] establish that a correction of systematically underestimated lengths of case durations would not markedly reduce overutilization of ORs. Tancrez et al. [141] proposed an analytical approach that takes into account both stochastic operating times and random arrivals of emergency patients. They showed how the probability of overtime in the OR changes as a function of the total number of scheduled operations per day. As shown in Table 5.2, we relate underutilization to undertime and overutilization to overtime, although they do not necessarily represent the same concept. Utilization actually refers to the workload of a resource, whereas undertime or overtime includes some timing aspect. Hence, it is possible to have an underutilized set of ORs, although with overtime in some of the ORs. In some manuscripts it is unclear which view is applied. Therefore, we group underutilization with undertime and similarly overutilization with overtime.

Adan et al. [2] formulated an optimization problem that minimizes the deviation from a targeted utilization level of the OR, the ICU, the medium care unit, and nursing hours. The deviation is measured as the sum of overutilization and underutilization. They recommend using stochastic time durations as well as a broad perspective on supporting resources such as the ICU and the wards. Pandit and Dexter [116] defined rules to determine whether an OR should be staffed for 8 or 10 h. They concluded that in case the average OR time is less than 8 h 25 min, 8 h staffing should be planned, while in case the average OR time is over 8 h 50 min, 10 h of staffing is needed. For averages in between, the full analysis of McIntosh et al. [102] should be performed.

Augusto et al. [7] minimized the sum of the makespan (completion time) of patients undergoing surgery. Makespan in general defines the time span between the entrance of the first patient and the finishing time of the last patient. Since minimizing the makespan often results in a dense schedule, deviations from the plan can result in complications requiring adjustments to the schedule. An example is the arrival of an emergency patient to the hospital which results in a dense schedule with no free ORs available. In a case like this, a likely scenario includes the deferral of an elective patient, who will have to be reassigned to another surgery slot at a later point in time. Zonderland et al. [168] used queuing theory to investigate the trade-off between cancelations of elective surgeries due to semi-urgent surgeries and unused OR time. They provide a decision support tool, which assists the scheduling process of elective and semi-urgent cases. General reasons for patient deferrals in one specific hospital are discussed by Argo et al. [4].

In addition to emergency patient arrivals, other reasons can cause cancelation of elective surgeries. For example, a fully occupied PACU would prohibit patients who have already completed surgery from leaving the OR. A blocked OR will delay the service of preceding elective surgeries, which as a final measure may have to be canceled. This situation can be avoided if the OR schedule is constructed in a way that resource utilization is leveled. Similar approaches may be used for other resources. For instance, Ma and Demeulemeester [94] maximize the availability of the number of expected spare beds and investigate bed occupancy levels at wards, whereas Van Houdenhoven et al. [152] target the ICU.

Some of the articles in the literature used methods that have not been covered in the previous paragraphs. Does et al. [53] used Six Sigma to decrease the tardiness of surgeries, which are performed first on a day. Applied to two hospitals in the Netherlands, substantial savings were achieved and the number of surgeries was increased by 10% without requiring additional resources. Epstein and Dexter [44] introduced a method through which analysts can screen for the economic impact of improving first-case starts. In [2, 77] nursing hours are considered, while [128] consider the number of open ORs, and [141] the number of disruptions.

## 5 Decision Delineation

A variety of planning and scheduling decisions are studied in the literature. In Table 5.3, we provide a matrix that indicates what type of decisions are examined, such as the assignment of a date (e.g., on Friday, February 25), a time indication (e.g., at 2 p.m.), an OR (e.g., OR 1, OR of type B), or the allocation of capacity (e.g., 1 h of OR time). The manuscripts are further categorized according to the decision level they address, i.e., to whom the particular decisions apply. We distinguish between the discipline (e.g., pediatrics), the surgeon, and the patient level. Figure 5.5 shows that a large part of the literature aims at the patient level, whereas the discipline as well as the surgeon level receives less attention.

The discipline level unites contributions in which decisions are taken for a medical specialty or department as a whole. This frequently involves the construction of a cyclic timetable, which aims at minimizing the underallocation of a specialties' OR time with respect to its predetermined target time. On a lower level, the target is a surgeon group or a single surgeon. In Denton et al. [39], surgeries consecutively carried out by one surgeon define a surgery block, which are subsequently assigned to ORs. The problem is formulated as a stochastic optimization model, which balances the cost of opening an OR with the cost of overtime.

On the patient level, the decision variables are formulated on the individual patient or patient type. Fei et al. [58] describe a two stage approach, where in the first stage patients are assigned to days and rooms, and the exact daily sequence of surgeries (timing aspect) is set in the second stage. This is a common way of patient

**Table 5.3** Type and level of decisions

| | Discipline level | Surgeon level | Patient level | Other |
|---|---|---|---|---|
| Date | [13, 19, 20, 24, 33, 49, 65, 130, 132, 144, 145, 167] | [14–16, 25, 74, 80, 120, 142] | [2, 25, 33, 34, 36, 48, 49, 54–59, 64, 65, 68, 69, 72, 73, 78–83, 87–90, 94, 105–107, 111, 115, 118–121, 123, 127, 128, 132, 136, 140, 143–146, 152, 153, 155, 161, 164, 166] | [54, 143, 156] |
| Time | [13, 33, 49, 65, 71, 132, 145] | [11, 14–16, 25, 37] | [5–7, 11, 25, 27, 28, 33, 37, 38, 40, 41, 49, 57–59, 65, 70, 71, 73, 75–79, 83, 86, 91, 96, 97, 99, 110, 123, 127, 128, 132, 137, 145, 147, 150] | [11, 12] |
| Room | [19, 20, 33, 62, 130, 132, 144, 145, 167] | [11, 15, 16, 39, 74, 80, 120, 142] | [11, 23, 27, 28, 33–35, 41, 50, 51, 55–60, 64, 68–70, 73, 76–80, 83, 86, 87, 90, 96, 98, 99, 106, 107, 110, 111, 115, 118, 120, 121, 123, 127–129, 132, 137, 143–145, 152, 153, 161, 163, 164] | [11, 143] |
| Capacity | [22, 24, 33, 49, 65, 67, 71, 130, 132, 139, 145, 167] | [11, 18, 25, 37, 43, 45, 46, 80, 84, 120] | [2, 11, 22, 25, 33, 36, 37, 49, 54, 64, 65, 68, 71, 72, 80, 94, 105, 108, 114, 120, 121, 132, 138, 141, 145, 168] | [11, 12, 52, 54, 61, 92, 93, 109, 117, 122, 126, 131, 154] |
| Other | [151] | [120] | [3, 7, 35, 54, 111, 120, 136, 143, 157, 162, 168] | [52, 54, 143, 148] |

For example, articles dealing with the sequencing problem are found in column 3 and row 2. Articles dealing with the problem generally referred to as the assignment step are found in column 3 and row 1. Defining patient capacity requirement for a given day of the week are articles found in column 3 and both row 1 and row 4

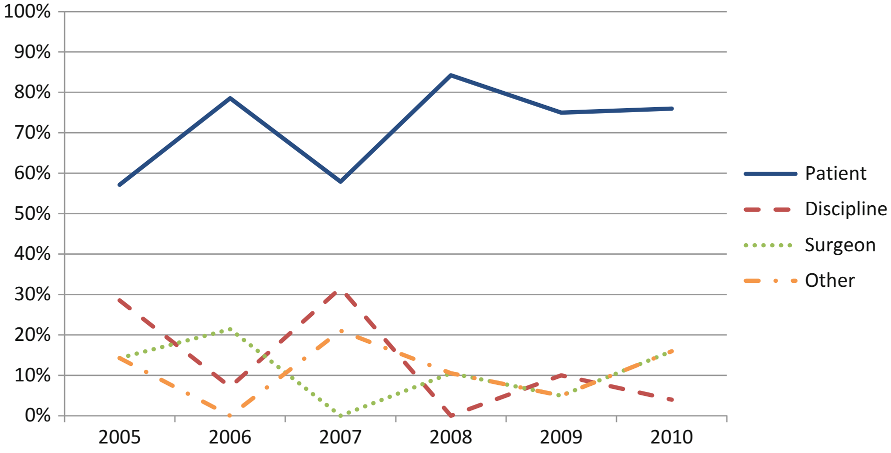


**Fig. 5.5** In the literature, most decision problems relate to the patient level. A typical problem setting, for example, consists of finding the optimal patient to day/OR assignment

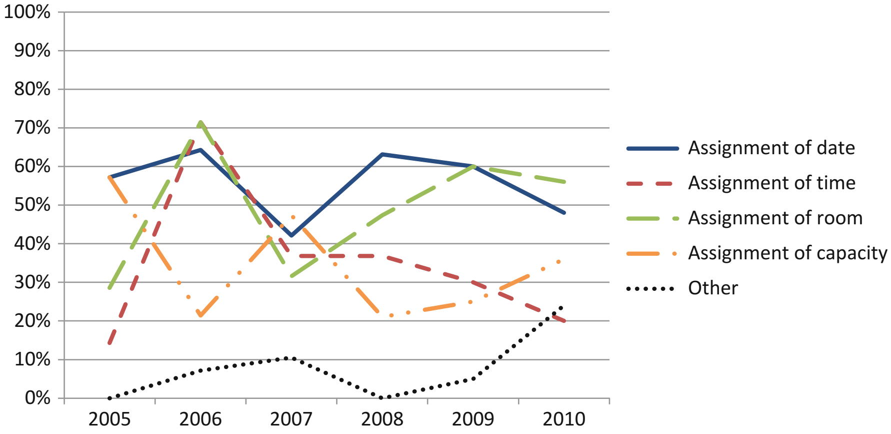


**Fig. 5.6** Solution approaches related to the assignment of dates and rooms are increasingly popular in the literature, whereas the time assignment step (e.g., sequencing) is slightly less popular than it used to be

scheduling, as the assignment of the day and the room of a given surgery is more easily planned ahead than the exact time, which is often fixed close to the actual surgery date.

Date, room, time, and capacity questions as shown in Table 5.3 are answered on all three decision levels. Figure 5.6 shows that the share of manuscripts dealing with decisions on exact times is decreasing, whereas date, room, or capacity problems are continuously addressed in the literature. A capacity problem is discussed by Masursky et al. [101] who forecasted long-term anesthesia and OR workload.

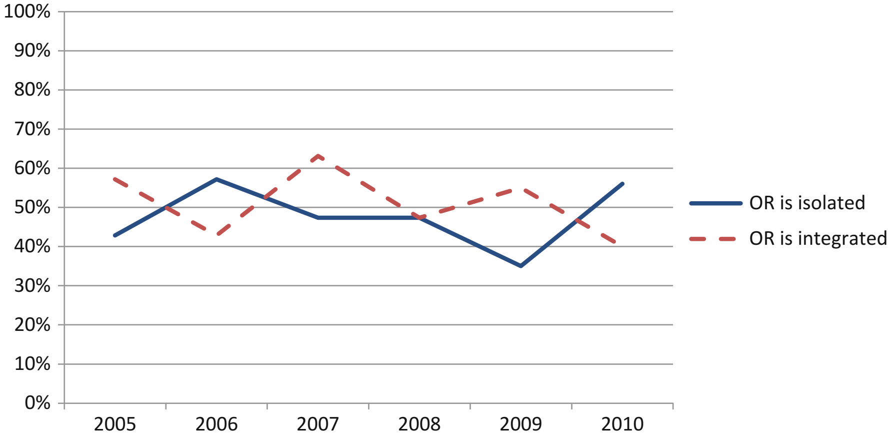


**Fig. 5.7** Every second article deals with the OR planning and scheduling process in an integrated way

They concluded that forecasting future workload should be based on historical and current workload-related data and advise against using local population statistics. The problem of forecasting workload is addressed by Gupta et al. [67] as well. In their case study, simulation is used to answer capacity-related questions. They concluded that a one-time infusion of capacity in the hope to clear backlogs will fail to reduce waiting times permanently, while targeting extra capacity to highest urgency categories reduces all-over waiting times including those of low urgency patients. In situations where arrival rates increased, even if only within a specific urgency class, waiting times increased dramatically and failed to return to the baseline for a long time.

We added both a row and a column (other) to Table 5.3 to provide entries for manuscripts that study OR planning and scheduling problems in a way that is not well captured by the main matrix. Manuscripts that are categorized in this column or row examine, for instance, capacity considerations with regard to beds [92, 131], OR to ward [143], and patient to week assignments [168] or different timing aspects, such as the amount of recovery time spent within the OR [7].

As OR planning and scheduling decisions affect facilities throughout the entire hospital, it seems to be useful to incorporate facilities, such as the ICU or PACU, in the decision process in an attempt to improve their combined performance. If not, we believe that improving the OR schedule may worsen the efficiency of those related facilities.

Figure 5.7 shows that the ratio of manuscripts between 2005 and 2010, which deal with the OR in an integrated way, and those which deal with the OR in an isolated way, are oscillating around the 50% mark. This is surprising as we would expect to observe an increasing interest in integrated approaches. Whether a manuscript uses an integrated or an isolated approach can be looked up in Table 5.4.

**Table 5.4** In an integrated OR, supporting facilities such as the ICU, PACU, and wards are considered

| | |
|---|---|
| Isolated OR | [5, 10, 11, 19, 20, 22, 33, 34, 37–39, 41, 45, 48–52, 55–60, 64, 65, 67–69, 73, 74, 83, 84, 87–91, 98, 99, 105, 107, 110, 111, 113, 114, 117–120, 122, 127, 128, 132, 138, 141, 146–148, 150, 151, 157, 161–164, 166, 168] |
| Integrated OR | [2, 3, 6–9, 12–16, 18, 23–25, 27, 28, 36, 40, 43, 46, 54, 58, 61, 62, 65, 68, 70–73, 75–79, 82, 86, 94, 96, 97, 100, 106, 108, 109, 115, 121, 123, 126, 130, 131, 136, 137, 139, 140, 142, 143, 145, 152–156, 167] |

The problem of the congested PACU, which previously had been scarcely addressed in the literature, received more attention between 2008 and 2010. In this problem, patients are not allowed to enter the fully occupied PACU and are therefore forced to start their recovery phase in the OR itself, keeping it blocked. Iser et al. [77] used a simulation model to tackle the problem and compare overtime occurring in the OR with PACU-specific performance measures. Augusto et al. [7] showed the benefit of using a mathematical model to plan ahead the exact amount of recovery time a patient will spend within the OR. As it is typical for highly utilized systems, there is a sensitive relationship between overall case volume, capacity (of the PACU), and the effect on waiting time (to enter the PACU). This relationship is described in more detail by Schonmeyr et al. [131] using queuing theory.

Besides the PACU, a downstream facility which could affect the function of the OR is the ICU. Kolker [82] reduced diversion of an ICU to an acceptable level by defining the maximum number of elective surgeries per day that are allowed to be scheduled along with the competing demand from emergency arrivals. Litvak et al. [92] went a step further and tackled the ICU capacity problem in a cooperative framework. In their model, several hospitals of a region jointly reserve a small number of beds in order to accommodate emergency patients and achieve an improved service level for all patients.

## 6 Uncertainty

One of the major problems associated with the development of accurate OR planning and scheduling strategies is the uncertainty inherent to surgical services. Deterministic planning and scheduling approaches ignore uncertainty, whereas stochastic approaches explicitly try to incorporate it. In Table 5.5, we list the relevant manuscripts classified according to the type of uncertainty they incorporate.

As shown in Fig. 5.8, stochasticity in the form of uncertain patient arrivals and surgery durations is frequently used in the OR literature. If we narrow the literature to recent contributions, which explicitly incorporate nonelective patients; we see that over 80% of the methods incorporate some sort of uncertainty. Nonelective patient arrivals are in most cases impossible to predict in advance and additionally occupy a random amount of OR time, which often leaves OR managers with no

**Table 5.5** Methods frequently take stochasticity into account

| | |
|---|---|
| Deterministic | [6, 7, 10, 14, 15, 18–20, 24, 27, 28, 33–35, 43, 46, 52, 55–59, 64, 68, 70, 73, 75, 77–80, 83, 84, 86, 97, 99, 107, 108, 110, 111, 114, 115, 118, 120, 121, 123, 127, 128, 130, 136, 137, 140, 143–146, 150, 152, 155, 156, 164, 166, 167] |
| Stochastic | |
| Arrival | [3, 9, 13, 16, 22, 23, 25, 36, 48, 49, 54, 62, 65, 67, 71, 72, 74, 81, 82, 87, 89, 90, 93, 94, 100, 105, 109, 117, 121, 122, 126, 132, 138, 139, 141, 142, 145, 150, 154, 162, 163, 167, 168] |
| Duration | [2, 3, 5, 8, 9, 11–13, 16, 22, 25, 36–41, 49, 50, 54, 62, 65, 67–69, 71, 72, 76, 81, 87–89, 91, 93, 94, 96–98, 100, 105, 106, 109, 113, 117, 119, 122, 126, 131, 132, 139, 141, 142, 145, 148, 150, 151, 153, 154, 157, 161–163, 167] |
| Other | [3, 5, 9, 25, 45, 46, 72, 92, 94, 106, 126, 162] |

Most common forms are duration and arrival uncertainty

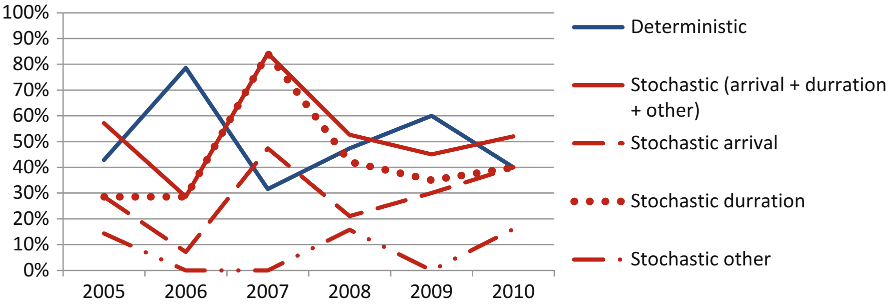


**Fig. 5.8** Uncertainty incorporation. Implicit in the figure is the fact that if duration uncertainty is accounted for, it is often the case that arrival uncertainty is considered as well

option but to keep a safety margin to accommodate them [141]. In contrast, the arrival of elective patients contains a lower amount of uncertainty, and as shown in Fig. 5.9, is frequently considered as deterministic in the literature.

Duration uncertainty is a central element in Denton et al. [39] as well as in Batun et al. [11]. In Denton et al. [39] decisions include the number of ORs to open and surgery block to OR assignments, whereas in Batun et al. [11] this is supplemented by patient sequencing and setting surgeon start times. Both methods aim at minimizing OR opening and OR overtime costs, where Batun et al. [11] additionally consider surgeon idle times. The functional difference between their methods lies in the way surgery to OR assignments are carried out. In Denton et al. [39], the common practice of assigning a surgery block to a single surgeon (block scheduling) is followed, whereas Batun et al. [11] consider the scenario of pooled ORs, and therefore surgeons are allowed to switch between ORs. OR pooling allows carrying out surgeries in parallel as the main surgeon only needs to be present during the critical part of the surgery and can move to the next patient before the close-up of the patient.

A timing aspect, which is different from the actual surgery duration but is characterized by large variations, is the patient length of stay (LOS) in the PACU,

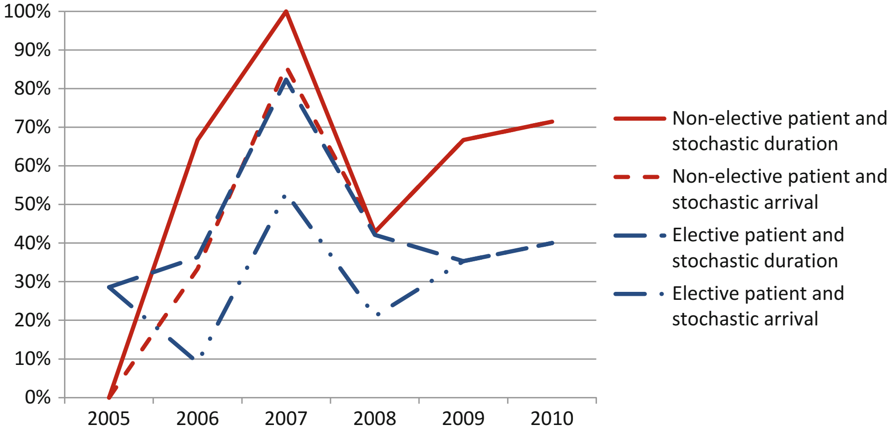


**Fig. 5.9** Stochasticity in the elective and nonelective patient setting. If uncertainty is considered in an elective or nonelective patient setting, stochastic aspects are in either case slightly more often applied to surgery durations than to patient arrival times. (Percentages are calculated in ratio to the total number of articles dealing with the respective patient type.)

ICU, or ward. The variability of patient time spent in the ward is considered by Ma and Demeulemeester [94] in which the rate of patient misplacements caused by bed shortages is minimized.

It should be clear that operations research techniques are able to deal with stochasticity, especially simulation techniques (included in 66% of the stochastic literature) and analytical procedures (included in 22% of the stochastic literature) and that an adequate planning and scheduling approach may lower the negative impact of uncertainty. Mostly, studies assumed a certain level of variability, based on analyzing historical data, and use this information as input for models. However, only limited attention is paid to the reduction of variability within the individual processes. As an example, consider the estimation of surgery durations. Instead of the immediate determination of the distribution of a surgery duration, one should examine whether the population of patients for which the durations are taken into account is truly homogenous. If not, separating the patient population may result in a decreased variability even before the planning and scheduling phase is executed. As the estimation of surgery durations exceeds the scope of this literature review, we do not elaborate further on this issue.

## 7 Operations Research Methodology

The literature on OR planning and scheduling exhibits a wide range of methodologies that fit within the domain of operations research and that combine a certain type of analysis with some solution or evaluation technique. Table 5.6 provides an

**Table 5.6** Different solution techniques are used in the literature: analytical procedures (e.g., queuing theory or new vendor model), mathematical programming, dedicated branch-and-bound, scenario analysis (or sensitivity analysis), simulation, and heuristics

| | |
|---|---|
| Analytical procedure | [22, 37, 59, 62, 65, 88, 89, 92, 93, 105, 113, 114, 131, 141, 151, 157, 162, 168] |
| Mathematical programming | |
| Linear programming | [7, 37, 43, 46, 84, 108, 119] |
| Goal programming | [2, 18, 36, 115, 140] |
| Integer programming | [19, 20, 27, 33, 52, 110, 130, 142–145, 153] |
| Mixed integer programming | [11, 13, 16, 39, 68, 78–80, 87–89, 106, 107, 118, 120, 121, 123, 128, 167] |
| Column generation | [55, 57–59, 68, 73, 86, 87, 90, 152, 153] |
| Branch-and-price | [14, 28, 56] |
| Dynamic programming | [6, 7, 14, 28, 56, 73, 90, 105, 164] |
| Other | [6, 7, 13, 16, 45, 70, 99, 119] |
| Dedicated branch-and-bound | [27, 39, 111, 156] |
| Scenario analysis | [3, 5, 7–9, 12, 15, 18, 22, 23, 25, 34–41, 43, 46, 48–51, 54, 55, 57–59, 61, 62, 67, 69, 71, 72, 74–76, 78, 81, 82, 84, 86, 91–94, 96–98, 100, 106–109, 111, 113, 115, 117, 118, 121, 122, 126, 128, 130–132, 136, 138, 139, 141, 142, 145, 146, 148, 150–152, 154, 156, 157, 162, 163, 167] |
| Simulation | |
| Discrete-event | [3, 5, 8, 9, 12, 22, 23, 25, 36, 48–50, 54, 58, 60, 61, 67, 71, 72, 74, 76, 77, 81, 82, 92–94, 96–98, 100, 106, 109, 121, 122, 126, 132, 139, 142, 145, 148, 150, 154, 163, 167] |
| Monte Carlo | [22, 40, 46, 69, 87–89, 91, 111, 117] |
| Heuristics | |
| Improvement heuristics | |
| Simulated annealing | [13, 16, 40, 69, 88, 150] |
| Tabu search | [35, 73, 75, 88] |
| Genetic algorithm | [34, 58, 72, 73, 127, 128, 136, 137, 147, 161, 166] |
| Other | [19, 20, 38, 69, 87, 88, 90, 98, 106, 138, 150, 164] |
| Constructive heuristics | [5, 6, 13, 16, 33, 37–39, 49–51, 55, 58, 59, 64, 69, 76, 77, 83, 86–88, 90, 126, 142, 143, 150] |

overview of the ways in which OR planning and scheduling problems are analyzed. The table shows that mathematical programming and heuristics are frequently applied, generally to patient sequencing or assignment type of problems (e.g., patient to surgeon/OR assignment). These types of problems are combinatorial optimization problems.

In some approaches the impact of specific changes to the problem setting is examined. We refer to this type of analysis as scenario analysis since multiple scenarios, settings, or options are compared to each other with respect to the performance criteria.

As shown in Fig. 5.10, performing scenario analysis is popular, especially in the discrete-event simulation (DES) modeling literature. Scenario analysis can be done by plotting the efficiency frontier formed by respective scenarios' (possibly

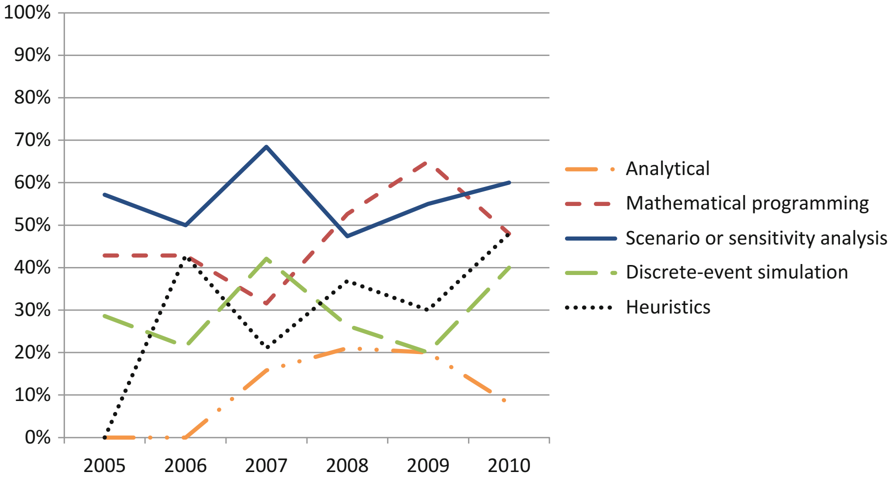


**Fig. 5.10** Only selected solution techniques are shown. Most articles include a scenario or sensitivity analysis

multidimensional) performance scores. This helps to identify and distinguish between advantageous and disadvantageous scenarios. The performance criteria most frequently used in the DES modeling literature are patient waiting time and different kinds of utilization measures such as the overtime related to the OR, ward, ICU, or PACU.

In Sect. 4, we expressed our surprise about the lack of an increasing use of integrated approaches. In the DES modeling literature, however, the proportion of integrated approaches does increase, as OR supporting facilities such as the PACU or ward are increasingly taken into account. We think that modeling the OR in an integrated way is an important step towards the construction of more realistic and applicable models.

An integrated DES model is introduced by Steins et al. [139], in which preoperative care as well as a PACU are considered. The arrival of case types, the surgery time, and the LOS in the PACU are represented as probabilistic distributions. The patients in their model are differentiated according to their urgency status, i.e., whether a patient is elective or nonelective. It is true in general that DES modeling approaches take explicit account of non-electives.

In the literature, the analytical approach is less often encountered than DES models. Aside from their differences on the methodological side, both DES and analytical methods are often related to a similar problem setting. In case of a stochastic environment where capacity questions have to be answered and non-electives possibly play a role, both of the approaches are useful. Tancrez et al. [141] define the amount of OR capacity, which is needed to accommodate for nonelective patients in a Markovian model setting. Simulation is used to show that the assumptions required to build the Markov chain have a minor influence on their final analytical results. In their work, the stochasticity in OR capacity is the consequence of randomly arriving nonelective patients occupying an uncertain amount of OR time. Also without nonelective patient arrivals, it is difficult to predict the required OR capacity on a day, as surgery durations are unknown in advance and can vary considerably in length. In Olivares et al. [113], the decision-making process of reserving OR capacity is investigated using the newsvendor model. In the analytical approach, an estimate is given of the cost placed by the hospital on having idle capacity and the cost of a schedule overrun. Their results reveal that the hospital under study places more emphasis on the tangible costs of having idle capacity than on the costs of a schedule overrun and long working hours for the staff.

As shown in Fig. 5.10, MPs are popular. As opposed to DES and analytical models, MPs, such as mixed integer programs, deal with combinatorial optimization problems. In the majority of cases ($>$60%), the objective function of the optimization problem includes under/overtime or under/overutilization. Those performance criteria are rarely used by themselves but are usually part of a multiple objective formulation. The use of multiple objectives in MPs, as is the case in general, is increasingly popular. In 2010, less than 20% of mathematical formulations found in the literature still restrict themselves to a single objective. Their popularity can be explained in two ways. First, the development of better solvers makes it increasingly practical to use them. Second, defining multiple objectives allows capturing stakeholder preferences more realistically.

Despite the increasing complexity of the objective functions of MPs, there are no indications that the same would be true in respect to their constraints. In other words, the variety of constraints used in MPs seems to be constant. The most frequently used constraints are resource related, which in many cases relate to the OR (under, regular, or overtime) or medical personnel. Regularly applied constraints, which do not focus on a given resource, are priority constraints (a high priority patient always needs to be served before a low priority patient), demand-related constraints (a given specialty needs to be given a certain amount of OR time), and release related (a patient belonging to a given category needs to be served before a given deadline). As in Min and Yih [106], the decisions in most of the mathematical programs apply to the elective patient. In their work, a stochastic mixed integer programming model is proposed and solved by a sampling-based approach. The surgery durations, the LOS, the availability of a downstream facility (ICU), and new demand are assumed to be random with known distributions.

In some cases mathematical programs are too difficult to solve within a reasonable time limit and therefore heuristics are proposed. In Fei et al. [59] a column generation-based heuristic is used to solve the patient scheduling problem. In their setting, a column corresponds to a feasible plan, in other words, the assignment of surgical cases to an OR. Roland et al. [128] propose a method, which includes the assignment of cases to ORs, planning days, and operating time periods. The NP-hard problem is tackled by means of a genetic algorithm. Similarly to mathematical programming, also heuristics are in most of the cases used for scheduling tasks involving the elective patient. Noteworthy is that in 2006, all heuristic methods found in the literature were time assignment problems whereas in 2010, as a result of a gradual decrease, this was true for only 20% of the articles as date and room assignment problems become more popular.

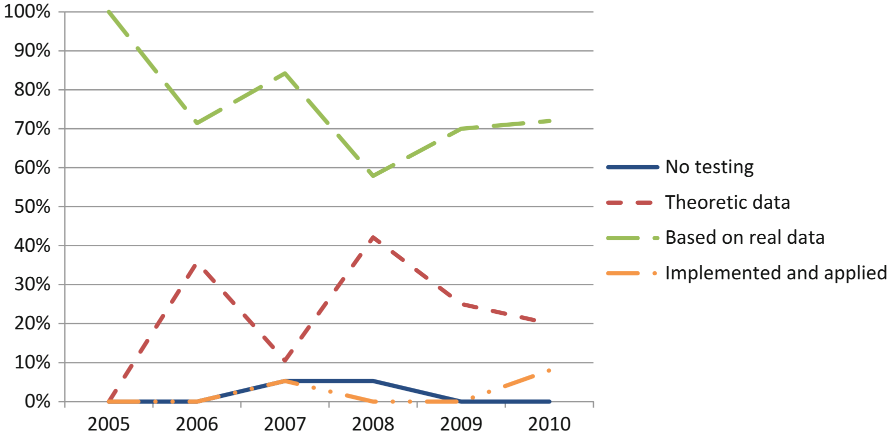


**Fig. 5.11** Even though most data used in the literature are based on real data, this does not mean that the methods are applied in reality

# 8 Applicability of Research in Practice

Many researchers provide a thorough testing phase in which they illustrate the applicability of their research. Whether applicability points at computational efficiency or at showing to what extent objectives may be realized, a substantial amount of data is desired. From Fig. 5.11 and Table 5.7, we notice that most of this data is on real healthcare practices. This evolution is noteworthy and results from the improved hospital information systems from which data can be easily extracted. Unfortunately, a single testing of procedures or tools based on real data does not imply that they finally get implemented in practice. Lagergren [85] indicates that the lack of implementation in the health services seems to have improved considerably. Figure 5.11 shows, however, that only a very small share of the articles report on actual implementation. An exception to this is Wachtel and Dexter [157] who introduce a web site, which is used by the hospital under study to decide on the exact times patients have to arrive to their surgery appointment. The problem tackled by the authors arises from the fact that a case is often started earlier than scheduled, but it cannot be known in advance if it will happen or not. Patient availability must therefore be balanced against patient waiting times and fasting times. Daily applicability is entailed by their method. However, there are problems, which have to be solved on a less frequent basis. An example is the application of a case mix model that is applied every year, clearly resulting in a different degree of implementation. A clear comparison of manuscripts on this aspect is hence not straightforward. Even if the implementation of research can be assumed, authors often provide little detail about the process of implementation. Therefore, we encourage the provision of additional information on the behavioral factors that coincide with the actual

**Table 5.7** Both theoretic and real data are frequently used for testing purposes

| | |
|---|---|
| No testing | [65, 161, 162] |
| Data for testing | |
| Theoretic | [6, 7, 13, 14, 37, 49, 51, 54–56, 64, 73, 77–79, 83, 86–91, 94, 96, 98, 110, 111, 120, 123, 136–138, 143, 164] |
| Based on real data | [2, 3, 5, 8–12, 15, 16, 22–25, 27, 28, 33–36, 38–41, 43, 45, 46, 48, 50–52, 57–62, 67–72, 74–76, 80–82, 84, 92, 93, 97, 99, 100, 105–109, 111, 113–115, 117–123, 126–128, 130–132, 139–141, 144–148, 150–157, 163, 166–168] |
| Implemented and applied | [18–20, 60, 70, 128, 141, 144] |

implementation. Identifying the causes of failure, or the reasons that lead to success, may be of great value to the research community [26].

In many contributions a problem is solved and applied to the problem setting specific to one single hospital, and it is unclear whether or to what extent a method is applicable to another setting. In order to justify the generality of their modeling assumptions, Schoenmeyr et al. [131] surveyed several hospitals. Introducing generalizable methods makes it easier to spread and implement good working operations research practices to more than one hospital.

Only limited research has been done to study which planning and scheduling expertise is currently in use in hospitals. Using a survey, Sieber and Leibundgut [133] reported that the state of OR management in Switzerland is far from excellent. A similar more recent exercise for Flemish (Belgium) hospitals is described in Cardoen et al. [30]. It seems contradictory that so little research is effectively applied in a domain as practical as OR planning and scheduling.

## 9 Opportunities for Future Research

Our review suggests that methods introduced in the literature are rarely implemented at hospitals and, if implemented, details usually remain unpublished. Both the problems of low success rates of implementations and the lack of reports could be mitigated by actively involving surgeons, head nurses, and IT personnel as coauthors. In order to avoid developing scheduling software that is only specified to the needs of one hospital, it might be wise that projects cover several hospitals. Including more than one hospital in a study provides, besides generalizability, also other opportunities: resource pooling on the level of emergency ORs, anesthesia rooms, equipment, or even nursing staff can lead to an integrated approach, which profits all participating hospitals. Integration is also an important concept within the hospital itself. Considering its importance, it appears there is an opportunity to study the role of supporting facilities such as the anesthesia department, PACU, ICU, and/or wards in an integrated way with the OR. Equally important as integrality is the incorporation of uncertainty. First and most prominently, uncertainty is accounted for in respect to patient surgery times, which are unknown until surgeries are actually realized. Second, it is unknown whether the operating room will be available at the planned surgery start as an emergency could occupy it. Third, while booking a surgery into a certain slot, it is unknown whether the slot might be needed to allocate a future more urgent patient. Even though we see that uncertainty is frequently incorporated, the question arises whether it should be a prerequisite for an algorithm to take aspects of uncertainty into account, i.e., is a strictly deterministic scheduling approach able to provide the robustness required in reality? These and other open questions present many opportunities for future research related to OR planning and scheduling.

## 10 Conclusion

In this chapter we have studied and described recent trends in the field of OR planning and scheduling. Based on the data, we found that most attention is given to elective patients, and even though often not stated explicitly, it is in many cases implied to be an inpatient setting. Less frequently occurring in the literature are methods which consider outpatients. This is surprising as outpatient care is gaining in importance and therefore we would expect to observe an increasing amount of literature dealing in this area. We also observed a gradual shift from determining the exact time of a surgery to problems related to date and/or room assignments. With respect to the performance measures considered, we found that overtime is the most popular measure and that preference-related criteria are gaining popularity. Noteworthy is the fact that preferences are increasingly used in multi-criteria settings. Mathematical programming, DES, and heuristics are the techniques most frequently used. It is also true that the majority of articles present results based on real data. However, it is important to note that this does not imply that the methods are applied in practice.

**Acknowledgements** We acknowledge the support given to us by the Fonds voor Wetenschappelijk Onderzoek-Vlaanderen (FWO-Vlaanderen) under project G-0309.10.

## References

1. Achieving operating room efficiency through process integration (2003) Healthc Financ Manage 57(3):S1–S8
2. Adan I, Bekkers J, Dellaert N, Vissers J, Yu XT (2009) Patient mix optimisation and stochastic resource requirements: a case study in cardiothoracic surgery planning. Health Care Manag Sci 12(2):129–141. doi:10.1007/s10729-008-9080-9
3. Antonelli D, Taurino T (2010) Application of a patient flow model to a surgery department. In: 2010 IEEE workshop on health care management (WHCM), p 6. doi:10.1109/whcm.2010.5441247

4. Argo JL, Vick CC, Graham LA, Itani KMF, Bishop MJ, Hawn MT (2009) Elective surgical case cancellation in the veterans health administration system: identifying areas for improvement. Am J Surg 198(5):600–606. doi:10.1016/j.amjsurg.2009.07.005
5. Arnaout JPM, Kulbashian S (2008) Maximizing the utilization of operating rooms with stochastic times using simulation. In: 2008 Winter simulation conference, vols 1–5, pp 1617–1623
6. Augusto V, Xie X, Perdomo V (2008) Operating theatre scheduling using lagrangian relaxation. Eur J Ind Eng 2(2):172–189
7. Augusto V, Xie XL, Perdomo V (2010) Operating theatre scheduling with patient recovery in both operating rooms and recovery beds. Comput Ind Eng 58(2):231–238. doi:10.1016/j.cie.2009.04.019
8. Ballard SM, Kuhl ME (2006) The use of simulation to determine maximum capacity in the surgical suite operating room. In: Proceedings of the 2006 winter simulation conference, vols 1–5, pp 433–438
9. Barkaoui K,Dechambre P, Hachicha R (2002) Verification and optimisation of an operating room workflow. In: Proceedings of the 35th annual hawaii international conference on system sciences, pp 2581–2590. doi:10.1109/hicss.2002.994236
10. Basson MD, Butler T (2006) Evaluation of operating room suite efficiency in the veterans health administration system by using data-envelopment analysis. Am J Surg 192(5):649–656. doi:10.1016/j.amjsurg.2006.07.005
11. Batun S, Denton BT, Huschka TR, Schaefer AJ (2010) Operating room pooling and parallel surgery processing under uncertainty. Informs J Comput: ijoc.1100.0396. doi:10.1287/ijoc.1100.0396
12. Baumgart A, Zoeller A, Denz C, Bender HJ, Heinzl A, Badreddin E (2007) Using computer simulation in operating room management: impacts on process engineering and performance. In: Proceedings of the 40th annual hawaii international conference on system sciences, p 10
13. Belien J, Demeulemeester E (2007) Building cyclic master surgery schedules with leveled resulting bed occupancy. Eur J Oper Res 176(2):1185–1204. doi:10.1016/j.ejor.2005.06.063
14. Belien J, Demeulemeester E (2008) A branch-and-price approach for integrating nurse and surgery scheduling. Eur J Oper Res 189(3):652–668. doi:10.1016/j.ejor.2006.10.060
15. Belien J, Demeulemeester E, Cardoen B (2006) Visualizing the demand for various resources as a function of the master surgery schedule: a case study. J Med Syst 30(5):343–350
16. Belien J, Demeulemeester E, Cardoen B (2009) A decision support system for cyclic master surgery scheduling with multiple objectives. J Sched 12(2):147–161. doi:10.1007/s10951-008-0086-4
17. Blake JT, Carter MW (1997) Surgical process scheduling: a structured review. J Soc Health Syst 5(3):17–30
18. Blake JT, Carter MW (2002) A goal programming approach to strategic resource allocation in acute care hospitals. Eur J Oper Res 140(3):541–561
19. Blake JT, Dexter F, Donald J (2002) Operating room managers' use of integer programming for assigning block time to surgical groups: a case study. Anesth Analg 94(1):143–148
20. Blake JT, Donald J (2002) Mount Sinai hospital uses integer programming to allocate operating room time. Interfaces 32(2):63–73
21. Boldy D (1976) Review of application of mathematical-programming to tactical and strategic health and social-services problems. Oper Res Q 27(2):439–448
22. Bowers J, Mould G (2004) Managing uncertainty in orthopaedic trauma theatres. Eur J Oper Res 154(3):599–608. doi:10.1016/S0377-2217(02)00816-0
23. Bowers J, Mould G (2005) Ambulatory care and orthopaedic capacity planning. Health Care Manag Sci 8(1):41–47
24. Calichman MV (2005) Creating an optimal operating room schedule. AORN J 81(3):580–588
25. Cardoen B, Demeulemeester E (2008) Capacity of clinical pathways—a strategic multi-level evaluation tool. J Med Syst 32(6):443–452. doi:10.1007/s10916-008-9150-z

26. Cardoen B, Demeulemeester E (2011) A decision support system for surgery sequencing at uz leuven's day-care department. Int J Inf Technol Decis 10(3):435–450. doi:10.1142/S0219622011004397
27. Cardoen B, Demeulemeester E, Belien J (2009) Optimizing a multiple objective surgical case sequencing problem. Int J Prod Econ 119(2):354–366. doi:10.1016/j.ijpe.2009.03.009
28. Cardoen B, Demeulemeester E, Belien J (2009) Sequencing surgical cases in a day-care environment: an exact branch-and-price approach. Comput Oper Res 36(9):2660–2669. doi:10.1016/j.cor.2008.11.012
29. Cardoen B, Demeulemeester E, Belien J (2010) Operating room planning and scheduling: a literature review. Eur J Oper Res 201(3):921–932. doi:10.1016/j.ejor.2009.04.011
30. Cardoen B, Demeulemeester E, Van der Hoeven J (2010) On the use of planning models in the operating theatre: results of a survey in Flanders. Int J Health Plann Manage 25(4):400–414. doi:10.1002/hpm.1027
31. Centers for Medicare & Medicaid Services OotA (2010) National health expenditures and selected economic indicators, levels and annual percent change: calendar years 2004–2019
32. Centers for Medicare & Medicaid Services OotA, National Health Statistics Group (2009) The nation's health dollar, calendar year 2009: where it went
33. Chaabane S, Meskens N, Guinet A, Laurent M (2006) Comparison of two methods of operating theatre planning: application in Belgian hospital. In: Proceedings of the 2006 international conference on service systems and service management, vols 1 and 2, pp 386–392
34. Conforti D, Guerriero F, Guido R (2010) A multi-objective block scheduling model for the management of surgical operating rooms: new solution approaches via genetic algorithms. In: 2010 IEEE workshop on health care management (WHCM), p 5. doi:10.1109/whcm.2010.5441264
35. Dekhici L, Belkadi K (2010) Operating theatre scheduling under constraints. J Appl Sci 10:1380–1388
36. Dellaert N,Jeunet J (2008) Hospital admission planning to optimize major resources utilization under uncertainty. Paper presented at the 3rd World Conference on Production and Operations Management
37. Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. IIE Trans 35(11):1003–1016. doi:10.1080/07408170390230169
38. Denton B, Viapiano J, Vogl A (2007) Optimization of surgery sequencing and scheduling decisions under uncertainty. Health Care Manag Sci 10(1):13–24
39. Denton BT, Miller AJ, Balasubramanian HJ, Huschka TR (2010) Optimal allocation of surgery blocks to operating rooms under uncertainty. Oper Res 58(4):802–816. doi:10.1287/opre.1090.0791
40. Denton BT, Rahman AS, Nelson H, Bailey AC (2006) Simulation of a multiple operating room surgical suite. In: Proceedings of the 2006 winter simulation conference, vols 1–5, pp 414–424
41. Dexter F (2000) A strategy to decide whether to move the last case of the day in an operating room to another empty operating room to decrease overtime labor costs. Anesth Analg 91(4):925–928
42. Dexter F, Birchansky L, Bernstein JM, Wachtel RE (2009) Case scheduling preferences of one surgeon's cataract surgery patients. Anesth Analg 108(2):579–582. doi:10.1213/ane.0b013e31818f1651
43. Dexter F, Blake JT, Penning DH, Sloan B, Chung P, Lubarsky DA (2002) Use of linear programming to estimate impact of changes in a hospital's operating room time allocation on perioperative variable costs. Anesthesiology 96(3):718–724
44. Dexter F, Epstein RH (2009) Typical savings from each minute reduction in tardy first case of the day starts. Anesth Analg 108(4):1262–1267. doi:10.1213/ane.0b013e31819775cd
45. Dexter F, Ledolter J, Wachtel RE (2005) Tactical decision making for selective expansion of operating room resources incorporating financial criteria and uncertainty in subspecialties' future workloads. Anesth Analg 100(5):1425–1432. doi:10.1213/01.Ane.0000149898.45044.3d

46. Dexter F, Lubarsky DA, Blake JT (2002) Sampling error can significantly affect measured hospital financial performance of surgeons and resulting operating room time allocations. Anesth Analg 95(1):184–188. doi:10.1213/01.Ane.0000018821.20416.5a
47. Dexter F, Macario A, Ledolter J (2007) Identification of systematic underestimation (bias) of case durations during case scheduling would not markedly reduce overutilized operating room time. J Clin Anesth 19(3):198–203. doi:10.1016/j.jclinane.2006.10.009
48. Dexter F, Macario A, Lubarsky DA (2001) The impact on revenue of increasing patient volume at surgical suites with relatively high operating room utilization. Anesth Analg 92(5):1215–1221
49. Dexter F, Macario A, O'Neill L (2000) Scheduling surgical cases into overflow block time—computer simulation of the effects of scheduling strategies on operating room labor costs. Anesth Analg 90(4):980–988
50. Dexter F, Traub RD (2002) How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. Anesth Analg 94(4): 933–942
51. Dexter F, Traub RD, Macario A (2003) How to release allocated operating room time to increase efficiency: predicting which surgical service will have the most underutilized operating room time. Anesth Analg 96(2):507–512. doi:10.1213/01.ane.0000042450.45778.ab
52. Dexter F, Wachtel RE, Epstein RH, Ledolter J, Todd MM (2010) Analysis of operating room allocations to optimize scheduling of specialty rotations for anesthesia trainees. Anesth Analg 111(2):520–524. doi:10.1213/Ane.0b013e3181e2fe5b
53. Does R, Vermaat TMB, Verver JPS, Bisgaard S, Van den Heuvel J (2009) Reducing start time delays in operating rooms. J Qual Technol 41(1):95–109
54. Everett JE (2002) A decision support simulation model for the management of an elective surgery waiting system. Health Care Manag Sci 5(2):89–95. doi:10.1023/a:1014468613635
55. Fei H, Chu C, Meskens N (2009) Solving a tactical operating room planning problem by a column-generation-based heuristic procedure with four criteria. Annu Oper Res 166(1):91–108. doi:10.1007/s10479-008-0413-3
56. Fei H, Chu C, Meskens N, Artiba A (2008) Solving surgical cases assignment problem by a branch-and-price approach. Int J Prod Econ 112(1):96–108. doi:10.1016/j.ijpe.2006.08.030
57. Fei H, Combes C, Chu C, Meskens N (2006) Endoscopies scheduling problem: a case study. Paper presented at the INCOM 06, 17 May 2006
58. Fei H, Meskens N, Chu C (2010) A planning and scheduling problem for an operating theatre using an open scheduling strategy. Comput Ind Eng 58(2):221–230. doi:10.1016/j.cie.2009.02.012
59. Fei HY, Meskens N, Combes C, Chu CB (2009) The endoscopy scheduling problem: a case study with two specialised operating rooms. Int J Prod Econ 120(2):452–462. doi:10.1016/j.ijpe.2007.09.016
60. Ferrand Y, Magazine M, Rao U (2010) Comparing two operating-room-allocation policies for elective and emergency surgeries. In: 2010 winter simulation conference (WSC 2010), pp 2364–2374. doi:10.1109/wsc.2010.5678933
61. Ferrin DM, Miller MJ, Wininger S, Neuendorf MS (2004) Analyzing incentives and scheduling in a major metropolitan hospital operating room through simulation. In: Proceedings of the 2004 winter simulation conference, vols 1 and 2, pp 1975–1980
62. Gonzalez P, Herrero C (2004) Optimal sharing of surgical costs in the presence of queues. Math Meth Oper Res 59(3):435–446. doi:10.1007/s001860400350
63. Guerriero F, Guido R (2011) Operational research in the management of the operating theatre: a survey. Health Care Manag Sci 14(1):89–114. doi:10.1007/s10729-010-9143-6
64. Guinet A, Chaabane S (2003) Operating theatre planning. Int J Prod Econ 85(1):69–81. doi:10.1016/S0925-5273(03)00087-2
65. Gupta D (2007) Surgical suites' operations management. Prod Oper Manag 16(6):689–700. doi:10.1111/j.1937-5956.2007.tb00289.x
66. Gupta D, Denton B (2008) Appointment scheduling in health care: challenges and opportunities. IIE Trans 40(9):800–819. doi:10.1080/07408170802165880

67. Gupta D, Natarajan MK, Gafni A, Wang L, Shilton D, Holder D, Yusuf S (2007) Capacity planning for cardiac catheterization: a case study. Health Policy 82(1):1–11. doi:10.1016/j.healthpol.2006.07.010
68. Hans E, Nieberg T, van Oostrum JM (2007) Optimization in surgery planning. Medium Econometrische Toepassingen 15:20–28
69. Hans E, Wullink G, van Houdenhoven M, Kazemier G (2008) Robust surgery loading. Eur J Oper Res 185(3):1038–1050. doi:10.1016/j.ejor.2006.08.022
70. Hanset A, Meskens N, Duvivier D (2010) Using constraint programming to schedule an operating theatre. In: 2010 IEEE workshop on health care management (WHCM), p 6. doi:10.1109/whcm.2010.5441245
71. Harper PR (2002) A framework for operational modelling of hospital resources. Health Care Manag Sci 5(3):165–173. doi:10.1023/a:1019767900627
72. Helm JE, Lapp M, See BD (2010) Characterizing an effective hospital admissions scheduling and control management system: a genetic algorithm approach. In: 2010 winter simulation conference (WSC 2010), pp 2387–2398. doi:10.1109/wsc.2010.5678935
73. Hongying F, Meskens N, Chengbin C (2007) An operating theatre planning and scheduling problem in the case of a “block scheduling”: strategy. In: 2006 International conference on service systems and service management (IEEE Cat No06EX1406). CD-ROM, p 7
74. Hongying F, Meskens N, El-Darzi E (2010) Evaluating alternative surgery plans with discrete-event simulation model. In: 2010 IEEE workshop on health care management (WHCM), p 6. doi:10.1109/whcm.2010.5441241
75. Hsu VN, de Matta R, Lee CY (2003) Scheduling patients in an ambulatory surgical center. Nav Res Logist 50(3):218–238. doi:10.1002/nav.10060
76. Huschka TR, Denton BT, Gul S, Fowler JW (2007) Bi-criteria evaluation of an outpatient procedure center via simulation. In: Proceedings of the 2007 winter simulation conference, vols 1–5, pp 1489–1497
77. Iser JH, Denton BT, King RE (2008) Heuristics for balancing operating room and post-anesthesia resources under uncertainty. In: 2008 Winter simulation conference, vols 1–5, pp 1601–1608
78. Jebali A, Alouane ABH, Ladet P (2006) Operating rooms scheduling. Int J Prod Econ 99(1–2):52–62. doi:10.1016/j.ijpe.2004.12.006
79. Jebali A, Hadj-Alouane AB, Ladet P (2003) Performance comparison of two strategies for operating room scheduling. In: International symposium on computational intelligence and intelligent informatics
80. Kharraja S, Albert P, Chaabane S (2006) Block scheduling: toward a master surgical schedule. In: Proceedings of the 2006 international conference on service systems and service management, vols 1 and 2, pp 429–435
81. Kim SC, Horowitz I (2002) Scheduling hospital services: the efficacy of elective-surgery quotas. Omega-Int J Manag Sci 30(5):335–346
82. Kolker A (2009) Process modeling of icu patient flow: effect of daily load leveling of elective surgeries on icu diversion. J Med Syst 33(1):27–40. doi:10.1007/s10916-008-9161-9
83. Krempels K-H, Panchenko A (2006) An approach for automated surgery scheduling. In: Sixth international conference on the practice and theory of automated timetabling
84. Kuo PC, Schroeder RA, Mahaffey S, Bollinger RR (2003) Optimization of operating room allocation using linear programming techniques. J Am Coll Surgeons 197(6):889–895. doi:10.1016/j.jamcolsurg.2003.07.006
85. Lagergren M (1998) What is the role and contribution of models to management and research in the health services? A view from Europe. Eur J Oper Res 105(2):257–266. doi:10.1016/s0377-2217(97)00233-6
86. Lamiri M, Augusto V, Xie XL (2008) Patients scheduling in a hospital operating theatre. In: 2008 IEEE international conference on automation science and engineering, vols 1 and 2, pp 627–632
87. Lamiri M, Dreo J, Xiaolan X (2007) Operating room planning with random surgery times. In: Proceedings of the 3rd annual IEEE conference on automation science and engineering, pp 521–526

88. Lamiri M, Grimaud F, Xie XL (2009) Optimization methods for a stochastic surgery planning problem. Int J Prod Econ 120(2):400–410. doi:10.1016/j.ijpe.2008.11.021
89. Lamiri M, Xie XL, Dolgui A, Grimaud F (2008) A stochastic model for operating room planning with elective and emergency demand for surgery. Eur J Oper Res 185(3):1026–1037. doi:10.1016/j.ejor.2006.02.057
90. Lamiri M, Xie XL, Zhang SG (2008) Column generation approach to operating theater planning with elective and emergency patients. IIE Trans 40(9):838–852. doi:10.1080/07408170802165831
91. Lebowitz P (2003) Schedule the short procedure first to improve or efficiency. AORN J 78(4):651–654
92. Litvak N, van Rijsbergen M, Boucherie RJ, van Houdenhoven M (2008) Managing the overflow of intensive care patients. Eur J Oper Res 185(3):998–1010. doi:10.1016/j.ejor.2006.08.021
93. Lovejoy WS, Li Y (2002) Hospital operating room capacity expansion. Manag Sci 48(11):1369–1387
94. Ma G, Demeulemeester E (2010) Assessing the performance of hospital capacity planning through simulation analysis, KBI_1031st edn. Katholieke Universiteit Leuven, Leuven
95. Magerlein JM, Martin JB (1978) Surgical demand scheduling—review. Health Serv Res 13(4):418–433
96. Marcon E, Dexter F (2006) Impact of surgical sequencing on post anesthesia care unit staffing. Health Care Manag Sci 9(1):87–98
97. Marcon E, Dexter F (2007) An observational study of surgeons' sequencing of cases and its impact on postanesthesia care unit and holding area staffing requirements at hospitals. Anesth Analg 105(1):119–126. doi:10.1213/01.ane.0000266495.79050.b0
98. Marcon E, Kharraja S, Simonnet G (2003) The operating theatre planning by the follow-up of the risk of no realization. Int J Prod Econ 85(1):83–90. doi:10.1016/S0925-5273(03)00088-4
99. Marcon E, Kharraja S, Smolski N, Luquet B, Viale JP (2003) Determining the number of beds in the postanesthesia care unit: a computer simulation flow approach. Anesth Analg 96(5):1415–1423. doi:10.1213/01.ane.0000056701.08350.b9
100. Marjamaa RA, Torkki PM, Hirvensalo EJ, Kirvela OA (2009) What is the best workflow for an operating room? a simulation study of five scenarios. Health Care Manag Sci 12(2):142–146. doi:10.1007/s10729-008-9073-8
101. Masursky D, Dexter F, O'Leary CE, Applegeet C, Nussmeier NA (2008) Long-term forecasting of anesthesia workload in operating rooms from changes in a hospital's local population can be inaccurate. Anesth Analg 106(4):1223–1231. doi:10.1213/ane.0b013e318167906c
102. McIntosh C, Dexter F, Epstein RH (2006) The impact of service-specific staffing, case scheduling, turnovers, and first-case starts on anesthesia group and operating room productivity: a tutorial using data from an Australian hospital. Anesth Analg 103(6):1499–1516. doi:10.1213/01.ane.0000244535.54710.28
103. Medpac (2010)Report to congress: medicare payment policy
104. Milliman (2011) 2011 milliman medical index
105. Min DK, Yih Y (2010) An elective surgery scheduling problem considering patient priority. Comput Oper Res 37(6):1091–1099. doi:10.1016/j.cor.2009.09.016
106. Min DK, Yih Y (2010) Scheduling elective surgery under uncertainty and downstream capacity constraints. Eur J Oper Res 206(3):642–652. doi:10.1016/j.ejor.2010.03.014
107. Molina JM, Framinan JM (2009) Testing planning policies for solving the elective case scheduling phase: a real application. In: 35th international conference on operational research applied to health services, Leuven
108. Mulholland MW, Abrahamse P, Bahl V (2005) Linear programming to optimize performance in a department of surgery. J Am Coll Surgeons 200(6):861–868. doi:10.1016/j.jamcollsurg.2005.01.001
109. Niu Q, Peng Q, ElMekkawy T, Tan YY (2007) Performance analysis of the operating room using simulation. In: CDEN and CCEE conference

110. Nouaouri I, Nicolas JC, Jolly D (2009) Scheduling of stabilization surgical cares in case of a disaster. In: Proceedings of the 2009 IEEE international conference on industrial engineering and engineering management (IEEM 2009), pp 1974–1978. doi:10.1109/ieem.2009.5373510
111. Noyan Ogulata S, Erol R (2003) A hierarchical multiple criteria mathematical programming approach for scheduling general surgery operations in large hospitals. J Med Syst 27(3):259–270. doi:10.1023/a:1022575412017
112. OECD (2011) Total expenditure on health 2011
113. Olivares M, Terwiesch C, Cassorla L (2008) Structural estimation of the newsvendor model: an application to reserving operating room time. Manag Sci 54(1):41–55. doi:10.1287/mnsc.1070.0756
114. van OostrumJM, Parlevliet T, Wagelmans APM, Kazemier G (2008) A method for clustering surgical cases to allow master surgical scheduling. Erasmus University Rotterdam, Econometric Institute, Rotterdam
115. Ozkarahan I (2000) Allocation of surgeries to operating rooms by goal programing. J Med Syst 24(6):339–378. doi:10.1023/a:1005548727003
116. Pandit JJ, Dexter F (2009) Lack of sensitivity of staffing for 8-hour sessions to standard deviation in daily actual hours of operating room time used for surgeons with long queues. Anesth Analg 108(6):1910–1915. doi:10.1213/ane.0b013e31819fe7a4
117. Paoletti X, Marty J (2007) Consequences of running more operating theatres than anaesthetists to staff them: a stochastic simulation study. Br J Anaesth 98(4):462–469. doi:10.1093/Bja/Aem003
118. Pariente JMM, Torres JMF, Cia TG (2009) Policies and decision models for solving elective case operating room scheduling. In: CIE: 2009 international conference on computers and industrial engineering, vols 1–3, pp 112–117
119. Pérez Gladish B, Arenas Parra M, Bilbao Terol A, RodrIguez UrIa MV (2005) Management of surgical waiting lists through a possibilistic linear multiobjective programming problem. Appl Math Comput 167(1):477–495. doi:10.1016/j.amc.2004.07.015
120. Persson M, Persson JA (2006) Optimization modelling of hospital operating room planning: analyzing strategies and problem settings. In: Annual conference of OR applied to health services
121. Persson M, Persson JA (2009) Health economic modeling to support surgery management at a Swedish hospital. Omega-Int J Manag Sci 37(4):853–863. doi:10.1016/j.omega.2008.05.007
122. Persson MJ, Persson JA (2010) Analysing management policies for operating room planning using simulation. Health Care Manag Sci 13(2):182–191. doi:10.1007/s10729-009-9122-y
123. Pham DN, Klinkert A (2008) Surgical case scheduling as a generalized job shop scheduling problem. Eur J Oper Res 185(3):1011–1025. doi:10.1016/j.ejor.2006.03.059
124. PierskallaWP, Brailer DJ (1994) Applications of operations research in health care delivery. In: Pollock SM, Rothkopf MH, Barnett A (eds) Operations research and the public sector. Elsevier, Amsterdam, pp 469–505
125. Przasnyski ZH (1986) Operating room scheduling. A literature review. AORN J 44(1):67–79
126. Ramis FJ, Palma JL, Baesler FF (2001) The use of simulation for process improvement at an ambulatory surgery center. In: Proceeding of the 2001 winter simulation conference (Cat No01CH37304), vol 1402, vol (xxxiii + xx + 1678), pp 1401–1404. doi:10.1109/wsc.2001.977462
127. Roland B, Di Martinelly C, Riane F (2006) Operating theatre optimization: a resource-constrained based solving approach. In: Proceedings of the 2006 international conference on service systems and service management, vols 1 and 2, pp 443–448
128. Roland B, Di Martinelly C, Riane F, Pochet Y (2010) Scheduling an operating theatre under human resource constraints. Comput Ind Eng 58(2):212–220. doi:10.1016/j.cie.2009.01.005
129. Ruey-Kei C, Yu-Chen Y (2010) Fuzzy-based dynamic scheduling system for health examination. In: 2010 international conference on machine learning and cybernetics (ICMLC 2010), pp 636–641. doi:10.1109/icmlc.2010.5580551

130. Santibanez P, Begen M, Atkins D (2007) Surgical block scheduling in a system of hospitals: an application to resource and wait list management in a British Columbia health authority. Health Care Manag Sci 10(3):269–282
131. Schoenmeyr T, Dunn PF, Garnarnik D, Levi R, Berger DL, Daily BJ, Levine WC, Sandberg WS (2009) A model for understanding the impacts of demand and capacity on waiting time to enter a congested recovery room. Anesthesiology 110(6):1293–1304
132. Sciomachen A, Tanfani E, Testi A (2005) Simulation models for optimal schedules of operating theatres. Int J Simul 6:26–34
133. Sieber TJ, Leibundgut DL (2002) Operating room management and strategies in Switzerland: results of a survey. Eur J Anaesth 19(06):415–423. doi:10.1017/S0265021502000662
134. Slack N (1999) The Blackwell encyclopedic dictionary of operations management. Blackwell, Oxford
135. Smith-Daniels VL, Schweikhart SB, Smith-Daniels DE (1988) Capacity management in health-care services—review and future-research directions. Decis Sci 19(4):889–919
136. Souki M, Ben Youssef S, Rebai A (2009) Memetic algorithm for operating room admissions. In: CIE: 2009 international conference on computers and industrial engineering, vols 1–3, pp 519–524
137. Souki M, Rebai A (2009) Memetic differential evolution algorithm for operating room scheduling. In: CIE: 2009 international conference on computers and industrial engineering, vols 1–3, pp 845–850
138. Stanciu A, Vargas L, May J (2010) A revenue management approach for managing operating room capacity. In: 2010 winter simulation conference (WSC 2010), pp 2444–2454. doi:10.1109/wsc.2010.5678940
139. Steins K, Persson F, Holmer M (2010) Increasing utilization in a hospital operating department using simulation modeling. Simul-Trans Soc Model Simul 86(8–9):463–480. doi:10.1177/0037549709359355
140. Tan YY, ElMekkawy TY, Peng Q, Oppenheimer L (2007) Mathematical programming for the scheduling of elective patients in the operating room department. In: CDEN/C$^2$E$^2$, Winnipeg
141. Tancrez J-S, Roland B, Cordier J-P, Riane F (2009) How stochasticity and emergencies disrupt the surgical schedule. In: McClean S, Millard P, El-Darzi E, Nugent C (eds) Intelligent patient management, vol 189. Studies in computational intelligence. Springer, New York, pp 221–239. doi:10.1007/978-3-642-00179-6_14
142. Tanfani E, Testi A (2010) Improving surgery department performance via simulation and optimization. In: 2010 IEEE workshop on health care management (WHCM), p 6. doi:10.1109/whcm.2010.5441255
143. Tanfani E, Testi A (2010) A pre-assignment heuristic algorithm for the master surgical schedule problem (mssp). Annu Oper Res 178(1):105–119. doi:10.1007/s10479-009-0568-6
144. Testi A, Tanfani E (2009) Tactical and operational decisions for operating room planning: efficiency and welfare implications. Health Care Manag Sci 12(4):363–373. doi:10.1007/s10729-008-9093-4
145. Testi A, Tanfani E, Torre G (2007) A three-phase approach for operating theatre schedules. Health Care Manag Sci 10(2):163–172
146. Testi A, Tanfani E, Valente R, Ansaldo G, Torre G (2008) Prioritizing surgical waiting lists. J Eval Clin Pract 14(1):59–64
147. Tsoy G, Arnaout JP, Smith T, Rabadi G (2004) A genetic algorithm approach for surgery operating rooms scheduling problem. Manag Dangerous World 299–304
148. Tyler DC, Pasquariello CA, Chen CH (2003) Determining optimum operating room utilization. Anesth Analg 96(4):1114–1121. doi:10.1213/01.Ane.0000050561.41552.A6
149. Value in health care: Current state and future directions (2011)
150. van der Lans M, Hans E, Hurink JL, Wullink G, van Houdenhoven M, Kazemier G (2006) Anticipating urgent surgery in operating room departments. BETA working paper WP-158, University of Twente
151. Van Houdenhoven M, Hans E, Klein J, Wullink G, Kazemier G (2007) A norm utilisation for scarce hospital resources: evidence from operating rooms in a Dutch university hospital. J Med Syst 31(4):231–236. doi:10.1007/s10916-007-9060-5

152. Van Houdenhoven M, Oostrum JMV, Wullink G, Hans E, Hurink JL, Bakker J, Kazemier G (2008) Fewer intensive care unit refusals and a higher capacity utilization by using a cyclic surgical case schedule. J Crit Care 23(2):222–226. doi:10.1016/j.jcrc.2007.07.002
153. van Oostrum JM, Van Houdenhoven M, Hurink JL, Hans EW, Wullink G, Kazemier G (2008) A master surgical scheduling approach for cyclic scheduling in operating room departments. Or Spectrum 30(2):355–374. doi:10.1007/s00291-006-0068-x
154. VanBerkel PT, Blake JT (2007) A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. Health Care Manag Sci 10(4):373–385
155. Velasquez R, Melo T, Kufer KH (2008) Tactical operating theatre scheduling: efficient appointment assignment. Oper Res Proc 2007:303–308
156. Vissers JMH, Adan IJBF, Bekkers JA (2005) Patient mix optimization in tactical cardiothoracic surgery planning: a case study. IMA J Manag Math 16(3):281–304304. doi:10.1093/imaman/dpi023
157. Wachtel RE, Dexter F (2007) A simple method for deciding when patients should be ready on the day of surgery without procedure-specific data. Anesth Analg 105(1):127–140. doi:10.1213/01.ane.0000266468.09733.4d
158. Wachtel RE, Dexter F (2008) Tactical increases in operating room block time for capacity planning should not be based on utilization. Anesth Analg 106(1):215–226. doi:106/1/215 [pii]
159. Wachtel RE, Dexter F (2009) Influence of the operating room schedule on tardiness from scheduled start times. Anesth Analg 108(6):1889–1901. doi:10.1213/ane.0b013e31819f9f0c
160. Wachtel RE, Dexter F (2009) Reducing tardiness from scheduled start times by making adjustments to the operating room schedule. Anesth Analg 108(6):1902–1909. doi:10.1213/ane.0b013e31819f9fd2
161. Wang D, Xu JP (2008) A fuzzy multi-objective optimizing scheduling for operation room in hospital. Ind Eng Eng Manag 614–618
162. Wang QN (2004) Modeling and analysis of high risk patient queues. Eur J Oper Res 155(2):502–515. doi:10.1016/s0377-2217(02)00916-5
163. Wullink G, Van Houdenhoven M, Hans EW, van Oostrum JM, van der Lans M, Kazemier G (2007) Closing emergency operating rooms improves efficiency. J Med Syst 31(6):543–546. doi:10.1007/s10916-007-9096-6
164. Ya L, Chengbin C, Kanliang W (2010) Aggregated state dynamic programming for operating theater planning. In: 2010 IEEE international conference on automation science and engineering (CASE 2010), pp 1013–1018. doi:10.1109/coase.2010.5584004
165. Yang Y, Sullivan KM, Wang PP, Naidu KD (2000) Applications of computer simulation in medical scheduling. In: Proceedings of the fifth joint conference on information sciences, vols 1 and 2, pp A836–A841
166. Yu W, Jiafu T, Gang Q (2010) A genetic algorithm for solving patient-priority-based elective surgery scheduling problem. In: Life system modeling and intelligent computing international conference on life system modeling and simulation, LSMS 2010, and International conference on intelligent computing for sustainable energy and environment, ICSEE 2010, pp 297–304|xvi + 518. doi:10.1007/978-3-642-15597-0_33
167. Zhang B, Murali P, Dessouky MM, Belson D (2009) A mixed integer programming approach for allocating operating room capacity. J Oper Res Soc 60(5):663–673. doi:10.1057/palgrave.jors.2602596
168. Zonderland ME, Boucherie RJ, Litvak N, Vleggeert-Lankamp C (2010) Planning and scheduling of semi-urgent surgeries. Health Care Manag Sci 13(3):256–267. doi:10.1007/s10729-010-9127-6

# Chapter 6
# The Modeling, Analysis, and Management of Intensive Care Units

**Theologos Bountourelis, M. Yasin Ulukus, Jeffrey P. Kharoufeh, and Spencer G. Nabors**

## 1 Introduction

An intensive care unit (ICU) is a limited-capacity, resource-intensive unit in a healthcare facility designed to provide continuously monitored, intensive care and temporary support to critically ill patients with a broad range of health conditions. Some of these conditions include (but are not limited to) acute or moderate respiratory failure, chronic respiratory failure, infections or severe infections, sepsis and severe sepsis infections, renal failure, neurological conditions, and bleeding and clotting. Patients are typically admitted to the ICU via the emergency department, postoperatively from various surgical theaters, out of surgical or medical units, or they are transferred directly from other hospitals. The length of stay in an ICU can range from a few hours to days or even weeks. When the patient no longer requires intensive care or intensive monitoring, he/she is typically transferred from the ICU to a so-called "step-down" unit, which provides less critical care, before being moved to a regular hospital bed and ultimately being discharged from the hospital. Although most ICU patients eventually recover, some may die despite receiving the best possible medical and nursing care.

A very significant proportion of overall hospital operating costs can be attributed to the ICUs. The average daily cost of ICU care is higher than that of an average non-monitored unit due to increased staffing, invasive procedures, expensive associated therapies, and expensive equipment (Milbrandt and Kersten 2008). It has been

T. Bountourelis • M.Y. Ulukus • J.P. Kharoufeh (✉)
Department of Industrial Engineering, University of Pittsburgh,
1048 Benedum Hall, Pittsburgh, PA, USA
e-mail: bountourelis@gmail.com; myu1@pitt.edu; jkharouf@pitt.edu

S.G. Nabors
Department of Critical Care Medicine, University of Pittsburgh Medical Center,
Pittsburgh, PA, USA
e-mail: naborssg@upmc.edu

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_6, © Springer Science+Business Media New York 2013

estimated that in the past 15 years, the cost of caring for critically ill patients has risen to 1% of the total gross domestic product in the USA or 20% of total hospital costs (Halpern et al. 2004). By 2010 more than one-third (36.2%) of all Medicare hospitalizations had ICU or critical care unit (CCU) care at some point during the hospital stay. In a recent 2005 study, Kahn and Angus (2006) noted that more than half of the increase in total Medicare hospitalizations over the study period was due to additional ICU hospitalizations. While it is comforting to note that the operating cost of ICU care has remained fairly stable ($2,616 versus $2,575 for 1994 and 2004, respectively), reimbursement for these costs has not remained stable (Dasta et al. 2005). Cooper and Linde-Zwirble (2004) have noted that only 83% of costs were paid for ICU patients as compared to 105% for floor patients (a generic term for non-ICU patients), generating a $5.8 billion dollar loss to hospitals when ICU care is required. Ultimately, with a rising proportion of costs and increasing associated operational losses, hospital administrators must address this shift in healthcare resource management, as well as the cost impact, of caring for critically ill patients.

While many researchers and practitioners have devised models for analyzing some area of the hospital (e.g., the emergency department), models for analyzing and managing ICUs are relatively sparse. This may be due to the fact that ICUs present unique modeling challenges. First, patients are admitted to the ICU from a number of disparate sources (via the emergency department (ED), postoperatively from various surgical theaters, as transfers from surgical or medical units, and as diversions from other hospitals). Second, a heterogeneous patient population complicates the task of estimating the length of stay in the ICU. For example, dramatic differences can be seen between medical and surgical patients or coronary care and sepsis patients. Moreover, the evolution of physiological indicators of health can vary dramatically between patients experiencing the same illness. Third, daily operations within the ICU are typically driven by a number of internal factors including clinical practices and policies, staff scheduling, bed availability, and the persistent influx of new critical patients. Finally, because ICUs ultimately discharge their patients to inpatient units that offer varying levels of care, the status of those units (e.g., staffing levels, bed availability, resource availability) directly affects bed occupancy rates and admission and discharge decisions within the ICU.

It is vital to gain an understanding of day-to-day ICU operations in order to identify opportunities for improvement and to quantify potential gains by either altering the existing configuration of the ICU or by examining the policies and procedures that dictate admission and discharge decisions in the face of uncertain demand and constrained resources. Ideally, models of ICUs should incorporate the concepts of patient acuity, heterogeneous patient populations, interactions with other units in the hospital, unit staffing, and patient flow within constrained systems. By uniting the expertise of operations researchers and medical practitioners, opportunities to realize significant improvements in the management and operation of ICUs now abound. The main objective of this chapter is to highlight significant contributions toward that end and to describe a specific case study related to ICU modeling currently being undertaken by the authors.

The remainder of the chapter is organized as follows: Sect. 2 summarizes and expounds upon three of the most challenging issues related to modeling and analyzing ICU operations. Section 3 provides a brief overview of the most important indicators of ICU performance. Next, Sect. 4 surveys the literature related to the modeling and analysis of healthcare facilities with a particular emphasis on studies most relevant to ICUs. Finally, in Sect. 5, we describe an ongoing project related to ICU operations at a major healthcare facility in Pittsburgh, Pennsylvania.

## 2 ICU Issues and Challenges

In Sect. 1 it was argued that ICUs are unique among hospital units in that all of the patients admitted to the ICU are critically ill and, therefore, require specialized care and resources. In this section, we more fully elucidate the factors that make modeling, analyzing, and managing ICUs challenging.

**Patient Mix.** Some hospitals utilize ICUs that are dedicated to specific patient types. For example, a single hospital might possess a medical ICU as well as a surgical ICU that only handles surgery patients. Smaller hospitals may have only a single ICU that handles all patient types and illnesses so that prioritization of patients is a critical issue. Irrespective of the configuration, all ICUs have only a finite (and relatively small) number of beds. As noted in Sect. 1, several types of illnesses can send a patient to the ICU, and depending on the type and acuity of the illness, the regimen prescribed to the patient, and physiological differences between patients, it may be difficult to model a "typical" ICU patient's flow through the system. Even patients suffering from the same illness can have vastly different experiences in the ICU. This inherent heterogeneity makes it difficult to assess an "average" patient's length of stay (LOS) or the overall impact on other units of the hospital.

Demand for an ICU bed can arrive from the emergency department, post-operatively (from surgical theaters), out of surgical or medical units, or from other hospitals (in the form of diversions). Demand generated by the emergency department, inpatient, non-monitored units, or from other hospitals is *unscheduled* and typically very critical; a delay of only a few hours can have a disproportionately negative effect on the patient's future health status. On the other hand, a surgical patient is typically a *scheduled* bed request, and the patient often resides in a post-surgery recovery room where he/she receives the appropriate level of care until an ICU bed is available. A delay on such a patient's admission to the ICU might have only a limited impact on the patient's future health status.

**Patient Blocking.** As the most resource-intensive unit in a healthcare facility, the ICU cannot (and should not) be analyzed in isolation from other upstream or downstream inpatient units and hospital departments (e.g., non-monitored units, the emergency department). Although their cases are unique, the vast majority of

ICU patients follow a similar flow through the system. As soon as the patient is medically able to leave the ICU, a request is made by the attending physician to transfer the patient to a *telemetry* or *step-down* bed which provides monitored care that is less intensive than the care received in the ICU. From the step-down bed, the patient is usually transferred to a regular hospital bed from which he/she is ultimately discharged from the hospital (assuming recovery). However, patients may have trouble getting into, and out of, the ICU due to the *blocking phenomenon*. Blocking in ICUs can occur in the following ways. First, an arriving patient who requires intensive care may be denied admission to the ICU due to bed unavailability (i.e., all ICU beds are occupied, or the unoccupied beds may not be suitable for the new patient). In such cases, an existing ICU patient may be (prematurely) discharged to a step-down unit to accommodate the new patient; otherwise, the arriving patient must be diverted to another hospital, thereby delaying the critical care they need. The second type of blocking is experienced by existing ICU patients who cannot be transferred out of the ICU due to unavailability of a step-down or hospital bed. That is, patients who are medically able to leave the ICU might be forced to stay there until a downstream bed is available. Obviously, this type of blocking serves to exacerbate the former type as fewer ICU patients can be admitted. It is important to note that patients who are prematurely transferred out of the ICU may be forced to reenter it if their medical condition worsens at any downstream stage. (Such patients are often referred to as *bounce backs*.) Blocking also stems from a shortfall of external nursing care facilities or due to institutional rules and procedures. For example, if transfers to external facilities are prohibited during the weekend, step-down and hospital beds can all become occupied over the weekend, thereby blocking the transfer of ICU patients.

In extreme cases, patients must receive ICU-level care in non-monitored beds. For instance, equipment may need to be moved to the emergency department to administer ICU-level care and stabilize the patient until an ICU bed is available. Blocking has devastatingly negative effects on the operation and flow of the ICU and other hospital units. For example, unavailability of an ICU bed may cause surgery cancelations for procedures that require post-surgery intensive care. For these reasons, performance measures related to blocking have been proposed in order to assess its likelihood and overall cost impact. We will further discuss ICU patient blocking in Sects. 3 and 5.

**Clinical Practices and Procedures.** Modeling an ICU (and its interacting units) is complicated by clinical practices and procedures that are often employed to control the flow of patients into and out of the ICU. These practices can be viewed as "rules of thumb" for daily ICU operations, and they can vary dramatically across hospitals or across doctors within the same hospital. For example, a medical ICU (MICU) might accept post-surgery patients if no beds are available in the surgical ICU (SICU). As noted earlier, some ICUs may discharge patients early if the ICU occupancy reaches its peak (Diwas and Terwiesch 2012). Patient demographics, proximity to other clinical facilities, and the level of outpatient care available may also affect the length of stay and the patient mix. For instance, it was reported in Friedman and Steiner (1999) that the length of stay in the ICU is shorter for uninsured patients than for other privately insured patients. So a common set of rules for transferring or discharging patients does not exist, thereby complicating the task of modeling these important dynamics. The operations research modeler must decide which rules to include and which to exclude, for each rule added to the model can potentially add one or more control parameters that need to be estimated from data. For example, it is clear that not *all* surgery patients are diverted to the MICU following surgery, but the exact proportion that is diverted may be difficult to estimate, even for clinical staff members who are involved in day-to-day decision making. Out of necessity, these parameters are usually assigned initial values based on an educated guess and are subsequently updated when the (simulation or analytical) model is calibrated. But if the number of unknown parameters is significant, the process of calibrating the model can be nontrivial and very time consuming. We elaborate upon these challenges further in Sect. 5.

Section 3 provides a succinct discussion of important ICU measures of performance. Because our main interest lies in the operational aspects of the ICU, we primarily focus on those issues as opposed to medical outcomes.

## 3 Performance Measures of ICUs

Historically, research related to measures of ICU performance has been focused on one of two areas: (1) medical/clinical outcomes or (2) administrative/institutional outcomes. For the former, many investigators have pursued objective measures of clinical outcomes such as morbidity, mortality rates, or quality-of-life indicators. For the latter, investigators have typically focused on operational aspects of the ICU with an emphasis on capacity planning, length of stay, bed occupancy rate, and overall daily operating costs. While there has been some overlap between these two broad categories, when considering secondary outcomes or studies focused on volume status or regionalization (see Angus and Black 2004; Harrison et al. 2004; Marcin and Romano 2004; Tarnow-Mordi et al. 2000), most in-depth reviews or original research have independently focused on either clinical or administrative outcomes, but not both. Of course, some clinical outcomes are closely tied to administrative outcomes. For example, Iapichino et al. (2004) have shown empirically that higher ICU occupancy rates lead to higher mortality rates. For an extensive survey and in-depth discussion of measures of ICU performance, the interested reader is referred to Garland (2005).

While clinical and administrative outcomes may differ, most hospitals share a common set of goals for the ICU. These include (but are not limited to) a reduction in the patient's length of stay, a reduction in the mortality rate, a reduction in the rate of ICU readmissions, a reduction in errors or adverse events, and an overall increase in the quality of care received by patients. However, these improvements do not come without significant cost, so decision makers are often faced with multiple conflicting objectives of improving patient care and doing so in a cost-effective manner. Many of the performance-based goals are closely related to the operational aspects of the ICU, which is the focus of this chapter. Therefore, we now elucidate some of the critical operational measures of performance of ICUs:

1. *Bed occupancy rate*. This measure is commonly referred to as *bed utilization* and can be interpreted as the long-run fraction of ICU beds that are occupied. As noted by Green (2006), many hospitals target 85% bed utilization to justify their existing capacity levels or to make the case for capacity expansion. However, such a high utilization rate may lead to undesirable outcomes, such as excessive patient blocking.
2. *Length of stay*. The length of stay (LOS) in the ICU is usually measured by the number of hours or days spent there by the patient. (In Sect. 5, we will argue that the LOS should be viewed as the sum of two separate durations: the medical length of stay plus the time spent waiting for a transfer.) It is desirable to decrease the LOS in the ICU in order to reduce daily costs; however, a premature transfer can lead to detrimental health outcomes and potentially greater congestion (see Garland 2005).
3. *Blocking probability*. This measure represents the long-run likelihood that a patient arriving to the ICU is unable to secure a bed or that an existing ICU patient who is medically able to move cannot move due to unavailability of a downstream bed. Assessing this probability is a nontrivial task requiring significant and detailed data collection related to the timing of transfer requests and actual patient movements. Ideally, only a very small fraction of patients will experience blocking (of any kind); however, very few studies (cf. van Dijk and Kortbeek 2009; Dobson et al. 2010) have explored this critical measure of ICU performance in depth.
4. *ICU readmission rate*. This readmission rate is the long-run fraction of patients who are transferred out of the ICU and subsequently readmitted due to a deterioration of their health status either in a step-down or regular hospital bed or after being discharged from the hospital. Regarding this measure, Garland (2005) states the following:

   > Its potential value derives from observations that readmitted patients have higher mortality rates and longer lengths of stay. However, for it to be a meaningful surrogate requires the following: (1) a detrimental outcome occurred after the patient left that was due to a problem present prior to the original ICU discharge; and (2) it would not have occurred if the patient had remained longer in the ICU. There are no data that have demonstrated this. The optimal ICU readmission rate is unknown, and a low rate might actually indicate that, on average, patients are inappropriately remaining in the ICU longer than necessary, increasing the costs of care and their exposure to virulent pathogens.

   While it is indeed difficult to determine if a discharged patient would have derived benefit from a longer ICU stay, readmissions have a significant operational impact on the ICU as they impose additional demand on an already constrained system. Furthermore, the presence of readmissions requires that some thought be given to a patient admission priority scheme that balances the severity of the patients' conditions and timely delivery of critical care. For these reasons, we contend that the readmission rate can be a significant factor that should be considered when assessing ICU performance.

5. *Total average delay*. A patient is said to be *delayed* if they are medically able to be moved from their current unit to another (less intensive) unit but they must wait for the availability of a downstream bed. Typically, the delay can range from a few minutes to several hours or even days. It is an aggregate measure of overall system performance and can be used to estimate the effect of blocking on operating costs.

The complex task of collecting and analyzing data related to ICU performance measures is made easier when clinical information systems are available. Although some clinical information systems focus on important aspects such as computerized physician order entry and individual patient tracking information, few gather clinical information needed to provide a panoramic view of ICU performance and detailed medical or administrative outcomes (e.g., mortality rate, length of stay, severity of illness, clinical scores, nosocomial infections, adverse events, and adherence to good clinical practices). Those that do exist can be used for the daily care of patients while providing valuable data for in-depth analysis of ICU performance. As an example, the Epimed Monitor system is a continuous patient monitor system that tracks and stores patient data, usually vital signs and location data, for large-scale statistical analysis and research applications.

## 4 Methods for Modeling and Analyzing ICUs

In this section, we review the techniques predominantly used to model, analyze, and manage ICUs. Because the evolution of the ICU is both temporally dynamic and stochastic, it has most often been analyzed via stochastic discrete-event simulation models. However, some analytical queueing models have recently emerged that exemplify the value of simple mathematical expressions for assessing the salient features of ICU performance. Naturally, the models reviewed here consider different aspects of system performance. We begin with a review of relevant discrete-event simulation models.

**Simulation Modeling of ICUs.** Discrete-event simulation models of healthcare systems have existed for several decades and have been used for a variety of purposes (e.g., estimating and optimizing patient flow, estimating bed utilization, facility design/redesign, inpatient and outpatient scheduling, capacity planning, and staffing). One appealing aspect of a discrete-event simulation model is its ability to capture complex dynamics between interacting components of the healthcare system and assess important performance parameters under a variety of operational scenarios or design alternatives while considering resource constraints. Additionally, the level of detail (or fidelity) of a simulation model is at the discretion of the modeler.

Many different units of a healthcare facility can benefit from insights obtained through discrete-event simulation models; however, the ICU is an especially ideal candidate for simulation modeling for at least the following reasons: First, the ICU has a finite number of resources (namely, beds, staff, and equipment) that can be modeled fairly easily within most commercial simulation packages. It can also consider the impact of other hospital departments on the ICU. Second, a simulation model allows for a heterogeneous patient population originating from a variety of sources. Finally, it can be used as a tool to evaluate new unit design alternatives, unit redesign alternatives, staffing policies, and ICU admission, bumping, and discharge policies. A simulation model can also provide real-time animation of important, dynamic performance parameters such as the number of occupied beds, patient delay time, and the number of blocked patients waiting to enter or exit the ICU.

Several steps are involved in creating, executing, and using a discrete-event simulation model. An excellent summary of the basic procedure can be found in the text by Law and Kelton (2000). These steps are summarized as follows:

1. Identifying and formulating the problem
2. Collecting data to build a preliminary model
3. Assembling the appropriate model inputs using the collected data
4. Verifying the computer simulation model
5. Designing an experiment to evaluate the performance measure(s) of interest
6. Executing the replications of the simulation model
7. Statistically analyzing and interpreting the results of the experiments

This sequential procedure is well known in the operations research and industrial engineering communities where simulation models have played a pivotal role in the design, modeling, and analysis of manufacturing systems. However, the modeling of healthcare systems in general, and of ICUs in particular, presents additional challenges not faced in other arenas. As noted by Harper and Pitt (2004), healthcare projects vary in nature, and the modeler must consider the individual needs of the client organization. Additionally, the modeler must consider the complexity, interactivity, and resolution of the model; its generality, and the political context of the client organization, as well as the availability and quality of relevant data. Lowery et al. (1994) also note that:

> Patients are not parts, physicians and nurses are not machines, the clinical environment is not just another job shop, and providing care is not manufacturing health. The healthcare environment is far more complex than that.

The contemporary literature devoted to healthcare systems modeling discusses a number of challenges related to healthcare simulation that are pertinent to simulation of ICUs. Those challenges range from modeling the patient mix (patient heterogeneity), modeling LOS and its distribution, model calibration and validation, interpretation of simulation results, and dissemination of those results to the clinical community (cf. Carter and Blake 2005; Davies and Davies 1995; Harper 2002; Lowery 1996; Seymour 2001; Standridge 1999). Additionally, a number of review articles or surveys discussing the application of simulation methodology to healthcare delivery systems have been contributed; a representative sampling of these includes Brailsford et al. (2009); Fone et al. (2003); Jun et al. (1999); Mustafee et al. (2010); Jacobson et al. (2006). We especially draw the reader's attention to the cogent survey due to Jacobson et al. (2006). That work reviews a number of specific models and applications while providing an extensive bibliography. The bulk of contributions can be partitioned into two main categories: (1) estimating and optimizing patient flow and (2) reducing operating costs while ensuring a high level of patient care. We next review a few articles in each of these areas, particularly those most pertinent to ICU operations.

The first category is driven by the need of many healthcare providers to streamline patient flow in a way that minimizes delays and maximizes patient satisfaction. Hence, a significant part of the literature adopts certain patient flow characteristics as an objective and as a starting point for their analysis. In this context, they examine necessary organizational and resource changes as well as additional interventions needed to optimize patient admission rates and delay times. Cochran and Bharti (2006) presented the methodology underpinning the compilation and validation of a large-scale simulation model that includes the ICU and *telemetry units*. (A telemetry unit is one that is equipped with the capacity for continuous cardiac monitoring.) The validation was done using real hospital data for an existing facility. The model was subsequently used to solve a stochastic bed-balancing problem that leads to balanced bed utilization rates that minimize the blocking of beds across different units. Lowery (1992) presented a simulation model of the surgical suite and critical care areas of a large hospital. The model simulates the flow of patients through the ICUs, including the operating room, postanesthesia recovery unit, surgical ICU, intermediate surgical care unit, coronary care unit, intermediate coronary care unit, telemetry unit, medical ICU, and the ventilator unit. The primary objective was to assess the impact of critical care bed configurations on performance measures such as bed utilization, the number of patients denied admission, bumped, or accommodated on alternative units. A *bumped* patient is one whose current, or intended, location is taken for use by another presumably more critical patient. Harper and Shahani (2002) demonstrated the importance of modeling the various types of patient flows when simulating bed occupancies and patient rejection rates. They showed that the explicit modeling of patient mix results in higher-fidelity models that are able to capture the bed occupancy fluctuations over time.

The second broad category is motivated by the pressing need to simultaneously reduce hospital operating costs while maintaining (or exceeding) the quality of healthcare provided to patients. While some of the existing literature includes simulation models to inform resource allocation decisions in budget-constrained environments (see Jacobson et al. 2006 for a good sampling), models that consider the effect of bed and staffing allocations have received the most attention. The latter category of healthcare simulation models is especially relevant to the modeling, analysis, and management of ICUs.

Williams (1983) developed a simulation model to select the number of beds needed to meet the hospital standards of care. The model was calibrated with real patient data collected over a period of 12 months. The ICU is divided into a cardiac and a medical subunit. If the destination ICU subunit is full, arriving patients are transferred to the other, whereas if the ICU is full, the less severely ill patients are transferred to a regular ward to make room for the new patients. By varying the number of ICU beds between 7 and 15, the author reports the total number of admissions, transfers, premature discharges, and empty beds. Based on the simulation model, the hospital administration chose a configuration of 11 ICU beds that allowed for a premature discharge rate of 2.7%. A smaller unit increased this rate significantly, whereas a larger unit increased the ICU operating costs.

In Vassilacopoulos (1985), a simulation model was used to determine the number of beds required by inpatient units to satisfy several measures of operating efficiency such as high occupancy rates, immediate admission of emergency patients, and short patient waiting lists. Ridge et al. (1998) investigated the relationship between admission rules and rejection rates in a single ICU unit where the patients are classified as either emergency or nonemergency. Emergency patients are admitted upon arrival if there is an available ICU bed, and they are diverted to another ICU otherwise; nonemergency patients are accepted only if the number of free beds exceeds a minimum threshold. Patients who are denied admission retry as emergency patients after waiting a random amount of time. Ridge et al. (1998) considered a simplified version of the model for which an analytical solution can be derived. They created a simulation model for the simplified version and estimated the warm-up period. The warm-up period was used to execute the original simulation model to examine (1) the effect of varying the number of ICU beds, (2) the effect of the deferral period, (3) the effect of changing the number of beds for emergency conditions, (4) the typical number of free beds at midnight versus the day of the week, and (5) the typical free bed probability distribution. Highlighted is the nonlinear relationship between the number of beds, the occupancy levels, and the transfer rates. Additionally, they discussed the practice of reserving beds for emergency patients and how their model can be used as a capacity planning decision tool.

Harper and Shahani (2002) sought to unveil nonlinear relationships between bed allocation, bed occupancy rates, and the denied admission rate. The model was designed to highlight controllable variables that can improve the quality of patient care and staff working conditions, subject to budget constraints. Costa et al. (2003) illustrated the danger in using only average values to determine the number of ICU beds in the face of nonlinearity and system variability. Considered were the bed occupancy rate, emergency transfers (rejections), and elective deferrals. Using a simulation model, they examined three patient classes: emergency, elective (planned), and deferred. Emergency patients are always admitted if there is capacity; otherwise, they are diverted to another hospital or nursed temporarily at another unit in the same hospital. Elective or planned patients are usually deferred; however, if the number of times they are deferred exceeds a threshold value, their priority status becomes the same as an “emergency” patient. The severity of illness was measured by the APACHE II score, a commonly used system for classifying the severity of disease (cf. Knaus et al. 1985). Costa et al. (2003) then used Classification and Regression Tree (CART) analysis to group patients based on a number of explanatory variables with respect to a chosen criterion for the classification, such as length of stay. Consequently, four groups of ICU patients were identified (emergency short stay, emergency normal stay, elective short stay, and elective normal stay). The model was validated using real data from the Norfolk and Norwich University NHS Trust ICU. By considering alternative bed configurations, it was shown that using only average values, the bed requirements can be underestimated. Similarly, Shahani et al. (2008) presented a simulation model for a single critical care unit (CCU). Emphasis was given to the modeling of the patient mix using the CART method to obtain LOS distributions for statistically different patient categories. The model was used to implement changes in bed numbers, patient LOS, discharge policies, and the transfer of long-stay patients out of the CCU in order to explore their effects on bed occupancy and refused admissions. Lowery (1993) constructed and validated a critical care simulation model by examining the occupancy rates and the number of refused admissions to the unit due to lack of available beds.

It is a commonly held belief within the academic and clinical communities that healthcare simulation models are not generally followed up with a similar volume of implementation evidence (Taylor et al. 2009). Consequently, the true value of simulation modeling may not be fully understood or appreciated by the clinical community (Forsberg et al. 2011). This perceived disconnect between modeling and implementation has been discussed by a few researchers. Some have suggested practical ways to make simulation modeling an integral part of the clinical decision making process. For instance, Lowery et al. (1994) discussed the intrinsic barriers that exist in healthcare that make the application of simulation models more difficult than in traditional modeling settings, such as manufacturing systems. They provided remarks made by simulation and management professionals that highlight the involvement of clinical personnel in the model development process and the critical importance of managerial support when contemplating the implementation of complex, highly technical simulation tools. In Lowery (1996), the author highlighted various factors that affect the application of simulation models ranging from model complexity and input distributions to interpretation of findings. Eldabi (2009) discussed the difficult (and often contradictory) nature of healthcare problems and proposes a methodology towards successful implementation.

The manifold benefits of discrete-event simulation models for analyzing and managing ICUs are not limited to the statistical results that stem from these models or decisions that might arise from their analysis. As noted by Lowery (1996), the entire model development process forces the modeler to document the detailed operations of the unit in a systematic way while requiring justification for input modeling assumptions. This structured process is likely to reveal problem areas and opportunities for improvement, even before the model is created.

In summary, discrete-event simulation models of ICUs can be very effective for evaluating alternative bed configurations, operating policies, and performance parameters while incorporating detailed features of specific units. However, they also have some important shortcomings. For example, a great deal of data is required to initiate a simulation study, validation of the models can be very challenging, and implementation (execution) of the model might be very time consuming. Furthermore, implementation of the simulation model might be limited to only a few select individuals in the organization who are familiar with specialized simulation software packages or programming languages. Alternatively, analytical queueing models can be used to provide valuable insights into the salient features of these complex systems with fewer data requirements, though at a cost of model fidelity. We next discuss recent developments in analytical ICU queueing models.

**Queueing Models of ICUs.** In simple terms, a *queue* is a waiting line. The origin of the word can be traced to a sixteenth century Latin word *cauda*, which simply means "tail." We all experience queues in daily life, whether waiting at a traffic light, in line at the grocery store, or waiting to be seen by a doctor in the emergency room. Queues are formed when the demand for some kind of service exceeds the capacity to meet that demand. This shortfall in service capacity stems from random variations in the demand arrival process, as well as the service process. A *queueing system* can be defined as any system of flow wherein entities arrive to receive service of some kind, are processed by the system, and subsequently depart the system (or receive additional service in another part of the system). The entities that demand service are often called *customers*, and the resources that provide service are called *servers*. It is important to note that the customers need not be human beings; however, this term is often used generically to describe the entities that flow through the system. For example, in a healthcare facility, the customers are typically patients, but they might instead be requests for data, forms that must be processed, or hospital beds that must be prepared for the next patient. The servers might (respectively) be the physicians who care for patients, a computer that processes the form data, and a member of the nursing staff assigned to turn over the bed. An analytical *queueing model* is a mathematical representation of a queueing system, and *queueing theory* is that branch of applied probability concerned with the development and analysis of queueing models that represent stochastic service systems (Chap. 2).

Due to the nature of healthcare operations, queueing analysis can be a powerful tool to assess system performance and identify areas for improvement in the delivery of quality healthcare. It is ideal for modeling many units in the healthcare facility as delays abound in nearly all areas of these systems (e.g., waiting to check in, waiting to be seen by a doctor, waiting to be transferred out of the ICU, waiting for resources). Analytical queueing models differ from discrete-event simulation models in that they require far less data and can be used to develop simple analytical expressions for important measures of healthcare facility performance (e.g., average delay experienced by patients, resource utilization levels, and patient throughput). However, these models also impose much stronger assumptions than simulation models, some of which might be difficult to justify in practice (e.g., time-stationary patient arrival processes and/or steady-state operating conditions). Despite these restrictions, queueing models are extremely useful for describing the salient features of a service system, determining appropriate staffing or resource levels, and providing analytical expressions that can be used to mathematically optimize aspects of system performance. They can also be used to determine optimal policies (e.g., selecting the optimal *queueing discipline* or order in which queued customers are served). The primary objective of this subsection is to survey recent work related to queueing models in healthcare systems with a particular emphasis on ICUs. For an excellent summary of the use of queueing models in general healthcare contexts, the reader is referred to Green (2006).

ICUs provide care to critically ill patients, possess a limited number of beds, accommodate a variety of patient types, and require specialized services and (costly) resources. Because the number of ICU beds is limited, patients are not always admitted on a first-come-first-served (FCFS) basis, but rather on priority rules that consider the severity of the patients' illnesses, current bed occupancy levels, and other administrative rules. A queueing model can be very effective as an ICU design and/or performance analysis tool because it provides insights into two key measures of ICU performance: (1) patient admission delays or diversions (rejections) and (2) bed or other resource utilization levels. ICU admission delays (or diversions) can have potentially devastating outcomes for critically ill patients; a delay of only a few hours can lead to irreversible health degradation or even loss of life. Furthermore, admission delays for surgery patients in the ICU can result in downstream costs for the healthcare facility. Bed utilization has typically been the primary performance measure for bed capacity planning in ICUs (and other hospital wards) (cf. Green 2002), and bed occupancy rates are strongly correlated with the schedules of ICU clinical personnel, which comprise a significant portion of overall ICU expenditures. Both of these measures are prevalent in the models we review here.

Cooper and Corcoran (1974) sought to determine the number of beds required in a cardiac care unit of a major urban hospital to achieve a 5% patient rejection rate. The unit consists of two stations: the coronary care unit and the intermediate care unit. Two types of patients arrive to the coronary care unit, each with its own length of stay and arrival rate. Once the patient is discharged from the first unit, he/she is transferred to the intermediate care unit where less intensive care is provided. Patients are also assumed to die with some probability in each unit. The problem is considered as a network of loss systems (i.e., a queueing system with no waiting room) with exponentially distributed patient interarrival and service times. The rejection probabilities were computed exactly by solving a system of linear equations. The number of beds needed to achieve a 5% rejection rate in each of the units was computed for different patient arrival rates. However, the model was not validated with empirical data. Instead, the authors emphasize the potential benefits of using operations research techniques for decision making in healthcare environments.

Kim et al. (1999) analyze the admission and discharge processes of a particular ICU in a public Hong Kong hospital by first using a computer simulation model based on real data from the ICU. The authors answer the following questions: (1) How often will a full-capacity situation arise? (2) How long, on average, will patients have to wait for admission? and (3) To what extent might any untoward delays be relieved by additional capacity (more beds)? There are four different patient types: (1) ward patients, (2) accident and emergency patients, (3) operation theater-emergency patients, and (4) operation theater-elective patients. The ICU's admission decision takes two factors into consideration: the patient's "attributes" and the "state" of the ICU. If admitted, the patient joins a common queue and waits for service; when their service is completed, they release the system. Arrival rates, admission fractions, and average LOS for each patient type were estimated using actual data. Additionally, Kim et al. (1999) use an $M/M/c$ queueing model to compare the results of the simulation model. The most important part of the study compares the effects of different arrival streams for *elective* ICU patients. Specifically, simulation results revealed that the average number of queued patients, the maximum number of queued patients, and the average patient queueing time were all reduced by approximately 67% using a uniform arrival stream as opposed to a Poisson arrival process. Moreover, the average number of ICU diversions (rejections) dropped from 49 to 5.15, while the bed utilization rate was 69%, indicating the sufficiency of the current bed capacity. These observations present an important managerial insight for ICU units, namely, that a minor adjustment in the admissions policy can alleviate congestion, thereby obviating the need for additional bed capacity. Although the simulation model was validated with real data, a statistical analysis of these data was not provided.

Though not an ICU model, Gorunescu et al. (2002) present a general hospital bed capacity management tool. They present a technique for optimizing the number of beds needed to ensure an acceptable loss probability and a way to optimize the average daily cost, by balancing the trade-off between empty beds and denied admission rates. The performance measure is the long-run average cost per unit time. In their framework, patients arrive according to a Poisson process, and their LOS distribution is approximated by a phase-type (PH) distribution as the length of stay is highly variable. The parameters of the phase-type distribution are obtain via likelihood estimation techniques using real data from a large hospital. The proposed queueing model is the $M/PH/c/c$ loss model where $c$ represents the number of available hospital beds and the "PH" denotes the fact that the service time (LOS) follows a PH-distribution. Using this model, they compute the loss probability (fraction of arrivals that are denied admission), mean number of occupied beds, mean LOS, and bed utilization by using standard formulae. Additionally, they consider two optimization models. The objective of the first is to minimize the number of required beds, subject to a loss probability constraint. This problem can be solved using the inverse of the Erlang loss formula (cf. Ross 1996) with respect to $c$ (the number of beds). The second model seeks to minimize the average cost per unit time or maximize the average reward per unit time. The problem is cast as an inventory problem wherein a "holding cost" $h$ ($h > 0$) is incurred per day for each empty bed, and a fixed penalty cost $\pi$ ($\pi > 0$) is incurred for each patient that is turned away. It was shown that the optimal number of beds is sensitive to the cost ratio, $\pi/h$. The model was subsequently applied to the geriatric department of a London hospital, and it revealed that 150 beds were needed to achieve a 5% rejection rate. In fact, for cost ratios $10, 20, 30$, and 40, the optimal number of beds ranged from 140 to 160. The authors concluded that the $M/PH/c/c$ model is appropriate based on bed occupancy data collected in the geriatric department in 2000. One advantage of this model is that the policy depends only on relative costs as opposed to absolute costs, which are often very difficult to estimate.

Average bed occupancy rates have been the standard measure behind bed capacity decision making and planning. Green (2002) examines an alternative performance measure by considering the availability of beds upon demand, that is, the probability that an arriving patient finds an empty bed that grants him/her admission. Bed availability depends on a number of factors, for example, the hospital unit type, patient type, requested service, and hospital admission policies. In the context of a queueing model, bed availability can be viewed as the probability that arriving patient experiences zero delay. Using the classical Erlang delay model (namely, the $M/M/c$ model), the zero-delay probability can be expressed as a simple function of only the unit size (the number of beds $c$) and the bed utilization. This well-known result is used to determine the number of beds needed to meet various zero-delay probability targets and bed occupancy rates. Data from a New York hospital revealed that average bed occupancy levels for various obstetric and ICU units were below the commonly used standards. However, the queueing model revealed a different picture; only a small portion of those units have the necessary bed capacity to meet certain delay standards. For example, for a bed occupancy rate of 75% and a delay standard of 10%, only 40% of the state hospital obstetric units had sufficient capacity to achieve those standards. Similarly, for an occupancy rate of 85% and a delay standard of 10%, approximately 58% of the ICU units had insufficient capacity to meet the standards. This study illustrates that the average occupancy rate may be a misleading measure of the quality of patient care. Even if utilization rates meet commonly accepted standards, they can result in much higher rejection rates, especially for smaller units.

The stochastic nature of the arrival and service processes at an ICU makes patient scheduling a difficult task. Shmueli et al. (2003) focus on admission criteria that maximize a different measure of performance, namely, the expected incremental number of lives saved annually by operating the ICU. They define the *incremental surviving benefit* as the increase in probability of survival that a patient gains by being treated in the ICU as opposed to a regular hospital ward. It is assumed that upon arrival to the ICU, a patient's incremental survival benefit is drawn from some general distribution. Subsequently, they use the classical Erlang loss system ($M/M/c/c$) to obtain tractable steady-state probabilities and derive formulae for the ICU expected survival rate under three different admission policies: (1) a first-come-first-served policy (FCFS), (2) an FCFS threshold policy where patients are admitted if their incremental benefit is greater than a specified threshold, and (3) a FCFS threshold policy that may also depend on the current number of free beds. The incremental survival benefit distribution is estimated using real clinical data from the ICU of the Hebrew University-Hadassah Medical Center. It was concluded that an optimized FCFS threshold policy can result in a significant increase (18 lives or 17.9% improvement) in the incremental number of lives saved as compared to a FCFS policy. A FCFS threshold policy that depends on the current number of free beds offers only marginal improvement (1.4 lives or 1.2% improvement). However, these policies have some ethical implications as they do not promote parity. Nonetheless, the authors appeal to the notion that patients who are likely to recover should take priority over patients with a very poor recovery prognosis.

McManus et al. (2004) demonstrated that queueing models can capture the behavior of ICUs better than simple average calculations. They utilized two-year admission and LOS data from a busy medical and surgical ICU to estimate the parameters of an Erlang loss model ($M/M/c/c$). They calculated the rejection rates by varying the bed capacity and the patient arrival rates. It was demonstrated that, for a large unit operating at or near capacity, the model is able to accurately predict the ICU rejection rates. The correlation coefficient between predicted and observed rejection rates was 0.89. The model was also able to capture the exponential increase in rejection rates when utilization rates exceed 80–85%, which are typically observed in practice. They concluded that the stochastic nature of patient flow may lead capacity planners to underestimate the resources needed at busy ICUs, but that the Erlang loss model is able to accurately forecast bed capacity needs in high-utilization environments.

Griffiths et al. (2006) created a queueing model to represent the high variability in the LOS distribution and used the model as a capacity management tool. Considered are two primary performance measures: the mean number of patients in the ICU and the mean number of patients queued and waiting to enter. Using real data from a large teaching hospital, it was determined that the arrival process is Poisson but that the LOS distribution is highly skewed with a standard deviation that is substantially greater than its mean; therefore, the LOS distribution is approximated by a hyper-exponential distribution (a special type of PH-distribution whose probability density function is a linear combination of two exponential densities). Using the $M/H/c/\infty$ queueing model, the steady-state equations were derived and solved to obtain the primary queueing performance measures. Obtained were the steady-state probability distribution of the number of patients in the ICU, the mean number of patients in the ICU, and mean number of patients waiting to enter the ICU, each of which were validated using real data. For example, the steady-state mean number of patients in the ICU was computed to be 12.96 by the queueing model, whereas the empirical value was found to be 12.91 patients. The model was subsequently used to examine the impact of variation in the LOS and patient arrival rates. The model suggests that 17–18 beds are needed to ensure a delay probability of approximately 0.05. A one day decrease in the LOS decreases the average number of customers in the system from 12.91 to 9.87, and 22–23 beds are required when the arrival rate increases by 20%.

de Bruin et al. (2007) investigated the dynamics underlying the proportion of refused admissions (or blocking probability) $P_c$ and bed occupancy rates $\rho$, under different bed configurations in an emergency care chain that consists of units in tandem. Furthermore, the authors attempted to quantify the sensitivity of these measures to varying arrival and LOS rates. The paper was inspired by the emergency inpatient flow of cardiac patients in a university center that consists of three emergency units in tandem: A first cardiac aid (FCA), a coronary care unit (CCU), and a normal care (NC) unit. The FCA unit is intended to provide some rapid diagnosis, and upon completion of treatment, the patients exit or are forwarded to the CCU or the NC unit. The CCU patients either exit or proceed to the NC unit. From the NC unit, they can either return to the CCU or be discharged. The patients are divided into two distinct flows: those that enter the system at the FCA and those that enter at the CCU. Using computerized hospital records, arrival and LOS data were obtained and used to investigate the relationship between blocking probabilities and bed occupancy rates for a simple $M/M/\infty$ queue. The data revealed that 20–30% of the total LOS is due to blocking; therefore, solving the blocking problem is critical to improving patient flow. Then they focused their attention on a system that consists of a CCU and NC unit and explored the blocking probability at the CCU for various bed configurations. To that end, they formulated a two-dimensional Markov chain where the first component of the state space is the number of occupied beds at the CCU and the second component is the number of occupied beds at the NC unit. Using arrival and LOS data, they numerically obtained the blocking probabilities assuming various bed configurations. The infinite-server queueing model highlights the fact that the LOS distribution can be highly variable and possess a heavy tail as it is often the case that a small number of patients consume a disproportionate amount of resources. The analysis showed that rejections at the FCA are primarily due to unavailability of downstream beds. The study also indicated that larger units can attain higher occupancy rates while maintaining the same rejection rates (due to economies of scale), and target occupancy rate of 85% is only attainable by hospital units with more than 50 beds.

Litvak et al. (2008) proposed a cooperative solution for the ICU capacity problem. Several hospitals in a region jointly reserve a small number of beds for regional emergency patients. The main objective is to increase the patient acceptance rates. An overflow model, similar to queueing models for circuit-switched telephone systems with overflow capacity, was created. There are three types of patients: (1) elective patients, (2) internal trauma patients, and (3) regional trauma patients. Patient arrivals were assumed to follow a Poisson process, and the patient LOS distribution was assumed to be exponential. To approximate the blocking probabilities, the Equivalent Random Method (ERM) was utilized; however, standard ERM methods cannot be applied in this model because internal emergency patients cannot be rejected and elective patients are never sent to the overflow. Computed were the mean and variance of the overflow via the balance equations and corresponding probability generating functions. Additionally, a simulation model was built using a commercial simulator, eM-Plant, that allows for detailed acceptance rules and closely mimics actual patient flow in the ICUs, including general LOS distributions. Similar results were obtained using the ERM, so it was concluded that the ERM adequately captures the loss probability and the required number of regional beds. The model was then employed to show that cooperation between hospitals helps to achieve a high acceptance rate with a smaller number of beds. For instance, to achieve a 1% rejection rate, 11 beds are needed with cooperation as compared to 16 beds without cooperation.

Though not explicitly an ICU analysis, Cochran and Roche (2009) developed an implementable queueing network model for an emergency department (ED) with multiple areas and utilized the model as a capacity management tool. Different ED areas were sized using the average waiting time ($W_q$) and overflow probability ($P_c$) as quality-of-service targets. Nonhomogeneous arrival patterns, non-exponential service time distributions, and multiple patient types were all incorporated. The purpose of this work is to reduce abandonment from the ED stemming from excessive wait times. They use a strategy called Split Patient Flow (SPF) which aims to treat lower acuity patients in a separate queue from higher acuity patients. Five levels of patient acuity were defined based on LOS. Certain rules of the flow pattern of these patients were given, and routing matrices were calculated to estimate the patient flow to each ED area. The arrival rates were adjusted for seasonality effects. Using the Allen–Cunneen $M/G/c$ approximation, they approximate average waiting time. To estimate the overflow probability, they employed the Erlang loss model ($M/G/c/c$), where the number of servers ($c$) for each node is varied until the performance targets (including utilization) are attained. Blocking between ED areas was not considered. The model was used to redesign ED units of Banner Health hospitals, and the results of a Phoenix hospital were provided. The capacity requirements were computed for given quality-of-service targets (5% overflow probability and 75% utilization). The implementation of SPF yielded a 61% reduction in the number of patients who leave without treatment. A spreadsheet implementation of the model is currently employed by hospital decision makers.

ICUs and surgical (or operating) theaters (OT) are critically linked in a healthcare facility. Patients who complete surgery in the OT are often transferred to the ICU for post-surgery care. However, patients may be rejected upon arrival at the OT (or the ICU) due to limited ICU capacity. The probability that an arriving patient at the OT or ICU is admitted serves as an important performance parameter for the hospital. Rejection of such patients can have serious consequences on patient health and survival, hospital costs, and canceled operations. In van Dijk and Kortbeek (2009), van Dijk and Kortbeek examined a tandem queueing network model of the OT-ICU system wherein arriving patients are first admitted to the OT for surgery and then subsequently transferred to the ICU, or they are sent to the ICU directly. However, if the ICU is full, arriving patients are rejected (diverted), whereas completed OT patients remain in the OT waiting for an ICU bed. For this system, no exact analytical solution is available for the joint steady-state probability distribution of the bed occupancy levels in both units. However, assuming the ICU can be modeled as an Erlang loss system with $c$ beds, the authors prove that the overall rejection probability of the OT-ICU system can be bounded by the loss probabilities of the $M/G/c/c$ and $M/G/c-1/c-1$ queueing models, that is, the former provides a lower bound while the latter provides an upper bound. These bounds were validated using simulation experiments and can be especially useful for bed capacity planning at both units. A case study was provided in which the capacity requirements are computed to guarantee certain rejection rates; however, the results were not validated with empirical data.

de Bruin et al. (2010) analyzed capacity decisions of different hospitals units, including ICUs, using Erlang loss models. There are two types of patients: scheduled and unscheduled. Statistical tests showed that the arrival process of unscheduled patients can be modeled as a Poisson process; however, the arrival process of the scheduled patients is not Poisson. They imposed the Poisson assumption on the latter nonetheless for practical considerations. Using a detailed statistical analysis of LOS, the authors argued that decisions based on fixed LOS are misleading. An Erlang loss model ($M/G/c/c$) was used to compute the blocking probabilities and occupancy rates, and the model was validated using hospital data collected during 2004–2006. They developed a decision support tool that is easily implemented in a commercial spreadsheet program. The authors concluded that, if economies of scale are properly applied, by merging departments (via bed pooling) or mixing patient flows, both an acceptable service level (in terms of refused admissions) and an economically viable occupancy rate can be realized. That is, to achieve the same proportion of refused admissions (5%), 22 beds are needed in a merged unit as compared to 29 beds in total.

A common assumption in bed capacity planning models is that the patient arrivals are time stationary (i.e., that the mean rate of patient arrivals does not change over time). However, it is well known that this is not the case for most clinical wards. It is estimated that two-thirds of admitted ICU patients arrive during office hours, whereas most of the scheduled admissions occur during weekdays instead of the weekend. de Bruin and Bekker (2010) analyzed the impact of a piecewise constant patient arrival rate on the daily offered workload and the fraction of refused admissions for clinical wards using the $M_t/H/c/c$ queueing model. The Poisson arrival rate is assumed to be time varying, while service times follow a hyper-exponential distribution. Approximations for this system were obtained using the infinite-server $M/G/\infty$ model for which the queue workload has a tractable form. Since it was observed that the arrival pattern may vary on different time scales, the authors considered daily and weekly variations of the patient arrival rate. For daily variation, two different arrival rates and an exponential LOS distribution were assumed. Using the aforementioned approximation, they compared different scenarios by varying the arrival and LOS rates over a fixed traffic load. They reported that the impact of the arrival variation on the offered workload and patient loss fraction diminishes as the LOS increases. Thus, in clinical wards where the LOS is a multiple of the arrival cycle length, the fluctuations have a limited impact on the offered workload and patient loss fraction. In the case of arrival rates with a weekly variation pattern (i.e., when the arrival rate cyclic period is in range similar to that of the average LOS), the weekly load and loss fraction may fluctuate significantly over their average values. These values may be relatively small or undesirably large during certain days of the week, indicating that a varying arrival rate may have a profound effect on unit operations. Finally, the authors examined the effect of different LOS distributions. In the context of a weekly arrival pattern, they showed that the fluctuations in load and blocking probability were alleviated as the variance of LOS increases. They concluded that a more variable LOS has a stabilizing effect in the load and blocking probability.

It is evident that analytical queueing models can be very useful for modeling, analyzing, and managing ICU operations. This review of queueing ICU models reveals the prevalence of the Erlang loss and delay systems for modeling ICUs and their surrounding departments. Many of these demonstrate that modeling assumptions thought to be very restrictive are often valid in practice (e.g., unscheduled patient arrivals do, in many cases, follow a Poisson process). Moreover, it is evident that queueing approximations are particularly useful when considering the ICU in relation to other wards of the hospital. The final section of this chapter describes an ongoing ICU project currently underway involving both engineering and medical researchers at the University of Pittsburgh.

## 5 Case Study: Veterans Affairs Hospital

In this section, we briefly discuss recent efforts undertaken by the authors (as part of a larger, multidisciplinary research team) to closely examine issues related to the modeling, analysis, and management of ICUs. This collaborative project centers on ICU operations within the Veterans Affairs Pittsburgh Healthcare System (VAPHS) and is supported by the Veterans Engineering Resource Center (VERC) which was established at the VAPHS in 2009. The VAPHS is a major healthcare provider in the Pittsburgh area, serving the diverse medical needs of military veterans. The research team includes industrial engineering faculty, medical researchers, physicians, postdoctoral associates, and graduate students at the University of Pittsburgh and the University of Pittsburgh Medical Center (UPMC). Oversight was provided by management and industrial engineering staff members at the VAPHS.

Through consultation with a medical advisory board, patient congestion was identified as a major problem in the ICUs. As is the case for most hospitals, ICU patients in the VAPHS usually experience multiple unit transfers before or after their ICU stay. Patients are admitted to the ICU from the emergency department, an operating room, or some other inpatient unit. After being discharged from the ICU, they can be transferred to other monitored or non-monitored units; however, these transfers do not necessarily occur at the time they are requested (either by a nurse or physician). If the receiving (or assigned) unit is at capacity, or the necessary clinical staff are not available, patients experience a delay in their transfer. As in Sect. 2, we refer to this phenomenon as *patient blocking* and the associated time spent waiting to be transferred as *blocking time* or *delay*. Blocked patients (in the ICUs) unnecessarily occupy expensive resources that are no longer clinically needed; consequently, operating costs increase dramatically without improvement in healthcare outcomes. In Sect. 2, we noted that this congestion between the ICU and its interacting downstream units also serves to delay critical care for incoming ICU patients or to disrupt operations at other inpatient units (e.g., an emergency department patient may need to be diverted to another hospital).

Our review of existing ICU simulation and analytical models (in Sect. 4) revealed that very few models explicitly address the former aspect of blocking (namely, patient blocking at the ICU). In fact, the vast majority of the analytical models focused almost exclusively on the long-run fraction of patients who are denied admission to the ICU due to bed unavailability (with the exception of van Dijk and Kortbeek 2009; Dobson et al. 2010). Only a few simulation models explicitly considered patient blocking. Cochran and Bharti (2006) assumed that patients assigned to a downstream unit that is at capacity remain in their current bed until a bed becomes available. They considered alternative resource allocation schemes to satisfy certain blocking criteria. We contend that the literature related to patient blocking is sparse due to a dearth of data needed to identify (and quantify) patient blocking. Most hospital data systems (including those used by the VAPHS) record *actual* patient movements (namely, origin, destination, and time of transfer) as opposed to the time at which the patient *could have been* moved. Consequently, an analysis of patient LOS data that excludes blocking results in estimates of the *actual LOS* rather than the *medically–indicated LOS* of a patient. The problem is exacerbated when patients experience blocking during multiple unit transfers as the medically–indicated LOS spans the time from the last transfer, while the patient was at another unit, to the time of current transfer request.

**Study Objectives.** The primary objective of the VAPHS ICU study was to develop a simulation model that is able to (1) capture the interdependencies between ICUs and other inpatient units, (2) replicate patient blocking and medically–indicated LOS that are currently prevalent in the VAPHS, and (3) devise mathematical procedures to optimally manage the ICUs based on real-time evolution of patient physiology. The remainder of this section focuses on objectives (1) and (2) only. The starting point of our analysis was the VAPHS databases from which we extracted detailed patient movement information including blocking times and medically–indicated LOS. In the sequel, we describe a large-scale simulation model that replicates these key patient flow characteristics and accurately models patient blocking times and bed occupancy rates. The model was also used to examine (1) the short-term effects of sudden bed changes, (2) the long-term effects of bed capacity changes, and (3) alternative responses to a spike in the patient arrival rate due to an exogenous event (e.g., diversions from another hospital due to catastrophic events).

In light of the different ICU modeling alternatives discussed in Sect. 4, we chose to create a discrete-event simulation model for two important reasons. First, the VAPHS was interested in analyzing not only the probability that a patient experiences blocking but also the duration and potential cost impact of that blocking. Little to no work has been done in the queueing community to characterize blocking durations, so a simulation model emerged as the natural choice. Second, the VAPHS requested a high-fidelity model capable of capturing ICU interactions with other units in the hospital; therefore, the complexity of the model precluded the use of an analytical queueing model. Finally, the simulation model is designed to (ultimately) accommodate the evolution of the patients' physiological data to facilitate dynamic ICU decision making based on each patient's current health status.

**Description of the Facility.** The VAPHS facility is comprised of several ICUs as well as peripheral inpatient units. The units considered in our simulation model are as follows:

1. The medical ICU (MICU) which contains 9 beds.
2. The surgical ICU (SICU) which contains 12 beds.
3. The coronary critical care unit (CCU) which contains 18 beds.
4. A step-down unit (SDU) which contains 9 beds.
5. A monitored medical unit containing 15 beds.
6. A monitored surgical unit containing 12 beds.

These various units offer a wide range of monitored care. In the VAPHS, there also exist a number of non-monitored care units which were not modeled explicitly. Patients are admitted to the six units from a number of physical locations including the emergency department, operation recovery rooms, various clinical labs, or directly from home. After examining patient data files and consulting with the clinical advisory board, we compiled a detailed list of physical locations from which patients apply for admission to the individual units. Subsequently, we aggregated those locations into distinct categories we refer to as *patient sources*. Therefore, each admitted patient was mapped to exactly one of the patient sources labeled as: non-monitored medical and surgical inpatient units, emergency department (ED), operating room (OR), post-anesthesia room (PAR), other floors and home.

**Resources used for the study.** To facilitate the study, we extensively utilized (1) VAPHS clinical staff members who provided invaluable feedback regarding clinical practice and policies within the ICUs and otherwise and (2) VAPHS patient databases that contain explicit patient movement information. The clinical staff provided a description of the clinical practices and policies that dictate patient movement under different scenarios. For example, what happens to an emergency patient who is assigned to a full SICU? Or what happens to an ED patient who is assigned to a full medical monitored unit? With the help of the VAPHS clinical staff, we compiled a list of documented (and undocumented) clinical practices and policies that guide the rule base embedded within the discrete-event simulation model.

The other critical resources were two patient databases stored in the VAPHS data warehouse. The first database is a collection of entries corresponding to *patient transfer requests*. A transfer request is logged into the system by a nurse when a patient is medically cleared to transfer to another unit. Every patient entry contains the patient's current unit, the next assigned unit, the time at which the request was made, and the time at which the request was fulfilled. By examining the entries of each patient, we accurately identified a few critical pieces of information, namely, (1) patient arrival sources and (2) time-dependent patient arrival rates (based on time of day and the day of week). Using these data, we estimated the patient blocking times experienced at each unit by calculating the time difference between a patient transfer and the corresponding transfer request. This is precisely the blocking time experienced by the patient waiting to be transferred. The second database is a collection of entries corresponding to the patients' realized transitions. Each entry corresponds to a patient transfer and contains the relevant unit and the time this transition occurred. A transition can be an admission, a transfer between units, or a discharge from a unit or the hospital. From this information, we estimated the corresponding LOS for a patient's visited units by calculating the time difference between consecutive entries. Moreover, we calculated patient transition probabilities and throughput for every unit. By combining the two data files, a complete picture of the patient arrival processes to each unit, the total and medically indicated LOS, the discharge rates, and the actual blocking times emerged.

While these resources proved to be invaluable for the simulation study, due to the complexity of ICU operations, it was not possible to account for all of the possible scenarios that might be encountered. For example, certain extraneous rules often dictate patient movement, and these rules (which are neither documented nor easily described) usually stem from years of empirical experience. The clinical advisory board cited cases in which a transfer from the SICU to a SDU might be affected by the male-to-female ratio in that unit. In other instances, a patient transfer from the MICU to a non-monitored bed might be prohibited due to the patient's special need for the presence of clinical staff member.

Another complication is that the patient movement data can provide a biased view of patient transfers. For example, not all patient blocking instances result from downstream (or upstream) capacity issues. The advisory board noted that blocking delays less than three hours in duration are typically attributed to the patient's need for reevaluation by a physician and not a shortfall in capacity. Moreover, blocking at units such as the ED is often biased since a bed is requested almost immediately when a patient enters the ED. Therefore, a series of adjustments were made in the data analysis phase in order to compensate for these biasing effects. As is the case for any large-scale simulation project, some necessary simplifications and assumptions were needed, and these assumptions have to be considered when examining the model's output.

**Simulation Model Description.** The simulation model was coded in the software package OMNeT++ which is an event-driven C++ simulation library designed explicitly for building network simulators and used extensively by telecommunications engineers and researchers. This platform adopts an extensible and modular approach for building simulation models that makes it easy to incorporate complex experimental scenarios and configurations (e.g., the addition or removal of an entire clinical unit). The simulation model consists of several modules representing the patient sources and inpatient units. Patient entities originate at patient sources with varying intensities depending on the day of the week and the time of day. This is a necessary step to capture the differing arrival patterns for every patient source. For example, ED patients may arrive during early morning hours or during weekends, whereas surgery patients arriving from PAR may only arrive at particular time frames during the weekdays.

In the majority of existing ICU models, patient LOS data is used to fit a LOS distribution that generates realizations and transition probabilities during simulation runs. In order to improve model validity, some authors propose stratifying patients into different categories depending on the treatment specialty and patient attributes (e.g., age or some selected clinical indicators). This approach has yielded positive outcomes for studies involving single ICU simulation models where patients receive treatment and are discharged. However, in a large-scale simulation model, such an approach might not provide the fidelity needed to capture patient flow dynamics as patient entities occupy beds and transition from one unit to another. Patient transitions may be stochastically dependent on a number of factors including patient origin, the sequence of prior visited units, and the corresponding medically indicated LOS. By contrast, our simulation model exploited a large database containing *actual* patient instances as observed in the data files. Each patient instance was described by an arrival source, a sequence of visited units, and a corresponding sequence of medical LOS.

For every hour of the day and day of the week, the program retrieves the corresponding arrival rate for each patient source and generates the corresponding number of patients using a Poisson distribution. Subsequently, the model queries the database to randomly retrieve the number of patients for each source. The retrieved patients are inserted into the simulation model, and they transition through the different units based on their predetermined sequence of units and medical LOS. When patients reach the end of their medical LOS, they request a bed at the next assigned unit. If the unit is full, the model searches for a bed in other units based on predetermined rules that correspond to actual clinical practices. If a bed is not available in another unit, the patient is blocked until a bed becomes available. The simulation model tallies and reports patient blocking times and bed occupancy rates for all of the inpatient units of interest.

**Validation and Results.** The model validation process consisted of two steps: a *conceptual* validation that included demonstration of the model animation to the VAPHS clinical staff and an *operational* validation focusing on the model's quantitative outputs (see Bountourelis et al. 2011). The aim of the conceptual validation was to establish face validity of the model and its underlying logic. The high-fidelity simulation animation mimics (1) patient movements as they are admitted, transferred, and blocked in various monitored units, and (2) the visual display of various performance measures (e.g., the number of occupied beds and blocked patients) to ensure the observed variational patterns was consistent with the experience of the clinical staff. For the *operational* validation, we compared the simulation output statistics corresponding to patient blocking times (delay) and the number of occupied beds with the corresponding statistics retrieved from patient data files. Preliminary simulation runs indicated that the simulated blocking times and bed occupancy rates were consistent with the real data. However, a set of input parameters plays a pivotal role in the simulation model but cannot be quantified from the available data. These parameters needed to be calibrated to further improve the statistical proximity of the simulation output to the observed data.

We distinguished two types of input parameters—those pertaining to the construction of the simulation model and those that control patient movement. Parameters of the first type include, for example, the current number of beds in each unit. We varied the number of beds to account for *bed closures*, that is, the phenomenon of an empty bed not being used for reasons pertaining to staff availability and scheduling. Since accurate bed closure data were not available, the removal of bed capacity is an alternative way to approach the average bed availability. Parameters of the second type included, for example, the probability of a surgery patient being transferred upstream, that is, to a higher level of care due to a bed shortage at the destination unit. We introduced this parameter to replicate well-known clinical practices for which accurate estimates of the frequency of occurrence are not available.

Model calibration can be a time-consuming process due to the many dimensions of the parameter space induced by the total number and range of values of the input parameters. Furthermore, there is no standardized calibration procedure in the literature. After experimenting with various sets of parameter values, we identified a set of values under which the model output closely matched the output observed from the data. We focused on those units since these are the clinical areas were critical care patient blocking occurs the most. Figure 6.1 depicts the observed patient blocking in ED and SICU compared to the corresponding simulation output. For the ED and SICU, the quantile–quantile (Q–Q) plot shows that the simulated distribution of patient blocking times closely matches the historical distribution of patient blocking times. Given the inherent complexity of the ICU operations and the necessary simplifying assumptions, we concluded that the model performs satisfactorily in replicating the ED and critical care patient blocking times.

**Application of the Model.** Subsequently, our model was used to answer questions of interest to the VAPHS managerial and clinical staff. An important set of questions concerned the propagating effects of sudden bed removals from a particular unit. The clinical staff sometimes experience bed closures due to exogenous factors, such as equipment failures or building infrastructure problems. The simulation model allowed us to quantify the effects of sudden bed removal scenarios and the best ways to respond to such events. Additionally, the VAPHS were interested in capacity planning for all of the units. Therefore, we performed an analysis of the long-term effects of a large-scale bed reconfiguration on the delay time experienced by ED patients. The model can also potentially be used to evaluate so-called “hub-and-spoke” policies for different patient types depending on their arrival source and the unit of first admission. Alternatively, we may examine the impact of a “bursty” patient arrival process stemming from an epidemic in the community and the optimal way to handle the increased load by allocating additional resources. Ultimately, our aim is to use the model to help guide ICU admission and discharge decisions by explicitly taking into account a physiological indicator of the patient’s health status (e.g., the SOFA score) as it evolves temporally.

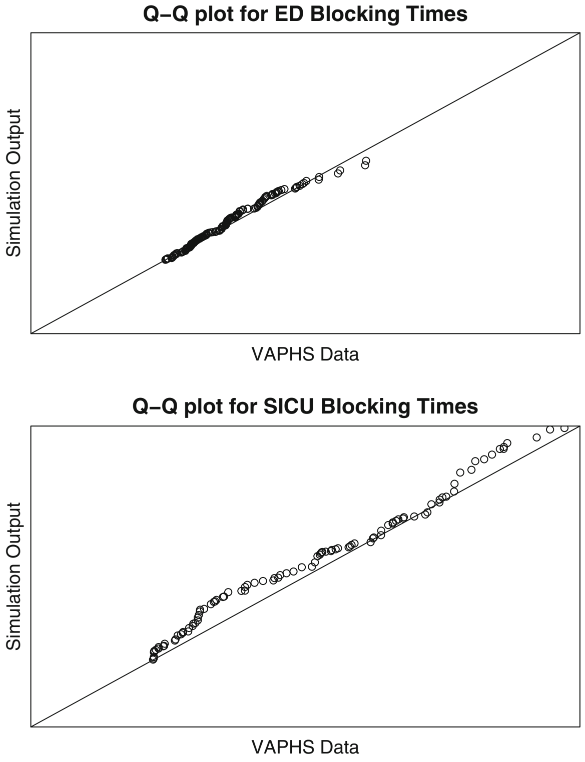


**Fig. 6.1** Q–Q plots of patient blocking times: simulated versus historical data

# 6 Conclusions and Open Research Challenges

The primary objective of this chapter was to highlight the unique challenges associated with designing, modeling, and managing ICUs. The ICU is an extremely important part of any hospital, and its interactions with other hospital units are indispensable. Due to the complex nature of the ICU, many operations researchers have employed discrete-event simulation and/or analytical queueing models to assess important performance measures, such as the patient rejection rate and length of stay in the ICU. However, many challenges and opportunities remain for the operations research community, and we highlight a few of those here.

Section 4 described differences between discrete-event simulation and analytical queueing models of ICUs. Simulation models allow for greater model fidelity, but can be difficult to validate and require a great deal of data, some of which may not be available. Queueing models require far less data and are (generally) very easy to implement, but the most accessible models usually impose strict modeling assumptions (e.g., stationary arrival processes). Networks of finite-capacity queueing stations (particularly loss systems) are not easy to analyze, and the modeler usually is forced to resort to approximation schemes.

Based on our review of existing models and our own experiences in modeling ICUs, we offer the following recommendations. First, a discrete-event simulation model should be employed if there is a strong desire to include detailed dynamics of patient movements, medically indicated LOS, and interactions between the ICU and the other units in the hospital. A simulation model allows for greater modeling flexibility, even though a number of parameters need to be best estimated (and later calibrated) for validation. However, if the modeler is primarily concerned with assessing the likelihood of blocking, or designing ICUs, an analytical queueing model (namely, the Erlang loss model Kharoufeh 2011) provides a very simple means by which to assess the number of beds needed to achieve a quality-of-service guarantee. The formulae for doing so can be easily implemented in a spreadsheet program or coded in any common programming language (e.g., C, C++, Visual Basic). These models can also assess the impact of increased patient arrival rates and/or the average LOS on the patient rejection rate and bed occupancy rates.

There are a few interesting areas of open research related to the design, modeling, and management of ICUs. First, in Sect. 5, we noted that values were assigned to a number of input parameters that clearly influenced patient movement but could not be estimated from data or clinical experience. To calibrate these parameters, a multi-dimensional state-space search was performed *empirically* using a time-consuming trial-and-error approach. Because each ICU is unique, one important research direction is the development of a standardized methodology particularly tailored to the calibration of large-scale ICU models. Such a methodology might include both empirical rules and calibration procedures based on state space search heuristics. Additionally, when using simulation models for capacity planning and/or ICU design or redesign, the computational burden can be cumbersome due the large number of design alternatives. Specifically, in a large-scale facility with many different units and levels of care, the task of optimal resource allocation usually requires the evaluation of an exponentially large set of alternative configurations with respect to the number of resources being allocated. Doing an exhaustive search of feasible configurations is very time consuming, even if parallel simulation techniques are used on computer clusters. Alternatively, one might consider customization of *metamodeling* techniques, or methods for *approximating* the input/output (I/O) relationships revealed by the simulation model. Metamodels are fit using I/O data by sampling only a subset of the possible configurations and may be used to predict the expected simulation output for configurations that have not been simulated. They can also be used for sensitivity analysis, what-if scenarios, and optimization of the simulated system (Ankenman and Staum 2010; Kleijnen 2007). One such promising direction is the problem of optimal resource allocation through the use of *Kriging* methods (Kleijnen 2009; Kleijnen and Sargent 2000).

With regard to analytical queueing models of ICUs, the most natural approach is a network of finite-capacity queueing stations arranged in tandem. This configuration serves to mimic the movement of patients from the ICU to step-down beds to hospital beds and then to discharge from the hospital (assuming recovery). However, networks of finite-capacity queueing systems are notoriously difficult to analyze; therefore, many researchers have resorted to useful bounds and/or approximations for these types of systems (van Dijk and Kortbeek 2009; Osorio and Bierlaire 2009 are good examples of relevant work). Most queueing models focus on computing the patient rejection rate and bed occupancy rate, but it is very difficult to characterize the distribution—or even the mean—of the delay experienced by patients due to blocking. Another daunting challenge for the analytical modeler is the inclusion of patients who must reenter the ICU because their health status deteriorates after they have been released from intensive care. Obtaining analytical results for even the simplest model (i.e., one in which all arrival processes are Poisson and all medical LOS distributions are exponential) is nontrivial. Many opportunities exist for analytical approaches that are able to quantify blocking and its adverse effects on the system.

**Acknowledgements** We are grateful to the editor and an anonymous referee for constructive comments that improved the content and presentation of this chapter. We also acknowledge, with gratitude, support from the Veterans Engineering Resource Center at the Veterans Affairs Pittsburgh Healthcare System (VAPHS) in Pittsburgh, Pennsylvania. We especially thank Mr. Bob Monte, Mr. Matt Jenkins, and Mr. Scott Moore for providing invaluable information related to ICU operations at the VAPHS. Special thanks are also due to Drs. Gilles Clermont and Mark Roberts who provided guidance on clinical issues related to the simulation model.

## References

Angus D, Black N (2004) Improving care of the critically ill: institutional and health-care system approaches. Lancet 363:1314–1320

Ankenman BNB, Staum J (2010) Stochastic kriging for simulation metamodeling. Oper Res 58(2):371–382

de Bruin A, Bekker R (2010) Time-dependent analysis for refused admissions in clinical wards. Ann Oper Res 178:45–65

de Bruin AM, Bekker R, van Zanten L, Koole GM (2010) Dimensioning hospital wards using the Erlang loss model. Ann Oper Res 178:23–43

de Bruin A, van Rossum A, Visser M, Koole G (2007) Modeling the emergency cardiac in-patient flow: an application of queueing theory. Health Care Manag Sci 10:125–137

Bountourelis T, Luangkesorn L, Schaefer A, Maillart L, Nabors S, Clermont G (2011) Development and validation of a large scale ICU simulation model with blocking. In: Jain S, Creasey RR, Himmelspach J, White KP, Fu M (eds) Proceedings of the 2011 winter simulation conference. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey pp 1143–1153

Brailsford S, Harper P, Patel B, Pitt M (2009) An analysis of the academic literature on simulation and modelling in health care. J Simul 3:130–140

Carter M, Blake J (2005) Using simulation in an acute-care hospital: easier said than done. In: Brandeau M, Sainfort F, Pierskalla W (eds) Operations research and health care, vol 70. Springer, Berlin, pp 191–215

Cochran J, Bharti A (2006) Stochastic bed balancing of an obstetrics hospital. Health Care Manag Sci 9:31–45

Cochran J, Roche K (2009) A multi-class queuing network analysis methodology for improving hospital emergency department performance. Comput Oper Res 36:1497–1512

Cooper JK, Corcoran TM (1974) Estimating bed needs by means of queuing theory. New Eng J Med 291:404–405

Cooper L, Linde-Zwirble W (2004) Medicare intensive care unit use: analysis of incidence, cost, and payment. Crit Care Med 32:2247–2253

Costa A, Ridely S, Shahani A, Harper P, Senna VD, Nielsen M (2003) Mathematical modelling and simulation for planning critical care capacity. Anesthesia 58(4):320–327

Dasta J, McLaughlin T, Mody S et al (2005) Daily cost of an intensive care unit day: the contribution of mechanical ventilation. Crit Care Med 33:1266–1271

Davies H, Davies R (1995) Simulating health systems: modelling problems and software solutions. Eur J Oper Res 87(1):35–44

van Dijk NM, Kortbeek N (2009) Erlang loss bounds for OT-ICU systems. Queueing Syst: Theory Appl 63:253–280

KC DS, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. Manuf Serv Oper Manag 14(1):50–65

Dobson G, Lee H-H, Pinker E (2010) A model of ICU bumping. Oper Res 58(6):1564–1576

Eldabi T (2009) Implementation issues of modeling healthcare problems: misconceptions and lessons. In: Rossetti MD, Hill RR, Johansson B, Dunkin A, Ingalls RG (eds) Proceedings of the 2009 Winter Simulation Conference. Institute of Electrical and Electronics Engineers, Piscataway, pp 1831–1839

Fone D, Hollinghurst S, Temple M, Round A, Lester N, Weightman A, Roberts K, Coyle E, Bevan G, Palmer S (2003) Systematic review of the use and value of computer simulation modelling in population health and health care delivery. J Public Health Med 25(4):325–335

Forsberg H, Aronsson H, Keller C, Lindblad S (2011) Managing health care decisions and improvement through simulation modeling. Quality Manag Health Care 20(1):15–29

Friedman B, Steiner C (1999) Does managed care affect the supply and use of ICU services? Inquiry 36(1):68–77

Garland A (2005) Improving the ICU: Part 1. Chest 127(6):2151–2164

Gorunescu F, McClean S, Millard P (2002) A queueing model for bed-occupancy management and planning of hospitals. J Oper Res Soc 53:19–24

Green L (2006) Queueing analysis in healthcare. In: Hall R (ed) Patient flow: reducing delay in healthcare delivery. Springer, Berlin, pp 281–307

Green LV (2002) How many hospital beds? Inquiry 39(4):400–412

Griffiths J, Price-Lloyd N, Smithies M, Williams J (2006) A queueing model of activities in an intensive care unit. IMA J Manag Math 17:277–288

Halpern N, Pastores S, Greenstein R (2004) Critical care medicine in the united states, 1985–2000: an analysis of bed numbers, use, and costs. Crit Care Med 32:1254–1259

Harper P (2002) A framework for operational modelling of hospital resources. Health Care Manag Sci 5(3):165–173

Harper P, Pitt M (2004) On the challenges of healthcare modelling and a proposed project life cycle for successful implementation. J Oper Res Soc 55:657–661

Harper PR, Shahani AK (2002) Modelling for the planning and management of bed capacities in hospitals. J Oper Res Soc 53(1):11–18

Harrison D, Lertsithichai P, Brady A et al (2004) Winter excess mortality in intensive care in the UK: an analysis of outcome adjusted for patient case mix and unit workload. Intensive Care Med 30:1900–1907

Iapichino G, Gattinoni L, Radrizzani D, Simini B, Bertolini G, Ferla L, Mistraletti G, Porta F, Miranda DR (2004) Volume of activity and occupancy rate in intensive care units: association with mortality. Intensive Care Med 30:290–297

Jacobson S, Hall S, Swisher J (2006) Discrete-event simulation of health care systems. In: Hall R (ed) Patient flow: reducing delay in healthcare delivery. Springer, Berlin, pp 211–252

Jun J, Jacobson S, Swisher J (1999) Application of discrete-event simulation in health care clinics: a survey. J Oper Res Soc 50(2):109–123

Kahn J, Angus D (2006) Reducing the cost of critical care: new challenges, new solutions. Am J Respirat Crit Care Med 174

Kharoufeh J (2011) The $M/G/s/s$ queue. In: Cochran J, Cox A, Keskinocak P, Kharoufeh J, Smith J (eds) Wiley encyclopedia of operations research and management science. Wiley, New York

Kim SC, Horowitz I, Young KK, Buckley TA (1999) Analysis of capacity management of the intensive care unit in a hospital. Eur J Oper Res 115:36–46

Kleijnen JPC (2007) Experimental design for sensitivity analysis, optimization, and validation of simulation models. Handbook of simulation. Wiley, New York, pp 173–223

Kleijnen JP (2009) Kriging metamodeling in simulation: a review. Eur J Oper Res 192(3):707–716

Kleijnen JPC, Sargent, RG (2000) A methodology for fitting and validating metamodels in simulation. Eur J Oper Res 120(1):14–29

Knaus W, Draper E, Wagner DP, Zimmerman J (1985) APACHE II: a severity of disease classification system. Crit Care Med 13(10):818–829

Law A, Kelton W (2000) Simulation modeling and analysis, 3rd edn. McGraw-Hill, New York

Litvak N, van Rijsbergen M, Boucherie R, van Houdenhoven M (2008) Managing the overflow of intensive care patients. Eur J Oper Res 185:998–1010

Lowery JC (1992) Simulation of a hospital's surgical suite and critical care area. In: Swain JJ, Goldsman D, Crain RC, Wilson JR (eds) Proceedings of the 1992 winter simulation conference. Institute of Electrical and Electronics Engineers, Piscataway, pp 1071–1078

Lowery JC (1993) Multi-hospital validation of critical care simulation model. In: Evans GW, Mollaghasemi M, Russell EC, Biles WE (eds) Proceedings of the 1993 winter simulation conference. Institute of Electrical and Electronics Engineers, Piscataway, pp 1207–1215

Lowery JC (1996) Introduction to simulation in health care. In: Charnes JM, Morrice DJ, Brunner DT, Swain JJ (eds) Proceedings of the 1996 winter simulation conference. Institute of Electrical and Electronics Engineers, Piscataway, pp 78–84

Lowery JC, Hakes B, Keller L, Lilegdon W, Mabrouk K, McGuire F (1994) Barriers to implementing simulation in health care. In: Tew JD, Manivannan S, Sadowski DA, Seila AF (eds) Proceedings of the 1994 winter simulation conference. Institute of Electrical and Electronics Engineers, Piscataway, pp 868–875

Marcin J, Romano P (2004) Impact of between-hospital volume and within-hospital volume on mortality and readmission rates for trauma patients in california. Crit Care Med 32:1477–1483

McManus M, Long M, Cooper A, Litvak E (2004) Queuing theory accurately models the need for critical care resources. Anesthesiology 100:1271–1276

Milbrandt E, Kersten A (2008) Growth of intensive care unit resource use and estimated cost in Medicare. Crit Care 36

Mustafee N, Katsaliaki K, Taylor S (2010) Profiling literature in healthcare simulation. Simulation 86(8–9):543–558

Osorio C, Bierlaire M (2009) "An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. Eur J Oper Res 196:996–1007

Ridge J, Jones S, Nielsen M, Shahani A (1998) Capacity planning for intensive care units. Eur J Oper Res 105:346–355

Ross SM (1996) Stochastic processes, 2nd edn. Wiley series in probability and mathematical statistics. Wiley, New York

Seymour D (2001) Health care modelling and clinical practice: theoretical exercise or practical tool? Health Care Manag Sci 4(1):7–12

Shahani A, Ridley S, Nielsen M (2008) Modelling patient flows as an aid to decision making for critical care capacities and organisation. Anesthesia 63(10):1074–80

Shmueli A, Sprung CL, Kaplan EH (2003) Optimizing admissions to an intensive care unit. Health Care Manag Sci 6:131–136

Standridge C (1999) A tutorial on simulation in health care: applications and issues, vol 1. IEEE Computer Society, Los Alamitos, pp 49–55

Tarnow-Mordi W, Hau C, Warden A et al (2000) Hospital mortality in relation to staff workload: a 4-year study in an adult intensive care unit. Lancet 356:185–189

Taylor S, Eldabi T, Riley G, Paul R, Pidd M (2009) Simulation modelling is 50! Do we need a reality check? J Oper Res Soc 60

Vassilacopoulos G (1985) A simulation model for bed allocation to hospital inpatient departments. Simulation 45(5):233–241

Williams S (1983) How many critical care beds are enough? Crit Care Med 11(6)

# Chapter 7
# Improving the Flow of Patients Through Healthcare Organizations

**Steven M. Thompson, Robert Day, and Robert Garfinkel**

## 1 Introduction

Healthcare organizations (HCOs) face a challenging operating environment where delivering highly complex, specialized services requires patients to interact with a number of different types of healthcare professionals in a variety of settings. At the same time, advances in laboratory, radiological, and surgical technologies have dramatically increased the number of available tests and procedures. Further compounding these challenges, economic trends and demographic changes are pushing HCOs to deliver higher quality care to more patients at lower total cost.

As a result of increasing financial pressures, HCOs are trying to decrease the cost of providing healthcare by eliminating waste and improving the utilization of existing healthcare resources. Even though HCOs aspire to achieve high levels of staff, equipment, and space utilization, variability in demand and uncertainty in treatment and test duration can result in situations where a given resource is not available at the time it is needed. This creates a bottleneck that can cause a patient to experience a delay in treatment. These bottlenecks can be due to a lack of:

- Physical capacity: The provider does not have a sufficient number of beds, screening rooms, scanners, etc., to accommodate all demand.
- Scheduling: The provider does not take advantage of opportunities to balance demand for healthcare services with availability.

S.M. Thompson (✉)
Robins School of Business, University of Richmond, 1 Gateway Road, Richmond, VA 23173
e-mail: sthomps3@richmond.edu

R. Day • R. Garfinkel
University of Connecticut, 2100 Hillside Road, U-1041, Storrs, CT 06269-1041
e-mail: Bob.Day@business.uconn.edu; Robert.Garfinkel@business.uconn.edu

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_7, © Springer Science+Business Media New York 2013

- Staff: The provider does not have a sufficient number of perhaps multiple types of clinicians to meet all demand.
- Equipment and supplies: Capacity and staff are available, but various equipment, supplies, medical devices, etc., needed to conduct a given test or procedure are not available at that time.
- Information: In many cases, a patient flow bottleneck can occur because the clinicians do not have access to key pieces of information, such as lab results, consult reports from specialists, or radiological images that must be available before treatment can be initiated or completed.

Disruptions in patient flow are not without consequence. At a minimum, it can result in poor patient satisfaction and lower staff morale. At worst, delays in the start or completion of medical treatment can negatively impact the clinical outcome. As a result, the management of patient flow is fundamentally focused on striking a balance between keeping capacity investments as low as possible while ensuring that the HCO has sufficient capacity to handle patient demand without experiencing bottlenecks.

Achieving "good" patient flow has different meanings in different settings. In an emergency department, "good" flow for critically ill patients means that the patients can be treated immediately after the ambulance arrives. In contrast, "good" flow in a walk-in clinic could mean that patient can be seen in less than 30 min. Likewise, some healthcare services, such as surgery, are multistage processes that involve a range of clinicians, and the time that elapses from the patient's first encounter with the surgeon to their discharge from the hospital could span a period of weeks or months. In all settings, HCOs must determine how much capacity to make available and subsequently determine how to allocate that capacity to achieve high levels of utilization and quality of care.

Capacity decisions are complicated by the fact that demand for healthcare services can be difficult to predict. Efforts to improve flow after capacity decisions have been made are typically focused on decreasing total cycle time and eliminating waste by:

1. Decreasing the amount of time a patient spends in given stages of the healthcare delivery process
2. Decreasing demand for urgent services with highly variable resource needs by emphasizing preventive and disease-state management services
3. Decreasing the number of stages in the process
4. Performing stages of the process in parallel
5. Decreasing the amount of time required for a patient to move from one stage in the process to another

All of these efforts to improve flow are challenging because they require managing change in a high-stress environment where practitioners are accustomed to high levels of autonomy with respect to how healthcare is delivered. Patient flow problems are pervasive, and prior research has explored a variety of settings and levels of detail using a variety of empirical and analytical methods.

The rest of this chapter is organized as follows. In Sect. 2, we will provide a background on patient flow in order to understand why patient flow problems are so prevalent and the factors that make improving patient flow challenging. In Sect. 3 we will provide a brief overview of operations research on a variety of patient flow problems in a variety of settings. In Sect. 4 we provide a specific example of the complexity of breaking down and modeling patient flow problems in the context of a hospital-based surgery practice. In Sect. 5 we describe some future research opportunities.

## 2 Background on Patient Flow

Improvement initiatives in healthcare are challenging because HCOs are simultaneously executing a production process while providing a customized service. Like many organizations in other industries, HCOs seek standardized processes in order to improve efficiency and obtain economies of scale. Unfortunately, achieving standardization is often an elusive goal. The fundamental problem is that delivering customized healthcare services often requires clinicians to adapt the process. A patient may have allergies, preexisting conditions, or even personal beliefs that require the clinicians to deviate from the "plan" and use different amounts of different resources than otherwise expected.

The patient's perception of the quality of the service experience is also an important consideration. While it is tempting to at least begin modeling the flow of patients as a production process, HCOs should not lose sight of the fact that the patients (and clinicians) involved in service delivery must be accepting of any changes to the process. In healthcare, service quality is measured in terms of outcomes (i.e., did the treatment work?) and in terms of the individual experience (i.e., did I feel comfortable, respected, and valued?). This complicates efforts to better align the availability of healthcare resources with the demand for healthcare resources. For example, consider a patient with an upper respiratory infection that arrives at the clinic she uses for routine healthcare. The patient is not feeling well and would like to be seen quickly. While it may be more efficient to have the patient evaluated by the next available physician, the patient may also have a strong preference to be seen by the physician who takes care of her routinely. Honoring that preference means that one physician will be idle while a queue forms for the other. Ignoring the patient's preference enables the HCO to obtain better utilization and performance in the short term, but risks long-term patient attrition.

## 2.1 Unique Characteristics of Healthcare Services

Techniques commonly used in other service industries to shape demand and improve resource utilization while maintaining steady flow are often difficult to implement in healthcare. For example, an airline can overbook a flight and turn away passengers they are not able to accommodate, perhaps compensating them with a free ticket on a future flight. A hospital is not able to overbook an ICU and turn away patients that cannot be accommodated. Likewise, a hotel can reduce rates and attract more customers to improve room utilization during slow times. A hospital is not able to improve operating room (OR) utilization by offering discounts on knee replacements.

The fundamental distinction between healthcare and other service industries is that healthcare is a service that nobody wants to receive. That is not to say that healthcare is not highly valued; rather it means that people consume healthcare services because they *need* them. For the most part, these services are not pleasurable, disrupt our daily lives, and in many cases are anxiety-inducing experiences. Therefore, efforts to improve the flow of patients should not lose sight of the fact that the object moving through the production process is a person who, at that moment, is in a very emotionally stressful environment. The implication for patient flow improvement initiatives is twofold. First, the fact that patients may be experiencing emotional stress and are in need of healthcare services highlights the importance of maintaining good patient flow. That is, effective patient flow results in faster service with fewer disruptions in the provision of care. Second, demand management practices, such as dynamic pricing and other yield management techniques, are rarely used in healthcare because none of the patients really want to be there in the first place.

The healthcare industry is also distinctive in terms of the mechanisms through which HCOs are reimbursed for the services they provide. In most health systems, once an individual has access to health insurance, their ability to consume healthcare services is not dependent on their wealth or income. HCOs are subsequently reimbursed by fiscal intermediaries, either a governmental agency, for example, Medicare in the USA, or the patient's private insurer. The HCO will typically receive a fixed amount from the fiscal intermediary for treating a patient with a given ailment and a small co-payment from the patient. This co-payment is often a very small percentage of the total cost of providing service and is often not a function of the cost of providing the service (e.g., a co-payment for inpatient hospitalization in the USA might be $250 regardless of the procedures that are performed or the number of nights the patient stays in the hospital). As a result of risk pooling and subsequent payment through fiscal intermediaries, price-based market segmentation strategies are rarely used to manage demand for healthcare. With a few notable exceptions, such as concierge medicine where individuals pay an annual fee to a medical practice in exchange for higher levels of service (Stillman 2010), price-based market segmentation strategies are not effective because the healthcare industry does not contain a pricing mechanism that allocates scarce resources to those that place the highest value on those resources.

## 2.2 Demand Uncertainty

For a given healthcare service, demand can be uncertain in both volume and urgency. Further complicating matters is that even the time and resources needed to provide a given healthcare service can be highly variable. Seasonal variation in disease-state complications (like exacerbations of chronic obstructive pulmonary disease and asthma), influenza outbreaks, innovations in medical therapies, demographic and population changes, and economic factors can all impact the demand for various healthcare services.

Likewise, the urgency of the need for healthcare services is also highly variable. A patient needing the arthroscopic removal of a bone spur can be asked to wait for weeks before receiving surgery. A patient experiencing chest pain and in need of a coronary artery bypass graft must receive care almost immediately. This variation in urgency would be less problematic if it were as condition-specific as those two prior examples suggest. Unfortunately, assessing the urgency of demand is not condition-specific because patient-specific factors, such as the presence of multiple concurrent disease states, can influence that assessment. For example, a hypertensive patient with a blood pressure of 160/90 is not considered urgent, but if that patient also has been previously diagnosed with an abdominal aortic aneurism, then the blood pressure must be lowered immediately.

The impact of variable demand and variable urgency on efforts to improve patient flow is that some efficient production strategies used in other industries are difficult or impossible to implement. For example, *level scheduling* involves fixing capacity at a specific level and then scheduling patients into the next available time slot. A level scheduling approach is challenging to implement in an environment where patient acuity may be a factor and patient satisfaction with service quality is partly based on wait time. In most acute care settings the common operational strategy is to "chase demand." For example, rather than trying to influence the demand for emergency services or predict demand levels for a given time period, an emergency department manager will focus on developing a flexible staffing schedule that enables her to increase/decrease the number of nurses in response to changes in patient volume and acuity. In this example, flexible staffing could be achieved by making use of a "float pool" where nurses are moved from one patient care unit to another based on need or by scheduling so-called prn nurses who can be cancelled without pay if the actual work load for a given day is low.

Ambulatory settings, such as physician offices and clinics, handle a predominantly nonurgent population but must also be prepared to respond to urgent demand. In some cases, a patient arrives at an ambulatory, nonurgent setting presenting with a life-threatening condition. For example, what the patient thought was dizziness due to an inner ear infection is diagnosed as a mild stroke by the physician or a patient experiencing shortness of breath goes to a clinic thinking she has pneumonia but is diagnosed with congestive heart failure. Furthermore, and related to the service dimension of healthcare, while a patient might present with a medical condition that does not require immediate treatment, the patient may still desire immediate treatment.

In an increasingly competitive healthcare environment, where many providers are seeking to grow volume, meeting patient quality of service expectations is becoming increasingly important. In the USA, the desire for quick service has led to the proliferation of walk-in clinics, such as Patient First, that are taking volume from private physician practices. These walk-in clinics adopt a chase demand staffing and scheduling strategy that is similar to the strategy used by emergency departments. However, in the case of the walk-in clinic, it is not that the patients need to be seen right away but rather that they *want* to be seen. Providers that ignore patient preferences for convenience and speed of service in favor of short-term improvements in operational efficiency risk losing volume to those providers that are more responsive.

The level of competition for patients varies across health systems. While competition is significant in many parts of the USA, competition in systems that are financed entirely through public funding may be lower. Some health systems, such as Sweden, are between the two extremes. In the case of Sweden, a mixed model is emerging where some individuals receive publically funded healthcare and others have shifted to private insurance in order to obtain better access. Nevertheless, even fully public systems like Canada must be responsive to patient satisfaction. In Canada, policy makers are under constant pressure to improve the availability of healthcare resources and reduce the length of time patients must wait before receiving service.

## 2.3 Availability, Scalability, and Flexibility of Resources

The provision of healthcare services involves a number of specialized resources including human resources (such as physicians and nurses), technological resources (such as CT scanners and lab specimen processing equipment), and physical resources (such as beds and rooms). If any of the required resources is not available, then the delivery process stops and the patient must wait. While many HCOs adopt a chase demand strategy in order to meet patient expectations for prompt and quick service, many of the resources they need to provide healthcare services vary in terms of scalability and flexibility, especially in the short term. For example, consider an increase in the demand for inpatient beds from patients suffering from pneumonia. Even if the hospital has adequate staff, if it lacks sufficient bed capacity, it is not able to admit the patients. Likewise, if beds are available but the hospital lacks staff, it is not able to admit the patients without risking poor outcomes due to mistakes that occur as clinicians are overloaded. However, ensuring bed and staff availability requires the HCO to make decisions well in advance of when those resources are expected to be utilized. Unexpected spikes in demand can be very challenging to meet and often result in crowding conditions. Unexpected decreases in demand can be financially straining as the HCO is unable to cut fixed costs in response to a decrease in variable revenue.

In addition to costs that are sometimes prohibitively high, long-term capacity decisions are also typically influenced by government regulation. In the USA, HCOs must complete a *Certificate of Need* prior to making investments in certain types of capacity (e.g., increasing the number of inpatient beds, operating room suites, emergency department bays, and even acquiring certain surgical and radiological technologies, requires governmental approval). In other health systems, capacity is predetermined by central or local government planning agencies. The overall consequence of government regulation is that increasing capacity may not be a feasible choice for a HCO.

### 2.4 Patient Acuity

HCOs, for the most part, do not operate on a "first-come, first-served basis." When a patient with a severe illness or condition arrives, they preempt other patients who may have been scheduled to receive treatment weeks or months in advance. For example, when scheduling surgical procedures into a set of surgical suites, the scheduler must consider the availability of beds in the postanesthesia care unit (PACU) where patients recover immediately after surgery. The PACU needs of a patient vary based on patient condition and the procedure being performed. If the patient needs to spend one or more nights in the hospital after the surgery, then the scheduler must also consider the availability of inpatient beds on various units in the hospital. If too many surgical procedures finish at the same time, the PACU can be overloaded, and some patients must be recovered in the surgical suite, delaying the start of the next case. If a patient needs to go to an intensive care unit (ICU) after PACU but there is no bed available, then the patient must wait in the PACU until a bed becomes available. This decreases PACU capacity by one bed and increases the risk that the PACU will not be able to accommodate all scheduled surgical procedures.

In summary, HCOs face many idiosyncratic factors that make them distinctive. While many firms in other industries must deal with uncertainty in demand, HCOs must also consider individual-level factors such as acuity and concurrent disease states. The mechanisms through which providers are reimbursed and the pervasiveness of financial intermediaries are also unique to healthcare and impact patient demand and HCOs ability to shape that demand. Operations researchers must consider these factors because they help define the objectives of the HCO and operational constraints.

## 3 Operations Research and Patient Flow

The operations research community has a long history of addressing the patient flow challenges faced by HCOs. Over the past few decades, hundreds of healthcare-related articles have appeared in operations research literature, and most have focused on specific operational challenges faced by healthcare providers (see Fries 1979; Brotcorne et al. 2003; Cayirli and Veral 2003; Eldabi et al. 2007; Brailsford and Vissars 2010 for extensive reviews of the operations research literature). A great deal of this prior operations research is related to patient flow.

In fact, any research that seeks to improve performance metrics such as wait time, cycle time, and time in queue is fundamentally addressing a problem related to patient flow. Over time, the problems and research methods applied have become increasingly more advanced, reflective of the increasingly complex operating environment faced by HCOs. As HCOs continue to evolve new organizational structures and the delivery of health services transitions into new settings, many new challenges, opportunities, and research questions will certainly arise.

### 3.1 A Typology of Research on Patient Flow

Most operations research methods applied to patient flow problems can be clustered into setting-specific streams. This is driven by the fact that different healthcare settings have different priorities and face different operational constraints. Methods are considered generalizable if multiple providers in that setting could utilize the proposed technique. This trend of setting-specific research also has likely been influenced by the tendency of healthcare clinical researchers to conduct setting-specific studies.

Much of the operations research work in healthcare is understandably conducted in collaboration with clinicians or is strongly influenced by research that appears in journals targeted towards clinicians and clinical researchers. Research appearing in these clinical outlets tends to focus on specific settings with the goal of disseminating best practices to managers in those settings. Consistent with evidence-based medicine, where actions are only prescribed if robust double-blind peer reviewed research supports their efficacy, many of these studies identify operational processes that have been shown to perform well and communicate them so that other HCOs might emulate the practice.

For example, Abraham and Reddy (2010) conduct a case study to evaluate the design of workflows that prevent effective interdepartmental coordination and delay the transfer of patients between units. The authors provide a clear set of best practices that have been shown to successfully reduce patient flow delays due to poor interdepartmental information sharing and capacity planning. Similarly, Pikard and Warner (2007) describe how demand forecasting and management can be integrated with staffing to improve patient flow through the hospital. In that case, the authors show how patient backlogs can be reduced by making staffing and scheduling decisions jointly. Dexter (2007) provides guidance on how managers of surgical services can integrate information from electronic medical record systems to improve staffing decisions by making sure PACU staffing levels reflect the expected volume of patients flowing out of the surgical suites. This guidance can help managers avoid situations where, for instance, the entire staff of a surgical suite is running into overtime because a single PACU nurse was sent home.

While the methods are very different, the research appearing in clinical outlets is useful to operations researchers for two reasons. First, it helps convey the objectives of HCOs and provides insight into the operational constraints that appear in different settings. Second, it can serve as a benchmark for measuring the impact of an alternative method. These industry benchmarks reflective of current best practice are often used as a basis for comparison in simulation studies. The majority of the operations research applications related to patient flow have subsequently followed clinical research streams and focused on improving flow within specific clinical environments.

## *3.2 Ambulatory Care Settings: Clinics and Physician Offices*

Clinics and physician offices have an operational environment that is distinctive from other HCOs that provide at least some acute care services. First, they are typically not open 24 hours per day, and all previously scheduled appointments for each day must be completed before the HCO can close. As a result, meeting scheduled demand while minimizing the occurrence of overtime, where staff are paid a premium rate for each hour they work beyond their scheduled shift, is a consideration. Second, the majority of the demand is not urgent so scheduling becomes an important mechanism to balance demand with supply. Third, patients expect to be seen and serviced quickly, regardless of whether they have a scheduled appointment, so total cycle time is an important performance metric.

Prior research on clinic operations has employed both simulation and analytical modeling. Glowacka et al. (2009) use association rule mining to find patterns in patient "no shows," where patients fail to appear for a previously scheduled appointment, in combination with simulation techniques to develop scheduling rules that improve patient flow by reducing patient waiting time and reducing within process idle time. Lumus et al. (2006) combine Monte Carlo simulation and value-stream mapping to develop a scheduling template that improves patient flow and resource utilization for a walk-in clinic. In that case, the authors develop scheduling guidelines to balance demand across the day in order to decrease clinician idle time and reduce patient wait times. Specifically they find that based on actual demand patterns, the clinic should deliberately leave openings on the schedule for last-minute requests, and the number of openings per hour should be lower in the morning and higher in the afternoon. Ashton et al. (2005) also use discrete event simulation to evaluate different scheduling and decision policies on the flow of patients through a multiservice clinic where any given patient may receive multiple services, while Davies and Roderick (1998) use discrete event simulation to estimate the resources that will be required to meet future demand for renal services.

Discrete event simulation has been a popular simulation method for clinic-based research largely because of its flexibility to model multistage processes, subject to uncertainty, that unfold over time and incorporate complex decisionrules (Jun et al. (1999) provide a detailed literature review of the applications of discrete event simulation in healthcare clinic settings). For example, Cote (1999) uses discrete event simulation to show that clinics providing complex, potentially multistage healthcare services with high utilization of some resources, such as examining rooms, do not necessarily exhibit longer waiting times. The purpose of these kinds of studies is to help HCOs determine the true sources of patient flow bottlenecks and allocate resources accordingly.

White et al. (2011) also utilize discrete event simulation to show that integrated management of capacity, patient flow, and scheduling can improve utilization while reducing patient waiting times. Specifically, they use discrete event simulation to show the impact of integrated scheduling and capacity allocation policies. They conclude that patient waiting time and total cycle time can be enhanced by scheduling shorter appointments that exhibit less variance in duration early in the day and longer, more variable appointments later in the day. They also find that overall utilization will remain high, regardless of scheduling mechanisms, provided capacity is allocated such that the physician is the bottleneck. Considering capacity planning and scheduling at the same time is uncommon; most other work in this stream falls into one of two categories. The first assumes that capacity is fixed and subsequently develops scheduling mechanisms to improve performance under a capacity constraint. The second assumes some distribution of demand and determines how much capacity is required to meet certain service levels.

Various analytical methods have also been brought to bear on the problem of improving the flow of patients in clinic settings. For example, Oddoye et al. (2009) combine queuing and goal programming to evaluate the trade-off between resource availability and patient flow metrics such as wait time and cycle time in a multistage medical assessment unit in the UK. The authors use simulation to identify different scenarios for removing patient flow bottlenecks and then submit the results of each simulated scenario to a goal program. The goal program then selects the scenario that best reflects the organizations priorities. Bretthauer and Cote (1998) illustrate how queuing and mathematical programming can be used to make capacity decisions in a variety of ambulatory healthcare settings including blood bank clinics. Jiang and Giachetti (2008) develop a queuing network model to evaluate the impact of process parallelization, staff specialization, and coordination on patient flow through a multiservice walk-in clinic. They find that process parallelization is beneficial in reducing total cycle time and variation in cycle time in settings where patients require more than one diagnostic procedure and more than one treatment procedure.

The scheduling literature has also considered numerous deterministic and stochastic ambulatory healthcare settings (see Cayirli and Veral (2003) for a review of the outpatient scheduling literature). Much of the work on patient scheduling has a direct impact on patient flow and performance metrics such as patient waiting time and length of queue figure prominently.

## 3.3 Emergency Departments

Unlike clinics, emergency departments are typically open 24 h per day, every day of the year. They also differ from clinics in that they service two groups of patients: those with urgent treatment needs and those with nonurgent treatment needs. Planning is complicated by the fact that the relative size of each group can vary considerably over the course of the day and on different days of the week. Emergency departments also face downstream bottlenecks that clinics do not face. In many cases a patient that is treated in the emergency department will need to be admitted to the hospital rather than sent home. As a result, the number of beds available in the hospital and how those beds are allocated can impact the flow of patients through the emergency department. If patients that need to be admitted to a hospital ward begin to backlog in the emergency department, the effective capacity of the emergency department is reduced and the amount of time newly arrived patients must wait prior to service increases.

Research related to emergency department (ED) operations has tended to focus on one of three levels: within-department, within-facility, and acrossfacilities. The first level represents research that looks specifically at the ED and does not consider other units in the hospital. For example, Cochran and Roche (2009) develop a queuing network to model the impact of emergency department resource availability (e.g., staff, # of treatment bays) on total treatment time, staff and room utilization, and queue length. The objective of this kind of study is to help ED managers determine how much capacity to make available to achieve patient flow performance targets.

At the second level, research that considers other units in the hospital seeks to identify and eliminate downstream bottlenecks that are preventing the ED from achieving its patient flow performance targets. Ceglowski et al. (2007) combine data mining and discrete event simulation to develop a model that enables hospitals to identify bottlenecks that form between the ED and other departments, but do not offer guidance on how to alleviate those bottlenecks. Thompson et al. (2009) combine integer programming and Monte Carlo simulation to develop a decision-support tool to assist hospital bed managers as they allocate newly arrived patients and reallocate existing patients to various specialized units in a hospital in order to avoid bed-related bottlenecks that can impact the ED.

At the third level, research tends to define the problem setting as one of multiple interrelated HCOs where decisions made by one can impact patient flow in the others. For example, Patrick (2011) employs simulation and Markov decision models to evaluate the impact of long-term care bed management policies on ED crowding. He finds that the shortage of long-term beds results in patients spending additional nights in the hospital. This prevents patients from leaving the ED, and crowding conditions quickly follow.

## *3.4 Hospital Inpatient Units*

Hospital inpatient units represent an enormous financial investment. Construction costs alone can range from tens of millions of US dollars to hundreds of millions, depending on the size of the facility. Subsequent staffing costs add considerably more to the total. The majority of these costs are fixed expenses that do not rise (or fall) in response to actual demand. That being the case, most research has been focused on trying to determine how much hospital capacity should be made available in order to achieve clinical and quality of service performance goals. Incorporating patient flow pathways, determining flow performance metrics, and understanding and modeling different flow patterns among different patient populations are important elements of these studies (Weiss et al. 1982).

As with most other healthcare settings, Monte Carlo and discrete event simulation figure prominently as research methods. Vissars et al. (2007) use discrete event simulation to evaluate the differential performance between chase demand and level scheduling for hospital admissions. They find that while level scheduling is more efficient in that it achieves higher levels of utilization, chase demand, where resources are made available as needed in response to fluctuations in demand, results in shorter wait times for patients. The trade-off is that efforts at achieving higher levels of utilization and efficiency may negatively impact patient flow. At a more unit-specific level of analysis, Ridge et al. (1998) use queuing and simulation to develop a stochastic model to support ICU capacity planning. In that setting, lack of available ICU beds can negatively impact patient flow because patients in need of intensive care are either held in the ED or transferred to another hospital. In either case, the patient experiences a delay before they are placed in a proper venue. Likewise, de Bruin et al. (2010) develop methods for determining how much capacity should be available on each inpatient ward. They also show that establishing the same target for bed utilization across all units can result in excessive refused admissions, especially for smaller units. Again, patients that are refused experience a delay in treatment. Depending on the specific medical condition, the patient may have to wait in the ED or be transferred to another hospital. Vissars (1997) conducts a case study and uses empirical methods and Monte Carlo simulation to develop a mechanism for allocating human and physical resources to various departments within a hospital. The mechanism is reflective of anticipated demand but also considers long-term changes in patient mix, demographics, and population size.

While capacity planning requires a detailed specification of how patients flow through the various units within the hospital, these flows can vary dramatically from one hospital to another. This variability has led to some work dedicated towards developing modeling conventions that can aid future research. Harper (2002) develops and illustrates an elaborate framework to guide the design of hospital-based simulation studies. Likewise, Lane and Husemann (2008) develop a framework and modeling convention to describe patient flow pathways in order to provide a detailed yet generalizable framework that can be used as the basis for future analysis, research, and planning. In a similar vein, Fackrell (2009) reviews the use of phase-type distributions to model the flow of patients within and among hospitals and discusses the technical implications for the design of future research studies. These works share a common theme in that they seek to provide a mechanism to evaluate the impact of capacity decisions on the flow of patients through the system.

## *3.5 Surgical Suites*

Improving the utilization and performance of surgical suites is one of the most important considerations for managers of HCOs that provide surgical services. Whether located in an ambulatory surgical center (ASC) that deals with nonurgent, nonlife-threatening conditions or in an acute care, hospital setting, surgical suites are associated with very high costs and potentially very highrevenue.

In the USA, the cost of running a single surgical suite has been estimated to be $20 *per minute* when all overhead costs are included (Dexter et al. 2002). On the other hand, surgical suites are a major source of revenue for hospitals and ASCs. The financial benefit from performing a surgical procedure is referred to as a contribution margin. The contribution margin is a measure of the revenue the hospital receives above the variable costs of supplies and devices needed to perform the surgical procedure. In the USA, the contribution margin for a given surgical procedure can range from a few hundred dollars to a few thousand dollars *per hour* of surgical time.

Providing surgical services differs from other healthcare services in that the patient's perception of the quality of service is less important than in other settings. While wait time and cycle time are very important considerations when evaluating the performance of an ED or walk-in clinic, patients receiving surgical services are usually sedated or anesthetized and do not perceive the passage of time. As a result, research focused on improving operating room performance tends to frame the problem as one of improving the throughput of a production process and places less emphasis on patient-specific quality of service considerations.

Again, simulation is a common research methodology (see Kuzdrall et al. 1974, one of the first studies to use computerized Monte Carlo simulation to improve throughput of surgical suites). More recently, Persson and Persson (2010) employ queuing and discrete event simulation to compare different policies for the allocation of available surgical suite time. In that problem setting, the hospital must decide how much time to allocate to elective, nonurgent cases and how much time to allocate to urgent cases. They model the impact of different case scheduling policies on utilization and throughput by examining the flow of patients through the entire process that begins with preoperative preparation, proceeds to the performance of the surgical procedure, and concludes with postoperative recovery in the PACU. On a related topic, but at a finer level of detail, Fredendall et al. (2009) combine empirical analysis and process improvement methodologies to develop workflows that improve patient throughput.

Some researchers have also used simulation to study the impact of pre-hospital policies on surgical suite throughput. For example, Vasilakis et al. (2007) use discrete event simulation to compare the impact of placing nonurgent patients on a single list and assigning them to the next available physician rather than using the current, common practice of assigning patients to physician-specific lists. They find pooled lists increase surgical suite throughput and reduce physician idle time. The benefit is identical to the benefit received when transitioning from a multiple queue/multiple server model to a single queue/multiple server model.

Operations researchers have also used analytical modeling techniques to improve surgical suite throughput and performance. Adan et al. (2009) develop a mixed integer linear programming model to determine how much surgical suite time should be allocated to different surgeons from different specialties while also considering downstream inpatient bed availability in order to achieve throughput objectives. In that setting, throughput is important for two reasons. First, if the surgical suites experience a backlog, then patients must wait longer in the preoperative holding unit or the procedure must be rescheduled for a later date. Second, surgeons place a very high value on their time. If a surgical procedure is not able to start at the scheduled time, then the surgeon is idle until the case can be performed. Price et al. (2011) utilize integer programming and discrete event simulation to develop a scheduling mechanism that balances new surgical procedures that require inpatient beds with anticipated discharges. The basic problem is that if a bed is not available, then the patient in need of that bed will linger in the PACU. If the patient remains in the PACU too long, then the PACU does not have the capacity to accommodate other patients scheduled for recovery. If patients are not able to enter PACU, then the anesthesiologist must recover the patient in the surgical suite. This delays the start of the next case, and if the problem persists, the entire case schedule can be disrupted. Similarly, Gupta 2007 uses mathematical programming to evaluate the impact of downstream resource availability on OR throughput and discusses the implications for scheduling and capacity planning.

### *3.6 Health System Planning*

While most operations research applications in healthcare are focused on improving patient flow via process improvement and HCO-specific capacity planning, another stream of research frames the patient flow problem at the level of the overall healthcare system. In some cases, the research question is focused on a particular health service. For example, Andersson and Varbrand (2007) use integer programming and simulation to develop decision heuristics for ambulance dispatch and relocation to minimize transit time for patients. Almost all of the numerous papers focused on the provision of emergency medical services (EMS) include one or both of two patient flow metrics: decreasing the amount of time that elapses before a patient receives

EMS and decreasing the total cycle time from the initial call for help to the time the ambulance arrives at a hospital.

A study by Asaduzzaman et al. (2010) is another example of system level analysis for a specific healthcare service. In that study the authors develop a queuing model that enables a planner to evaluate the implications of various capacity planning decisions on the availability of neonatal cots. In this case, the authors frame the problem such that all neonatal beds in the healthcare system are considered along with the time and cost of travel.

Other works at this level of analysis are more comprehensive in that they consider all health services, not just a single service like EMS. However, these studies tend to include less detail and treat demand in aggregate. For instance, van Zon and Kommer (1999) develop a dynamic LP that guides central planners as they seek to allocate and reallocate healthcare resources in response to changing health trends, demographic changes, population growth and shifts, and changes in case mix. At this level of analysis, researchers tend to model monthly or annual demand and avoid the day-to-day and hour-to-hour fluctuations that are common considerations for HCO-specific research.

## 4 Example: Patient Flow in a Hospital-Based Surgery Practice

To illustrate the challenges and opportunities associated with applying operations research techniques to patient flow problems, we take a closer look at the capacity planning and resource allocation decisions associated with the provision of surgical services in a typical US hospital. In essence, hospitals make a series of capacity decisions where decisions at one stage place a constraint on the maximum available capacity at a subsequent stage. As a result, patient flow bottlenecks can be the result of near-term and long-term capacity decisions and scheduling decisions that can impact how capacity is utilized. Figure 7.1, adapted from Olivares et al. 2008, illustrates the high-level macro-process and the frequency of each step. Note that while the macro-process shown below is general to most healthcare systems, the information and objectives considered at each stage can vary considerably between privatized healthcare systems and publically funded healthcare systems.

The first three steps all involve determining how much surgical suite capacity to make available, but they differ in terms of objective and/or the information that is considered. The final two steps are focused on scheduling urgent and nonurgent cases into available staffed capacity.

In order to guide scheduling, many hospitals have adopted the practice of block scheduling (Dexter et al. 2005). The block schedule specifies how much surgical suite time is allocated to each surgeon or surgical subspecialty for each day of a one- or two-week period. The block schedule then serves as a template that recurs over time. The block schedule serves two main purposes. First, from the perspective

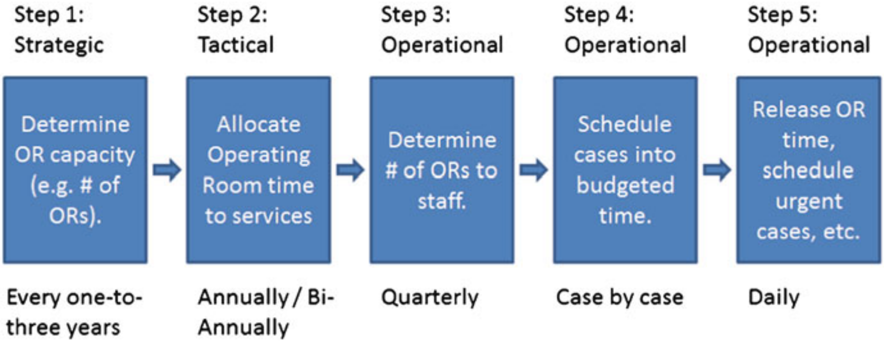

Step 1:
Strategic
Step 2:
Tactical
Step 3:
Operational
Step 4:
Operational
Step 5:
Operational
Determine OR capacity (e.g. # of ORs).
Allocate Operating Room time to services
Determine # of ORs to staff.
Schedule cases into budgeted time.
Release OR time, schedule urgent cases, etc.
Every one-to-three years
Annually / Bi-Annually
Quarterly
Case by case
Daily

**Fig. 7.1** Determining and allocating surgical suite capacity

of the hospital, it is used to guide the scheduling of surgical procedures and is helpful in identifying gaps in staffing. Second, it helps surgeons plan their time efficiently. This is very important because most surgeons maintain private practices and utilize the hospital for the provision of some, but not all, surgical procedures. The block schedule can have a significant impact on patient flow. If the time allocated to different surgeons and surgical subspecialties does not reflect actual demand distributions, the capacity is misallocated. As a result surgeons can suffer negative consequences including difficulty scheduling procedures and a greater likelihood that the case will be performed late in the day after other procedures have been finished.

The Integrated Block Sharing (IBS) mechanism developed in Day et al. (2012) is based on two central concepts that are essentially relaxations of assumptions made in prior research. The first relaxation is that the IBS mechanism allows block time to be granted in a variety of different configurations. Prior research assumes block time is either granted to individual surgeons or to entire surgical subspecialties (e.g., block time is allocated to, say, orthopedics or vascular surgery in general, and the schedule is silent on how that time is shared among individual surgeons). IBS allows for block time to be granted to individual surgeons, small groups of surgeons, or any other configuration that improves surgical suite utilization.

The second relaxation is that the IBS mechanism allows for nonurgent cases to be scheduled during a rolling two-week window of time rather than on the day the surgeon requests. That is, IBS assumes a level scheduling approach, and this assumption directly impacts how much block time surgeons need. Although demand is uncertain, the ability to load balance that demand means that a given level of service can be provided using fewer staffed hours of surgical suite time than if load balancing were not possible. The ability to load balance also impacts the decision of determining of how many rooms to open on a given day of surgery. Since requests for surgical suite time can be met at a later date, schedulers can avoid opening surgical suites to accommodate a single case in favor of putting the case on a different day where the room can be more fully utilized.

The impact on patient flow is twofold. First, since IBS allows groups of surgeons to “share” block time, lower-volume surgeons no longer have to wait until a few days before the actual surgery to schedule a case. This allows the patient to plan in advance and makes it possible to schedule any required preoperative tests or examinations in a manner that accommodates the patient’s schedule. Second, load balancing demand has the net effect of reducing the peak volume of the number of procedures performed each day and smoothing the variability in demand experienced by the PACU and inpatient units.

## 5 Future Research Opportunities

As healthcare systems strive to improve patient flow, operations research will play a strong role in shaping the operational processes of the future.

Across the industrialized world, a growing number of HCOs are gradually transforming into increasingly large organizations. There are a number of reasons for this growth. First, larger organizations have the ability to dedicate resources to initiatives that would not be economically feasible for smaller organizations. Investments in analytics and process improvement are a good example. Given the cost of highly trained, specialized operations research analysts, any investment in process improvement only makes sense if the improved process can be executed numerous times. Larger organizations, through sheer volume, are better suited to invest in process improvement initiatives because it is more likely they will yield a return on that investment.

As talent and opportunity collide, HCOs are identifying more areas that can benefit from the application of operations research methods. For example, HCA, Inc., a for-profit health system that owns and operates 164 hospitals and 106 free-standing surgical centers in the USA and UK, maintains a team of “management engineers,” who are essentially industrial engineers with extensive training in healthcare. That represents an investment that is far beyond the reach of a stand-alone community hospital. For HCA, the investment has paid off. HCA hospitals are frequently named among the most efficient in the USA with respect to achieving good outcomes, delivering high-quality service, and keeping costs low.

On the other hand, a natural consequence of an increase in size is a corresponding increase in bureaucracy, more complex internal controls, and a more challenging change management environment. While greater size provides the internal resources needed to identify and formalize improvement opportunities, larger organizations must remain capable to move ahead with new ideas.

The implications for operations researchers are that a growing number of HCOs are capable to evaluate and implement complex optimization models and simulations into clinical practice. As these organizations deploy more complex information systems, there are opportunities for OR researchers to develop algorithms and optimization routines to help ensure that limited healthcare resources are used as productively as possible. However, since larger organizations face a more complex change management environment, process improvement efforts must consider the needs and objectives of multiple stakeholder groups. While considering multiple objectives can increase the complexity of the problem, it can also result in solutions that, once found, are easier to implement into practice.

Another area of opportunity is that many HCOs are evolving into more complex organizations that provide a broader array of healthcare services. This trend is driven by a variety of factors including a desire to achieve economies of scale, obtain bargaining power over suppliers and insurance companies, and secure patient referral pipelines. The growth in complexity comes as HCOs expand horizontally and vertically (through internal growth initiatives or through mergers and acquisitions) across the healthcare value chain. Horizontal integration involves operating two or more organizations that perform the same function within the healthcare value chain. For example, three separate hospitals in a given metropolitan area might actually be a single HCO.

Vertical integration involves operating organizations that perform different functions within the healthcare value chain. For example, in the USA, hospitals and health systems have been purchasing clinics and private medical practices. The motivation for a hospital to purchase, say a cardiology practice, is to secure the pipeline of referrals and procedures generated by that practice. In the past the cardiologist may have performed some procedures at one hospital and other procedures at a different hospital. By acquiring the practice, the hospital is able to grow volume and revenue because it can now perform all of the physician's procedures. The resulting vertical integration also helps to ensure that the interests and priorities of physicians and hospitals are better aligned. In the past, the physician was indifferent as to where she performed the procedures. As an employee of the hospital, she shares an interest in the economic well-being of her employer.

Vertical integration provides HCOs access to information that was not available in the past and sets the stage for a level of collaboration among clinical providers that did not exist in the past. Regarding access to information, many of the studies mentioned earlier have treated demand as exogenous to the problem. That assumption was fine and did reflect the perceived operational environment of the HCO at that time. However, the actual situation was a bit more nuanced. Front line HCOs, like physician offices and clinics, have a great deal of information that is potentially useful to demand forecasting and resource planning. Unfortunately, downstream providers did not historically have access to that information because the front line HCOs had little incentive to share the information. As these different HCOs merge to become vertically integrated entities, they will have access to information that was never available to them in the past. This creates opportunities to develop new forecasting models, scheduling methods, and resource allocation policies that achieve better performance by taking advantage of new information.

By aligning the interests of formerly disparate HCOs, vertical integration also creates opportunities to improve coordination across the healthcare delivery process. In the past, downstream HCOs had to treat physicians as customers. Initiatives such as efforts to improve surgical suite utilization through the allocation of surgical suite time to various surgeons and efforts to decrease inpatient length of stay through more frequent physician assessments were all dependent on whether the involved physicians were supportive. Gaining that support is often difficult. In many cases where physicians are independent entities, they are understandably not overly enthusiastic about initiatives that required them to do more work or be more flexible without any additional compensation. Vertical integration provides an environment where the incentives to collaborate are stronger and the leaders of the HCO can exert a greater degree of managerial control.

The presence of more vertically and horizontally integrated organizations creates a number of new research opportunities. For example, vertical integration enables the HCO to more precisely integrate downstream resource availability with scheduling. In the past, private physicians did not consider the needs of any given hospital as they scheduled patient procedures. Once the physician practice has been integrated, new processes are possible. Likewise, horizontal integration creates the opportunity to pool similar resources. This would require new processes and procedures for determining how the pooled resources should be allocated.

Information technology is also creating new opportunities for operations research. While the specific drivers of information technology (IT) adoption vary across the health systems of industrialized nations, two trends have emerged. IT is becoming more prevalent and more mobile. HCOs now have access to more data than ever before, and it is increasingly available in real time. This enables improved performance analysis and also makes it possible to develop decision-support tools that, in combination with mobile computing devices, actually fit the workflow.

The increased flow of information can potentially decrease the amount of time needed to provide care, as laboratory and radiography results are made available more quickly. Increased information transparency can be used to help bed managers identify which inpatient units have beds and allocate patients strategically to avoid inadvertently creating patient flow bottlenecks. Improved information flow between HCOs will make it easier to transfer patients to the next level of care by making it easier for case workers to find suitable long-term care, rehabilitation, and psychiatric beds.

On the other hand, HCOs are challenged to transform increasingly large amounts of data into policies that positively impact patient flow. While the ability of information systems to collect and disseminate information is critical, mathematical modeling and operations research techniques are equally important in transforming data into actionable knowledge. For example, decisions on how to disperse patients to skilled rehabilitation facilities should reflect future expectations of discharges and admissions.

## 6 Conclusions

The healthcare industry has not historically been considered a pioneer in change management and process improvement. Long considered “behind” on information technology (IT) and having “discovered” Six Sigma and Lean decades after other industries, the healthcare industry has a reputation for rapid advances in medical technologies but slow progress on cost containment and efficiency metrics.

However, the catalyst for future progress is present: lower margins. One of the greatest barriers to process improvement, and the change management that ensues, is a lack of institutional support. While other industries are dominated by for-profit, private sector firms, many of the largest HCOs are public, not-for-profit firms that lack the internal incentives for-profit firms use to drive innovative behavior. In most cases, improving a process does not benefit any individual in the HCO in any measurable way. While patients might perceive the quality of service as "better" if they spend less time in the waiting room, those involved in the modified process did not receive bonuses or higher pay. The net effect of the lack of incentives was that many process improvement efforts were not pursued because the perceived benefit or reward was outweighed by the perceived cost of change.

HCOs are increasingly finding that decreased reimbursement levels from private insurers and government agencies are resulting in lower margins on the healthcare services they provide. Lower margins have also increased the urgency for improved delivery processes and higher levels of utilization. While many HCOs do not want to change, almost all will have to change in order to survive. The need to change sets the stage for operations researchers to explore new healthcare delivery models.

## References

Abraham J, Reddy MC (2010) Challenges to inter-departmental coordination of patient transfers: a workflow perspective. Int J Med Informat 79:112–122

Adan I, Bekkers J, Dellaert N, Vissars J, Yu X (2009) Patient mix optimization and stochastic resource requirements: a case study in cardiothoracic surgery planning. health-care Manag Sci 12:129–141

Andersson T, Varbrand P (2007) Decision support tools for ambulance dispatch and relocation. J Oper Res Soc 58:195–201

Asaduzzaman MD, Chaussalet TJ, Robertson NJ (2010) A loss network model with overflow for capacity planning of a neonatal unit. Ann Oper Res 178:67–76

Ashton R, Hague L, Brandreth M, Worthington D, Cropper S (2005) A simulation-based study of a NHS walk-in center. J Oper Res Soc 56:153–161

Brailsford S, Vissars J (2010) OR in healthcare: a European perspective. Eur J Oper Res 212: 223–234

Bretthauer KM, Cote MJ (1998) A model for planning resource requirements in health-care organizations. Decis Sci 29(1):243–270

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. Eur J Oper Res 147:451–463

Cayirli T, Veral E (2003) Outpatient scheduling in healthcare: a review of literature. Prod Oper Manag 12(4):519–549

Ceglowski R, Churilov L, Wasserfiel J (2007) Combining data mining and discrete event simulation for a value-added view of a hospital emergency department. J Oper Res Soc 58:246–254

Cochran JK, Roche KT (2009) A multi-class queueing network analysis methodology for improving hospital emergency department performance. Comput Oper Res 36:1497–1512

Cote MJ (1999) Patient flow and resource utilization in an outpatient clinic. Socioecon Plann Sci 33:231–245

Davies R, Roderick P (1998) Planning resources for renal services throughout UK using simulation. Eur J Oper Res 105:285–295

Day R, Garfinkel R, Thompson S (2012) Integrated Block Sharing: A Win-Win for Hospitals and Surgeons. Manuf Serv Oper Manag 14(4):567–583

de Bruin AM, Bekker R, van Santen L, Koole GM (2010) Dimensioning hospital wards using the Erlang loss model. Ann Oper Res 178:23–43

Dexter F, Blake JT, Penning DH, Sloan B, Chung P, Lubarsky DA (2002) Use of linear programming to estimate impact of changes in a hospital's operating room time allocation on perioperative variable costs. Anesthesiology 96:718–724

Dexter F, Ledolter J, Wachtel R (2005) Tactical decision making for selective expansion of operation room resources incorporating financial criteria and uncertainty in subspecialties' future workloads. Anesth Analg 100:1424–1432

Dexter F (2007) Bed management displays to optimize patient flow from the OR to the PACU. J Perianesth Nurs 22(3):218–219

Eldabi T, Paul RJ, Young T (2007) Simulation modelling in healthcare: reviewing legacies and investigating futures. J Oper Res Soc 58:262–270

Fackrell M (2009) Modelling healthcare systems with phase-type distributions. Health-care Manag Sci 12:11–26

Fredendall LD, Craig JB, Fowler PJ, Damali U (2009) Barriers to swift, even flow in the internal supply chain of perioperative surgical services department: a case study. Decis Sci 40(2): 327–349

Fries BE (1979) Bibliography of operations research in healthcare systems: an update. Oper Res 27(2):408–419

Glowacka KJ, Henry RM, May JH (2009) A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. J Oper Res Soc 60:1056–1068

Gupta D (2007) Surgical suites' operations management. Prod Oper Manag 16(6):689–700

Harper PR (2002) A framework for operational modelling of hospital resources. health-care Manag Sci 5:165–173

Jiang L, Giachetti RE (2008) A queueing network to analyze the impact of parallelization on patient cycle time. health-care Manag Sci 11:248–261

Jun JB, Jacobson SH, Swisher JR (1999) Application of discrete event simulation in health-care clinics: a survey. J Oper Res Soc 50(2):109–123

Kuzdrall PJ, Kwak NK, Schmitz HH (1974) The Monte-Carlo simulation of operating room and recovery room usage. Oper Res 22(2):434–440

Lane DC, Husemann E (2008) System dynamics mapping of acute patient flows. J Oper Res Soc 59:213–224

Lumus RR, Vokurka RJ, Rodeghiero B (2006) Improving quality through value-stream mapping: a case study of a physicians clinic. Total Qual Manag 17(8):1063–1075

Oddoye JP, Jones DF, Tamiz M, Schmidt P (2009) Combining simulation and goal programming for healthcare planning in a medical assessment unit. Eur J Oper Res 193:250–261

Olivares M, Terwiesch C, Cassorla L (2008) Structural estimation of the newsvendor model: an application to reserving operating room time. Manag Sci 54(1):41–55

Patrick J (2011) Access to long-term care: the true cause of hospital congestion? Prod Oper Manag 20(3):347–358

Persson MJ, Persson JA (2010) Analyzing management policies for operating room planning using simulation. health-care Manag Sci 13:182–189

Pikard B, Warner M (2007) Demand management: a methodology for outcomes-driven staffing and patient flow management. Nurse Leader 4:30–34

Price C, Golden B, Harrington M, Konewko R, Wasil E, Herring W (2011) Reducing boarding in a post-anesthesia care unit. Prod Oper Manag 20(3):431–441

Ridge JC, Jones SK, Nielson MS, Shahani AK (1998) Capacity planning for intensive care units. Eur J Oper Res 105:346–355

Stillman M (2010) Concierge medicine: a "regular" physician's perspective. Ann Intern Med 152(6):391–392

Thompson S, Nunez M, Garfinkel R, Dean M (2009) Efficient short-term allocation and re-allocation of patients to floors of a hospital during demand surges. Oper Res 57(2):261–273

van Zon AH, Kommer GJ (1999) Patient flows and optimal resource allocation at the macro-level: a dynamic linear programming approach. health-care Manag Sci 2:87–96

Vasilakis C, Sobolev BG, Kuramoto L, Levy AR (2007) A simulation study of scheduling clinic appointments in surgical care: individual surgeon versus pooled lists. J Oper Res Soc 58: 202–211

Vissars JMH (1997) Patient-flow based allocation of inpatient resources: a case study. Eur J Oper Res 105:356–370

Vissars JMH, Adan IJBF, Dellaert NP (2007) Developing a platform for comparison of hospital admission systems: an illustration. Eur J Oper Res 180:1290–1301

Weiss EN, Cohen MA, Hershey JC (1982) An iterative estimation and validation procedure for specification of semi-Markov models with application to hospital patient flow. Oper Res 30(6):1082–1104

White DL, Froehle CM, Classen KJ (2011) The effect of integrated scheduling and capacity policies on clinical efficiency. Prod Oper Manag 20(3):442–455

# Chapter 8
# Capacity Allocation and Flexibility in Primary Care

**Hari Balasubramanian, Ana Muriel, Asli Ozen, Liang Wang, Xiaoling Gao, and Jan Hippchen**

## 1 Introduction

Primary care providers (PCPs) are typically the first point of contact between patients and health systems. Broadly they include family physicians, general internists, geriatricians, and pediatricians. From a patient's perspective, PCPs provide the majority of care they receive during their lifetime. They are responsible for a variety of health services including preventive medicine, patient education, routine physical exams, and referrals to medical specialties for secondary and tertiary care. The benefits of a strong primary care system are well documented in the clinical literature. Papers in the health services literature, Shi et al. (2005) for example, have established that increased access to primary care (1) improves access to health services for relatively deprived population groups, (2) has a strong positive relationship with prevention and early management of health problems, and (3) leads to increased familiarity with patients and, consequently, to less wasteful expenditures due to inappropriate specialist care.

Despite its pivotal role, primary care physicians receive lower salaries than specialists, which has the effect of driving medical students away from pursuing primary care careers. This is one of the main reasons for shortages in primary care, which are common in many countries. A recent study by the American College of Physicians (American College of Physicians 2006) reports that primary care in the USA "is at grave risk of collapse due to a dysfunctional financing and delivery system." They also emphasize the growing demand for primary care; by 2015, "an estimated 150 million Americans will have at least one chronic condition."

H. Balasubramanian (✉) • A. Muriel • A. Ozen • L. Wang • X. Gao • J. Hippchen
Department of Mechanical and Industrial Engineering, University of Massachusetts, 160 Governors Drive, MA 01003, Amherst
e-mail: hbalasubraman@ecs.umass.edu; Muriel@ecs.umass.edu; aslozen@gmail.com; liangwanghust@gmail.com; xiaoling@engin.umass.edu; janhippchen@gmx.de

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_8, © Springer Science+Business Media New York 2013

Timely access to care and patient–physician continuity, the two metrics important to primary care practices, have been adversely affected due to these broader trends. The focus on timely access, or the ability to secure an appointment quickly, is well known in the operations research literature. The inability to get a timely appointment to a primary care physician increases the likelihood of patients visiting the emergency room (ER) (Rust et al. 2008). This hinders the appropriate management of chronic diseases that could have been effectively treated in a primary care setting. It also exacerbates the problem of long patient wait times in crowded emergency rooms.

Patient–physician continuity is one of the hallmarks of primary care and promotes a long-term relationship between the patient and the physician. Numerous studies have documented the importance of continuity to patients (see, e.g., O'Malley and Cunningham 2008; Atlas et al. 2000). Several studies (Gill and Mainous 1999) show that patients who regularly see their own providers are (1) more satisfied with their care, (2) more likely to take medications correctly, (3) more likely to have problems correctly identified by their physician, and (4) less likely to be hospitalized. The link between lack of continuity and increased ER use has been shown in Gill et al. (2000). Continuity is especially important for vulnerable patients with a complex medical history and mix of medications (Nutting et al. 2003)—patients with long-standing chronic conditions (e.g., diabetes, hypertension, and coronary artery disease). This forms a large percentage of the US population. Continuity is also beneficial for physicians, since their workloads are more focused (O'Hare and Corlett 2004).

How are timely access and patient–physician continuity related to capacity planning and allocation in primary care? When it comes to access to appointments, the two measures are often in conflict. It may be possible for a patient to get a same-day appointment but not necessarily with his/her own physician. Alternatively, a patient may get to see his/her own physician but only weeks later. The two extremes are illustrated in Fig. 8.1. Figure 8.1a shows the situation where patients see only their own physician, while in Fig. 8.1b all physician resources are pooled. In the former, continuity is perfect but timely access may be strained, while the latter results in high levels of timely access but patients may end up seeing unfamiliar physicians.

The focus of this chapter is on capacity planning and allocation for primary care practices at various levels of the planning hierarchy to balance timely access and continuity. For the purposes of this chapter, the term *capacity* refers to the number of appointment slots a physician has in a workday. *Capacity allocation* refers to the assignment of patient requests for an appointment to available appointment slots.

At the strategic level, we consider the impact of size and composition of a physician's panel on the ability to provide timely access and continuity. In primary care, a physician is responsible for the long-term, holistic care of the patients in his/her panel. The size and composition of a physician panel determines a physician's daily appointment burden. The management of panels thus determines capacity planning at the highest level. Next, at the tactical level, we illustrate how a multi-physician practice can manage the inherent flexibility of primary

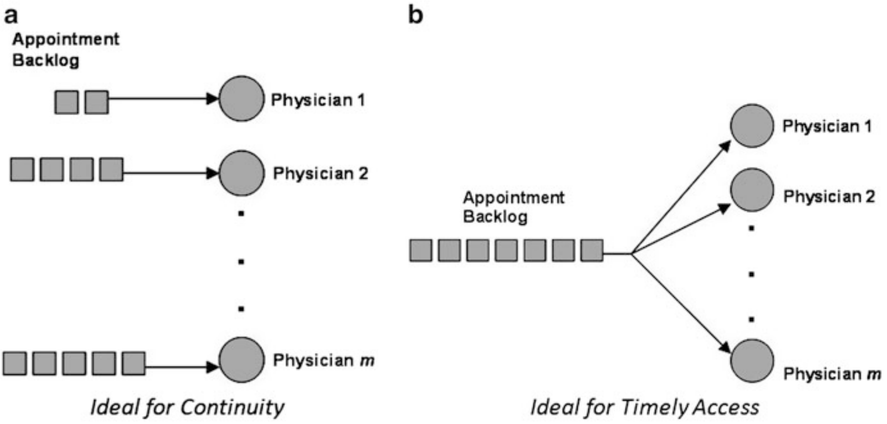


**Fig. 8.1** Dedicated versus pooled physician capacity illustrating the best cases for continuity and timely access, respectively

care physicians to see patients of other physicians to balance timely access and continuity. We study physician flexibility at the tactical level where two uncertain demand streams–*prescheduled* (nonurgent) appointments and *open-access* (same-day) appointments–must share the same capacity. Prescheduled appointments are booked in advance of the workday, while open-access appointments have to be fulfilled the same day. Finally, at the operational level, we also include a discussion of the dynamic context, where decisions about capacity allocation under flexibility have to be made as same-day requests for appointments arrive over the course of the day, that is, under incomplete demand information.

The remainder of this chapter is organized as follows. In Sect. 2, we discuss the literature relevant to capacity allocation in primary care. Section 3 provides the background on capacity planning at the strategic level and the design of physician panels and quantifies the impact of casemix using an example. In Sect. 4, we discuss flexibility in primary care and how it differs from flexibility in manufacturing and other service contexts. In Sect. 5 we provide a modeling framework for testing the impact of flexibility at the tactical level where physician capacity has to accommodate both prescheduled as well as same-day (open-access) appointments. We also provide computational results. In Sect. 6, we provide an outline of the decision framework for using flexibility in the dynamic case (as same-day requests come in over the course of a workday) and discuss other directions for future research. We conclude by summarizing our main findings in Sect. 7.

## 2 Literature Review

The application of operations research to appointment scheduling in healthcare is a growing area of research. We focus here only on the papers most relevant to our research in the primary care context. For a detailed discussion of the impact of primary care access on population health, we point the reader to key references from the health services literature, such as Shi et al. (2005); Macinko et al. (2007, 2003).

Over the last decade the adoption of *open access* (Murray and Berwick 2003), a scheduling policy which urges practices to provide same-day appointments irrespective of the urgency of the request, has brought to the forefront questions regarding appointment system design. What should physician panel sizes be to allow open access? What if patients prefer to have appointments at some future time rather than see a doctor the same day? These questions have necessitated the use of queueing and stochastic optimization approaches that provide guidelines to practices. For instance, Green et al. (2007) investigated the link between panel sizes and the *probability of overflow* or extra work for a physician under open access. They proposed a simple probability model that estimates the number of extra appointments that a physician can be expected to see per day as a function of his/her panel size. The principal message of their work is that for advanced access to work, supply needs to be sufficiently higher than demand to offset the effect of variability. In Green and Savin (2008) a queueing model was used to determine the effect of no-shows on a physician's panel size. They developed analytical queueing expressions that allow the estimation of physician backlog as a function of panel size and no-show rates. In their model, no-show rates increase as the backlog increases; this results in the paradoxical situation where physicians have low utilization even though backlogs are high—this is because patients that had to wait for long do not show up.

In Gupta et al. (2006) results of empirical study of clinics in the Minneapolis metropolitan area that adopted open access are presented. They provided statistics on call volumes, backlogs, and number of visits with own physician (which measures continuity) and discuss options for increasing capacity at the level of the physician and clinic. In Kopach et al. (2007) a discrete event simulation is used to study the effects of clinical characteristics in an open-access scheduling environment on various performance measures such as continuity and overbooking. One of their primary conclusions is that continuity in care is affected adversely as the fraction of patients using open-access increases. The authors mentioned provider groups (or physicians and support staff) working in teams as a solution to the problem. Robinson and Chen (2010) compared the performance of open access with a traditional appointment scheduling system. In the open access system, a practice has to deal with day-to-day variability but very few no-shows, while in the traditional appointment system, patients book their appointments well in advance with the result that day-to-day variability is smoothed but patients have a higher probability of no-show. Their numerical analysis reveals the open access generally outperforms the traditional appointment system when the objective function is a weighted average of patients' waiting time (lead time to appointment, the physician's idle time, and the physicians' overtime). Only when the patient waiting time is held in little regard or when the probability of no-show is small does the traditional system work better than the open-access system. Liu et al. (2010) proposed new heuristic policies for dynamic scheduling of patient appointments under no-shows and cancelations. They find that open access works best when patient load is relatively low.

The papers most closely related to the topic of this chapter are by Qu et al. (2007) and Gupta and Wang (2008). Qu et al. (2007) derived conditions under which a solution for the number of prescheduled appointments to reserve is locally optimal. In Sect. 5, we show a stronger result, guaranteeing global optimality, by first showing that our revenue maximization function has diminishing returns under mild assumptions. Gupta and Wang (2008) explicitly modeled many of the key elements of a primary care clinic. They considered scheduling the workday of a clinic in the presence of (1) multiple physicians; (2) two types of appointments, same-day as well as nonurgent appointments; and (3) patient preferences for a specific slot in a day and also a preference for physicians. The objective is to maximize the clinic's revenue. They use a Markov decision process (MDP) model to obtain booking policies that provide limits on when to accept or deny requests for appointments from patients. In terms of flexibility, their clinic is fully flexible with regard to both nonurgent and urgent appointments. The principal difference between their model and the capacity allocation framework proposed in Sect. 5 is that patient preference drives the scheduling of prescheduled appointments in Gupta and Wang (2008), while we try to balance prescheduled demand and same-day demand through physician flexibility and an explicit consideration of its effect on timely access and continuity.

# 3 Background on Primary Care and the Impact of Casemix

## *3.1 Patient Types*

At the strategic level of the three-part hierarchy defined in the Introduction, a physician builds a panel of patients. The physician's appointment burden depends on the (1) size, and (2) casemix or composition of the panel. A physician working full-time may have 1500–2000 patients. Casemix refers to the type of patients in the panel and can be characterized by various patient attributes, such as age and gender and the chronic conditions afflicting the patient, which play an important role in determining the distribution of visits. For example, a panel where the majority of patients are young and healthy will have a different appointment demand profile compared to a panel consisting mostly of elderly patients with chronic conditions.

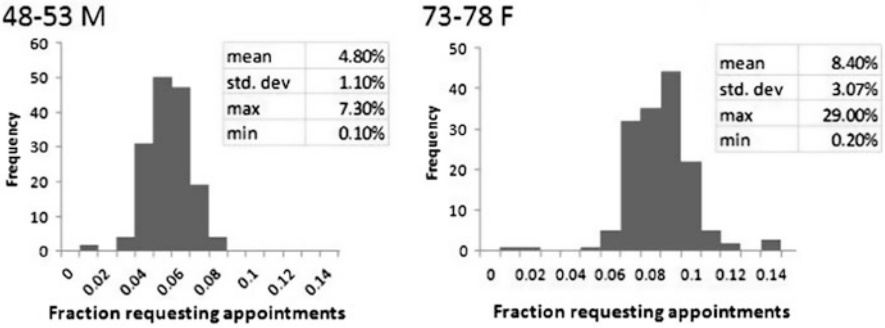


**Fig. 8.2** Histograms of the percentage (or fraction) of total patients requesting appointments in a week for two different patient age and gender categories

Patient classification can be useful for clinics because they enhance a practice's understanding of its population and disease trends and allow it to design its care models effectively. Furthermore, Barbara Starfield's seminal work about ambulatory care groups (ACGs) (Starfield et al. 2007) argued that understanding the role of patient clinical complexity in care utilization forms the cornerstone for effective resource planning and determining payment methods in healthcare.

Age and gender are the simplest patient classification in lieu of more disease-specific data. It has also been found to be generally effective (Murray and Berwick 2003; Balasubramanian et al. 2010). Figure 8.2 shows the distribution of the fraction of total patients requesting appointments in a week for two categories—males (48–53 years old) and females (73–78 years old)—based on historical data from 2004 to 2006 (156 weeks), from the Primary Care Internal Medicine Practice (PCIM) at Mayo Clinic, Rochester, MN. The two distributions show how appointment request rates can vary with gender and age. There are 708 males 48–53 y.o. (48–53 M) and 986 females 73–78 y.o. (73–78 F) empanelled in the practice. 8.4% of all 73–78 F patients request for appointments on average in a week as opposed to 4.8% of all 48–53 M patients. The standard deviation 73–78 F (3.07%) is more than double that of 48–53 M (1.1%).

In Naessens et al. (2011) the authors show that more than an individual chronic condition such as diabetes or hypertension, it is the number of simultaneous chronic conditions (or *comorbidities*) that predicts the consumption of healthcare costs. Furthermore, for primary care physicians, focusing on all comorbidities of a patient is more holistic than focusing in isolation on specific chronic conditions. Figure 8.3 shows mean and standard deviation of visit rates as a function of the number of patients under various counts of comorbidities. The data was simulated based on historical visits of 27,000 patients empanelled in the PCIM practice (Ozen and Balasubramanian 2012). Clearly, not only does the mean number of visits increase with the number of comorbidities; the variance does as well. For instance if a physician has 50 6-comorbidity patients, then he/she will have 450 appointment requests on average each year. If he/she has the same number of 0-comorbidity patients, he/she will have only 75 yearly visits on average.

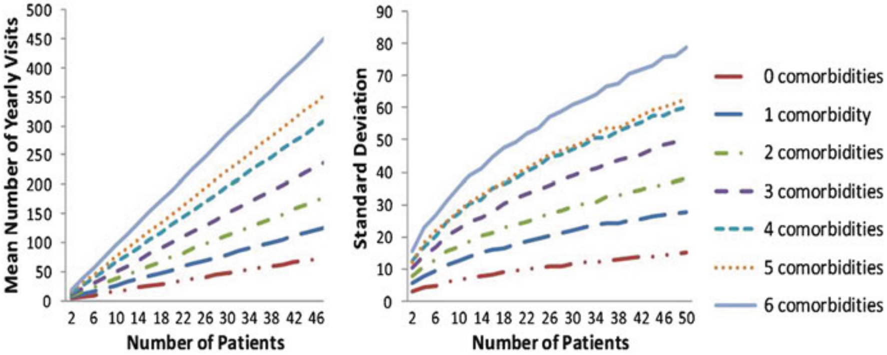


**Fig. 8.3** Mean and standard deviation of yearly visits for groups with different counts of comorbidities

**Table 8.1** Four physicians at PCIM, Mayo Clinic, and their patient casemix based on comorbidity count

| Physicians | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Panel size |
|---|---|---|---|---|---|---|---|---|---|
| Physician 1 | 260 | 249 | 226 | 161 | 108 | 42 | 14 | 3 | 1,063 |
| Physician 2 | 299 | 293 | 212 | 147 | 77 | 26 | 6 | 1 | 1,062 |
| Physician 3 | 214 | 253 | 223 | 177 | 115 | 44 | 21 | 5 | 1,053 |
| Physician 4 | 290 | 296 | 218 | 145 | 84 | 27 | 12 | 5 | 1,077 |

## *3.2 Example: Four Physicians*

We now consider an example of four physicians with approximately the same panel size (1,050 patients), but different casemixes, based on comorbidity counts. These panel compositions are shown in Table 8.1. The casemix can be used to simulate the distribution of daily visits for each physician by sampling for each comorbidity count from historical data. Once the daily visit distribution is obtained, the overflow for a given daily appointment capacity can be calculated. Overflow is simply the fraction of total samples in which the patients' visit requests exceed the available capacity of the physician. Patients that are not seen visit either an unfamiliar physician or an ER or may choose to wait to see the physician on another day. Thus, if overflow is high, both timely access and continuity are adversely affected.

Overflow can also be modeled in the following way. Consider a practice in which there are $J$ physicians and $M$ patient classes. First a practice determines $p_i$, the probability that a patient of class $i = 1,\ldots,M$ will request an appointment on any given day. This can be obtained by calculating the total visits generated by all patients of the class $i$ in the practice over a period of time—for example, two years—and dividing it by the number of unique class $i$ patients as well as the number of workdays in the two-year period. The method is similar to the one proposed in Green et al. (2007). Next, suppose $n_{ij}$ denotes the number of class $i$ patients in physician $j$'s panel. If we assume that each patient requests independently of

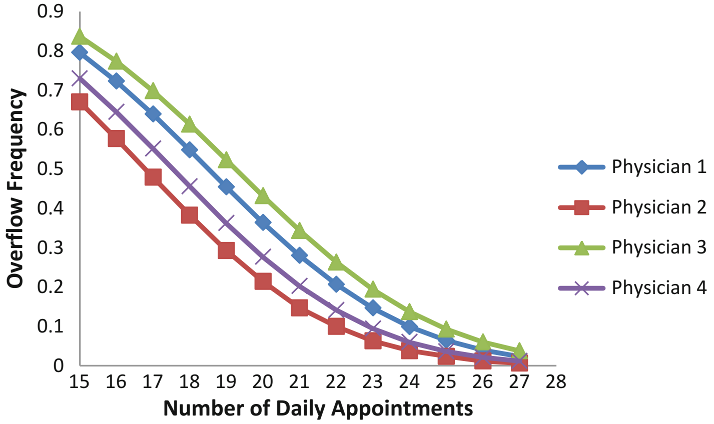


**Fig. 8.4** Overflow for the physicians as a function of the daily capacity (appointment slots)

others, then the total requests from each patient class for each physician can be modeled as a binomial random variable, with mean $n_{ij}p_i$ and variance $n_{ij}p_i(1-p_i)$. Going further, the mean demand for the entire panel is given by $\mu_j = \sum_{i=1}^{M} p_i n_{ij}$ and standard deviation $\sigma_j = \sqrt{\sum_{i=1}^{M} p_i(1-p_i)n_{ij}}$. We will use the normal approximation of a sum of binomial random variables. Then $O_j$, the overflow for physician $j$, is related to the percentile of the standard normal distribution, denoted by $\Phi$, in the following way: $O_j = 1 - \Phi(\frac{C_j - \mu_j}{\sigma_j})$. Here $C_j$ is the capacity of the physician, that is, the total daily slots that he/she has available in a day.

Note that this analysis is at the aggregate level—it does not consider the actual duration of appointments once patients are in the clinic, but tests whether the number of appointment slots (typically 20-min slots) a physician plans to have available in a day is sufficient. It also assumes that all appointments are of the same type. In reality, some appointment requests (such as follow-up appointments) are for a future day, while some are same-day requests. Nevertheless, if overflow is high for all appointments, then it is guaranteed that the timely access for both same-day as well as nonurgent future appointments with one's own PCP will be adversely affected.

The overflow for the four physicians of Table 8.1 as a function of the total capacity (daily appointment slots) is shown in Fig. 8.4. We calculated these overflow profiles using the binomial approximation described above, but it is also possible to obtain the same curves by sampling from historical visit data. For the same capacity, physician 3 and physician 1 have relatively high levels of overflow. This is because there are more patients with two or more comorbidities in their panels (see Table 8.1), and these patient groups generate a higher number of visits. This graph shows that it is inappropriate for clinics to make capacity decisions based solely on panel size. Casemix is also an important consideration. It is also clear that to keep overflow levels down to manageable levels, 20 or more appointment slots may be needed for each of the 4 physicians.

Such analysis allows practices to identify which physicians are overburdened. In the above case, it is clear that physicians 3 and 1 need to have their capacity enhanced—either by working extra hours in a day or by additional nurse practitioner support—or a reduction in the size of their panels. The long-term option for practices is to achieve a better balance among physicians by moving high-demand, high-variability patients from an overburdened physician to a physician with available capacity. More details about the panel redesign approach are presented in Balasubramanian et al. (2010). The paper shows that it is possible to reduce the wait time and the number of redirections to unfamiliar physicians by more than 35%.

The difficulties of redesigning panels are also discussed in Balasubramanian et al. (2010). Reallocating patients abruptly could damage existing patient–physician relationships. Rather the appropriate strategy would be to redesign panels when opportunities present themselves. Many primary care panels are dynamic. Patients enter and leave them all the time as they age, are diagnosed with new conditions, move out of the geographic area, and many other reasons. A useful by-product of this constant state of flux is that it affords continuous opportunities to make incremental changes to patient panels without disrupting the visit patterns of patients who already have strong ties to their PCP. For example, practices can leverage patients who have yet to decide on a PCP, new patients, and the turnover of existing patients. Patient surveys can be used to determine preferences and inclination towards change. In some cases, to minimize disruption, reassignment may simply be to another physician, whom the patient has seen almost as often as his/her own PCP, or to a physician within the same care team (if the care team consists of multiple physicians).

Another viable alternative to panel redesign is carefully managing physicians' ability to see patients of other physicians, depending on whether the requests are urgent/same-day requests or nonurgent requests. This management of physician flexibility forms the content of the next two sections.

## 4 Flexibility in Primary Care

The inherent flexibility of PCPs to see patients from other panels gives practices another lever to provide timely access to care. Using this flexibility, of course, comes at a cost: the resulting loss of continuity when a patient sees unfamiliar physicians. How should practices be designed and managed to use this flexibility to better balance timely access and continuity?

A practice can achieve maximum continuity of care by mandating that patients should see only their own provider. This, however, hampers timely access to care. At the other extreme, a practice may allow patients to see any provider. This is ideal for timely access, but hampers continuity of care. The two extremes are shown in Fig. 8.5a, b. In the first case, the providers are dedicated, while in the second the providers are fully flexible. Figure 8.5c–e show partially flexible configurations that offer a middle ground between (a) and (b). In each of them, a patient sees only one

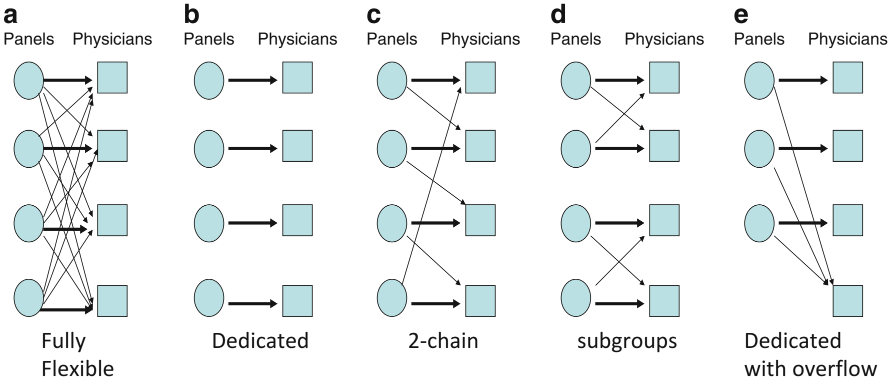


**Fig. 8.5** Figure illustrating different flexibility configurations that trade off continuity of care with timely access

physician other than his/her own PCP. Figure 8.5c is referred to as the 2-chain in the manufacturing flexibility literature (Jordan and Graves 1995) and allows demand variation to be absorbed effectively by the entire practice. In a 2-chain all physicians are directly or indirectly linked to each other, and this property can be exploited to cope with variability in demand. If demand is high for a particular physician, the allocations can be designed such so that the excess demand is shifted to a physician who has capacity available, even though the latter physician may not be directly linked to the former. While the 2-chain is a concept new to healthcare, practices do use the subgroup configuration (Fig. 8.5d). Physicians here may be divided into independent, self-contained teams (such as in the PCIM practice at Mayo Clinic and other academic primary care practices). The dedicated with overflow configuration of Fig. 8.5e is also common—here if the patient's PCP is unavailable, the patients tend to see an *overflow* physician or nurse practitioner (we have observed this setting at a small private practice as well as a community clinic in Western Massachusetts; academic medical centers also use this model).

In the ideal open-access world, there are no appointment types, such as urgent and nonurgent. All appointments are treated identically and scheduled the same day with the patient's PCP. However, the reality is that clinics have a fraction of their schedule available for open-access or urgent appointments. Such appointments, because of the perceived immediacy of need, are typically seen the same day, but not always by the patient's personal physician. The rest of the clinic's schedule consists of appointments booked a week or more in advance. We call these appointments prescheduled appointments. These nonurgent appointments are typically physicals or follow-up appointments for patients with chronic conditions. While the loss of continuity has to be minimized for all appointments, it can be appropriately sacrificed for urgent appointments needing immediate attention by introducing some form of flexibility. We will thus assume that flexibility applies only to urgent appointments. Nonurgent or prescheduled appointments are always seen by the patient's own provider.

We approach the design and management of the flexible practice at two different levels. At the tactical level in primary care, the design of flexibility needs to consider capacity allocation for the two streams of uncertain demand, nonurgent (prescheduled) and urgent (open access), each with different requirements for timeliness and continuity; while at the operational or dynamic level, at which same-day, urgent appointments are booked as patients call over the day, allocation decisions have to be made in real time without full knowledge of demand.

In the remainder of this section, we first summarize the main lessons from the flexibility literature, as they apply to the primary care context, and then consider the tactical and operational cases in detail.

## *4.1 Flexibility Literature Perspective*

Our study of flexibility in primary care practices builds upon the extensive literature on manufacturing flexibility and its more recent application to service systems and worker training and allocation. There are, however, key operational differences that make the application of flexibility to primary care worthy of further analysis: (1) two demand streams associated with each resource, where one (prescheduled demand) is realized before the other (open-access demand); (2) two conflicting objectives, timeliness and continuity of care; (3) no fixed cost associated with installing flexibility, but a loss in continuity for using it; and (4) appointments are booked over time, and thus future resource capacity is sequentially being allocated under partial demand information.

As in the case of cross-training in serial production lines (Hopp et al. 2004), flexibility improves efficiency in two main ways in the primary care environment. The first benefit is in what they refer to as *capacity balancing*: If physician panels are imbalanced with respect to the induced average number of visits to a physician per day, flexibility will allow the load to be shared between physicians, therefore improving overall timeliness of care and physician utilization. The second is in *variability buffering*: Even if the average workloads are balanced between physicians, variability in patient requests for a particular day/time will be better accommodated by a flexible environment. Hopp et al. (2004) compares a strategy that balances capacity using the minimum amount of cross-training with the chaining of skills in the sequence of the serial line. They find that skill-chaining strategies are more robust and more effective in variability buffering. The concept of chaining has received much attention since it was first introduced in the seminal work of Jordan and Graves (1995). In a single-period, multiproduct, multiplant production network, they show that the 2-chain (Fig. 8.2c) results in increased sales and capacity utilization, relative to the dedicated configuration (Fig. 8.2a), comparable to those achieved by a fully flexible system (Fig. 8.2b). That is, a few links, configured in the right way (2-chain), provide almost the same performance as the complete, fully flexible network. Furthermore, this strategic analysis has been extended recently to multistage supply chains (Graves and Tomlin 2003) and to a made-to-order environment where flexibility is also used to hedge against operational variability (Muriel et al. 2006). Chou et al. (2010) distinguishes between range (the different demand scenarios that can be accommodated) and response (the cost of doing so; that is, the cost of using secondary rather than primary resources for production/service) of flexible systems. They show that upgrading system response (i.e., building systems where physicians can handle other physician's panels at lower additional cost) outperforms improving system range (creating systems that can accommodate ever more extreme patient demand scenarios). This result suggests that in the primary care setting, the benefits of restricting the number of doctors that can see a particular patient (resulting in lower cost of service because of familiarity and thus increased response) are likely to outweigh the higher range provided by a fully flexible team care practice where any doctor can see the patient.

A number of computational reports in the literature (e.g., Jordan and Graves 1995) point out an increase in the marginal benefit associated with adding one more flexibility link (i.e., allowing one more panel to see a second physician) in forming the 2-chain, culminating with a markedly higher increase when the last link that closes the chain is put in place. Recently, Simchi-Levi and Wei (2012) prove that is always the case, and show that long chains are always superior to any other strategy where each product (panel) can be produced at two plants (can be assigned to two physicians). This suggests that the larger practices will benefit most by managing their inherent flexibility in the form of a long chain.

## 5 Example of Flexibility in Primary Care

In this section we provide a specific example of a model for evaluating the influence of flexibility in a primary care practice. We focus at the tactical level. We provide theoretical results relevant to the model and numerical experiments that illustrate the relative benefits of flexibility.

The basic question addressed by the model is how much of the physician's total daily workload should be dedicated to prescheduled versus urgent appointments. This will depend on how much flexibility the practice allows when allocating urgent patient demand. We thus need to address this question under different flexibility configurations, as illustrated in Fig. 8.2. This will also allow us to compare their resulting performance in terms of system revenue, continuity, and timely access. For that purpose, we develop a two-stage stochastic integer program that can accommodate any flexibility configuration and greedy, but exact, algorithms to quickly calculate the optimal capacity allocations in dedicated and fully flexible systems. The analytical and experimental results and conclusions summarized here are from Balasubramanian et al. (2010).

**Two-stage capacity allocation model**: We solve the capacity allocation problem for a single workday using a two-stage stochastic integer programming model.

We consider a general primary care practice with $M$ physicians, indexed by $i = 1,2,\ldots,M$, each with $N_i$ available appointment slots. Let $A$ be the set of all possible panel–physician links $(i, j)$ such that the open-access (same-day) requests of patients in panel $i$ (i.e., physician $i$'s panel) can be served by physician $j$. The set $A$ represents the particular flexibility configuration under consideration. Let $R_i^{\mathrm{p}}$ be the revenue associated with physician $i$ seeing one of his/her prescheduled patients and $R_{ij}^{\mathrm{o}}$ be the revenue associated with physician $j$ seeing an open-access patient of panel $i$, for any $(i, j) \in A$. The demand for prescheduled and open-access appointments is represented by a random vector $D = (D_1^{\mathrm{p}}, D_1^{\mathrm{o}}, \ldots, D_M^{\mathrm{p}}, D_M^{\mathrm{o}})$ where the superscript $p$ refers to prescheduled and $o$ to open access and the subscript indicates the primary care physician. Vector $D$ follows a discrete distribution that assigns a probability $q_s$ to each possible realization of demand, indexed by $s$, $s = 1,2,\ldots,S$. That is, $P[D = (d_{1s}^{\mathrm{p}}, d_{1s})^{\mathrm{o}}, \ldots, d_{Ms}^{\mathrm{p}}, d_{Ms}^{\mathrm{o}})] = q_s$. We introduce the following capacity allocation variables:

$N_i^{\mathrm{p}}$: Number of slots allocated for prescheduled demand of physician $i$.
$x_{is}^{\mathrm{p}}$: Number of patients allocated to physician $i$ under demand realization $s$.
$x_{ijs}^{\mathrm{o}}$: Number of open-access patients of panel $i$ assigned to physician $j$ under demand realization $s$, for all $i = 1,2,\ldots,M$ and $(i, j) \in A$.

The objective is to maximize the expected revenue of satisfying prescheduled and open-access appointments. We use binary decision variables to capture whether the prescheduled demand for a physician is less or greater than the corresponding $N_i^{\mathrm{p}}$ value. Inequality (8.3) ensures that $\phi_{iu_{is}} = 1$ if $d_{is}^{\mathrm{p}} < N_i^{\mathrm{p}}$. Inequality (8.4) ensures that $\phi_{iu_{is}} = 0$ if $d_{is}^{\mathrm{p}} > N_i^{\mathrm{p}}$. Equations (8.5) and (8.6) limit the number of prescheduled appointments to the allocated capacity and the realized demand, respectively. Inequalities (8.7) and (8.8) ensure that the total open-access appointments for any physician $j$ do not exceed remaining capacity, when $\phi_{iu_{is}} = 1$ and $\phi_{iu_{is}} = 0$, respectively. Inequality (8.9) limits the total number of open-access appointments scheduled from a panel to the realized demand for such appointments from that panel. Inequality (8.10) is the binary constraint.

$$(SIP) \quad \max \left\{ \sum_{s=1}^{S} \sum_{i=1}^{m} q_s \left\{ R_i^{\mathrm{p}} x_{is}^{\mathrm{p}} + \sum_{\{i,j\} \in A} R_{ij}^{\mathrm{o}} x_{ijs}^{\mathrm{o}} \right\} \right\} \tag{8.1}$$

s.t.

$$N_i^{\mathrm{p}} \le N_i \quad \forall i \tag{8.2}$$

$$N_i^{\mathrm{p}} \le d_{is}^{\mathrm{p}} + N_i \phi_{iu_{is}}, \quad \forall (i,s) \tag{8.3}$$

$$N_i^{\mathrm{p}} \ge d_{is}^{\mathrm{p}} \phi_{iu_{is}}, \quad \forall (i,s) \tag{8.4}$$

$$x_{is}^{\mathrm{p}} \le N_i^{\mathrm{p}} \quad \forall (i,s) \tag{8.5}$$

$$x_{is}^{\mathrm{p}} \le d_{is}^{\mathrm{p}} \quad \forall (i,s) \tag{8.6}$$

$$\sum_{i:(i,j) \in A} x_{ijs}^{\mathrm{o}} \le N_j - d_{js}^{\mathrm{p}} \phi_{ju_{js}} \quad \forall (j,s) \tag{8.7}$$

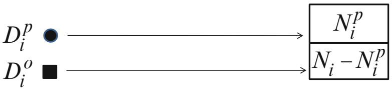


**Fig. 8.6** System configuration for a dedicated practice

$$\sum_{i:(i,j)\in A} x^{\mathrm{o}}_{ijs} \leq N_j - N^{\mathrm{p}}_j + \phi_{ju_{js}} N_j \quad \forall (j,s) \tag{8.8}$$

$$\sum_{i:(i,j)\in A} x^{\mathrm{o}}_{ijs} \leq d^{\mathrm{o}}_{is}, \quad \forall (i,s) \tag{8.9}$$

$$\phi_{iu_{is}} \in (0,1), \forall i, \quad \forall u_{is} = 0,1,\ldots,N_i \tag{8.10}$$

$$N^{\mathrm{p}}_i, x^{\mathrm{p}}_{is}, x^{\mathrm{o}}_{ijs} \geq 0, \quad \forall (i,j) : (i,j) \in A, \quad \forall s \tag{8.11}$$

We note in the above revenue optimization that $R^{\mathrm{o}}_{ij} > R^{\mathrm{p}}_{ij}$. This is because in a relative sense, open-access appointments are more valuable than prescheduled appointments, as explained in Balasubramanian et al. (2012). First, we note that open-access appointments, because they have such short lead times, tend to have much lower no-show rates. Second, if a prescheduled appointment results in a no-show, it can be substituted by an open-access appointment, while the reverse is not possible at such a short notice. Third, because prescheduled appointments are made generally a week or more in advance, the patient is likely to be flexible about choice of the appointment day, and thus this may result in postponed but not lost demand if denied timely access. An open-access patient, on the other hand, needs to see a physician immediately and hence is flexible in provider choice.

In the next two sections, we present analytical solutions to the capacity allocation problem for dedicated practices, where physicians can only see patients in their own panel, and fully flexible practices where open-access patients can be seen by any of the physicians in the practice. For large practices using partial flexibility such as the 2-chain configuration, unfortunately, the above stochastic program is too large to solve efficiently in practice. While the number of binary and integer variables is quite manageable, the sheer number of possible demand realizations makes the problem intractable. To overcome this issue, we will solve the problem using a computationally effective sample average approximation method proposed by Solak et al. (2010) for two-stage stochastic integer programming problems (see Sect. 5.1).

**The Dedicated Case**: In a dedicated practice, physicians can only serve the prescheduled and open-access patients from their own panel. They need to decide, however, a maximum number of appointment slots to make available to prescheduled patients, $N^{\mathrm{p}}_i$, so that enough capacity is reserved for the more lucrative open access ones. The system configuration is shown in Fig. 8.6.

Let $E[R_i(N^{\mathrm{p}}_i)]$ be the total expected revenue from the panel of physician $i$, as a function of $N^{\mathrm{p}}_i \in 0,1,2,\ldots,N$. Our goal is to find the optimal value of $N^{\mathrm{p}}_i$. The conditions for local optimality presented in Qu et al. (2007) for the problem of maximizing the expected number of patients consulted in a single-physician practice can be easily adapted to our revenue maximizing objective. In Balasubramanian et al. (2012), we show a stronger result, guaranteeing global optimality, by first showing that the objective function has diminishing returns under mild assumptions.

**Proposition 1.** *If $P[D_i^{\mathrm{o}} \geq N_i - N_i^{\mathrm{p}} | D_i^{\mathrm{p}} \geq N_i^{\mathrm{p}}]$ is nondecreasing in $N_i^{\mathrm{p}}$, the difference in revenue associated with increasing the number of prescheduled slots by one, $E[R_i(N_i^{\mathrm{p}}+1)] - E[R_i(N_i^{\mathrm{p}})]$, is nonincreasing in $N_i^{\mathrm{p}}$.*

The above condition holds when the demand for prescheduled and open-access appointments is independent of each other. Furthermore, it will be satisfied in most practical scenarios. Intuitively, for it to be violated, the probability of open-access demand being large would need to significantly decrease as the demand for prescheduled appointments grows; that is, the demand for open-access and prescheduled appointments would need to be heavily negatively correlated.

As a result of Proposition 1, we have that the expected revenue function exhibits diminishing returns, an analog of concavity for a discrete function, and thus its global maximum must occur at the largest integer $N_i^{\mathrm{p}} \leq N$ such that $E[R_i(N_i^{\mathrm{p}})] - E[R_i(N_i^{\mathrm{p}} - 1)] \geq 0$ leading to the following theorem.

**Theorem 1.** *If $P[D_i^{\mathrm{o}} \geq N_i - N_i^{\mathrm{p}} | D_i^{\mathrm{p}} \geq N_i^{\mathrm{p}}]$ is nondecreasing in $N_i^{\mathrm{p}}$, the optimal solution to the dedicated problem is the largest nonnegative integer $N_i^{\mathrm{p}} \leq N$ such that $P[D_i^{\mathrm{o}} \geq N_i - N_i^{\mathrm{p}} | D_i^{\mathrm{p}} \geq N_i^{\mathrm{p}}] \leq R_i^{\mathrm{p}}/R_i^{\mathrm{o}}$.*

The optimal solution can thus be easily obtained by calculating that probability starting at $N_i^{\mathrm{p}} = 0$ and increasing one unit at a time until it exceeds the threshold $R_i^{\mathrm{p}}/R_i^{\mathrm{o}}$. A binary search could also be used. Observe that in the case of independent open-access and prescheduled demands, the optimal value of $N_i^{\mathrm{p}}$ does not depend on the distribution of prescheduled demand for physician $i$.

**The Fully Flexible Case**: In a fully flexible practice, open-access patients can be seen by any available physician. In this case, the optimal number of slots to make available to prescheduled demand of the physicians, $N_1^{\mathrm{p}*}$ and $N_2^{\mathrm{p}*}$ in the case of two physicians, can still be found with a simple greedy algorithm. This is because the revenue function again exhibits diminishing returns as the number of slots offered to prescheduled patients is increased. For ease of exposition, we assume that all physicians have the same capacity of $N$ slots and that the revenue of an open-access appointment is identical for all physicians and panels and denoted by $R^{\mathrm{o}}$. We first consider the case of two physicians, $i$ and $j$. See Fig. 8.7.

**Proposition 2.** *If $P[D_i^{\mathrm{o}} + D_j^{\mathrm{o}} > 2N - (N_i^{\mathrm{p}} + \min(D_j^{\mathrm{p}}, N_j^{\mathrm{p}}) + 1) | D_i^{\mathrm{p}} \geq N_i^{\mathrm{p}} + 1]$ is nondecreasing in $N_i^{\mathrm{p}}$ and $N_j^{\mathrm{p}}$, the difference in revenue associated with increasing the number of prescheduled slots of physician $i$ by one, $E[R_i(N_i^{\mathrm{p}}+1, N_j^{\mathrm{p}}]) - E[R_i(N_i^{\mathrm{p}}, N_j^{\mathrm{p}}])$, is nonincreasing in $N_i^{\mathrm{p}}$ and $N_j^{\mathrm{p}}$.*

Observe that, as in the dedicated case, the conditions will hold when open-access and prescheduled demands are independent and in any practical scenario except for contrived cases where the demands for prescheduled and open-access appointments are significantly negatively correlated. Since the revenue function

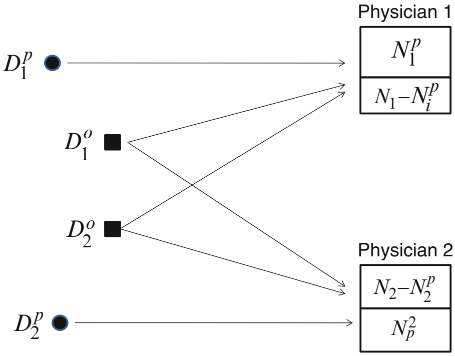


**Fig. 8.7** System configuration for a fully flexible practice

exhibits decreasing returns in both $N_i^{\mathrm{p}}$ and $N_j^{\mathrm{p}}$ under those mild conditions, which can be interpreted as concavity of the discrete revenue function, a greedy algorithm that keeps increasing one appointment slot at a time to the physician where it produces the highest system revenue will provide the optimal capacity allocation scheme.

Proposition 2, and therefore the optimality of a greedy algorithm, can be easily extended to the general case of $M$ physicians that fully share open-access demand. Full details of the theorems and the proofs are given in Balasubramanian et al. (2012).

## *5.1 Computational Experiments*

The exact greedy algorithms allow us to find the optimal capacity allocation and system revenues for dedicated and fully flexible practices. To test the performance of partial flexibility configurations (see Fig. 8.5), which promote continuity by restricting the number of doctors that a patient can be assigned to, we use the two-stage stochastic integer program (SIP). In what follows, we present a summary of the results emphasizing (1) the value of the 2-chain to improve open access while keeping acceptable levels of continuity and (2) how the optimal portion of clinic capacity reserved for open access changes as more flexibility is allowed when allocating open-access demand; for full details, please see Balasubramanian et al. (2012).

**Value of Partial Flexibility** : Following the findings of Bennett and Baxley (2009), we assume a typical no-show rate for prescheduled demand of 25% and a 10% no-show rate for open-access demand. Thus, an appointment slot given to an open access patient brings higher expected revenue, 0.9, as compared to revenue of 0.75 for scheduling one prescheduled patient. To encourage continuity in the system, we assume that there is a 0.05 cost of seeing patients from another physician's panel (the revenue of giving an appointment slot to one open-access patient not from a physician's panel is therefore $0.9-0.05=0.85$). While the no-show rates

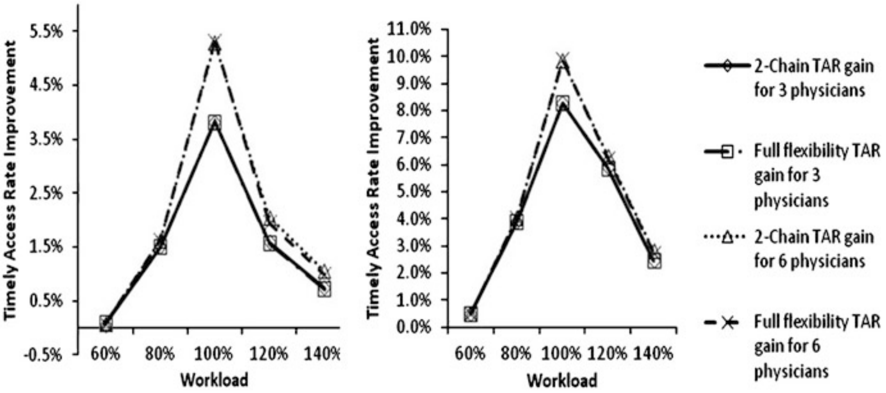


**Fig. 8.8** Comparison of timely access rate (TAR) improvement between 3 and 6 physicians in the symmetric (*left figure*) and asymmetric (*right figure*) cases

for the two types of appointments can be estimated from past data, the cost of diverting an open-access patient to a non-PCP physician is very difficult to quantify. Furthermore, in a limited flexibility environment, where the patient only sees at most one physician beyond his/her PCP, the actual cost of redirection is minimal, very different from that occurring in a large, fully flexible practice where care is significantly more fragmented and much harder to coordinate. For that reason, rather than comparing the expected revenues obtained under the different configurations, we focus here on the resulting timely access rates (TAR). We define TAR as the percentage of all patients, both prescheduled and open access, who get access to an appointment on the given workday.

Figure 8.8 shows the gains, relative to a dedicated practice, of implementing partial flexibility (configured as a 2-chain) and full flexibility to share open-access demand as system workload increases in practices with 3 and 6 physicians. Workload is defined as the ratio of the expected total demand for the clinic and total available capacity. Each physician has 24 appointment slots available in the day. The left graph, or symmetric case, involves a practice where all physicians face identical panel demand distributions (Poisson demands with a rate of 10 for prescheduled appointments and 14 for open-access demand). The right graph, or asymmetric case, considers a practice where physicians have varying panel compositions and therefore varying appointment burdens. This is common in practice. Senior and well-established physicians may have higher workloads since their panels are larger and include older, more complex patients, while physicians who have been recently hired may have lower workloads. In particular, we test a practice where physician 1 has an expected prescheduled demand of 6 and an expected open-access demand of 12 (low workload), physician 2 has an expected prescheduled demand of 8 and an expected open-access demand of 16 (balanced or full workload), and physician 3 has an expected prescheduled demand of 10 and an expected open-access demand of 20 (high workload). For the six-physician case, we merely double the 3-physician case, thus retaining the imbalances.

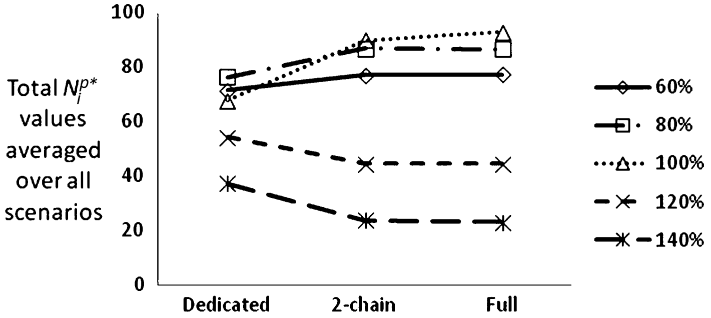


**Fig. 8.9** Trends in $N_i^{p*}$ values summed over all six physicians with prescheduled demands [6,8,10,6,8,10] and open-access demands [12,16,20,12,16,20]

The timely access rates of 2-chain flexibility and full flexibility are nearly the same no matter what the size and workload level of the system are. This is consistent with the results reported in the literature on flexibility in manufacturing settings. The difference is even lower in the healthcare setting that forms our test case, since we assume that prescheduled demand cannot be shared between physicians; flexibility can only be used for open-access demand. We also observe, as in the manufacturing literature, that the gains accrued through flexibility increase significantly as (1) the number of physicians increases from 3 to 6 and (2) the physicians have different workloads, that is, in the asymmetric case, when flexibility helps not only to accommodate demand variability but also to balance physician workloads.

These results suggest that flexibility provides an important lever for practices to increase their ability to accommodate open-access demand. Furthermore, the 2-chain configuration allows them to do so without severely compromising continuity and patient/physician bonds.

**Capacity Allocation** : The results above illustrate the value of flexibility. But how are capacity allocation decisions affected by the flexibility configuration used? What trends do they follow, if at all, and can the trends provide clues to capacity allocation decisions in practice? In our model, the capacity allocation is decided with the optimal first-stage variables, $N_i^{p*}$, which represent the capacity made available to prescheduled appointments. Figure 12 shows the average values for the entire clinic (i.e., for all the physicians) under different workloads and the three flexibility configurations for the 6-physician asymmetric case. We see the same trends by looking at the individual physicians' values (irrespective of the number of physicians, symmetry, and prescheduled to open-access demand ratios). Thus, Fig. 8.9 summarizes our conclusions about $N_i^{p*}$ values concisely.

In general, for the case of very low system workload, the total $N_i^{p*}$ values for the dedicated and flexibility configurations, not surprisingly, are very close. Since the demands are so low, the values are likely to be fairly robust at this level. As the system or clinic workload increases to 80% and 100%, the clinic as a whole reserves more prescheduled appointments in the flexibility cases than the dedicated case. This is a direct consequence of flexibility: Open-access appointments can be absorbed effectively by pooling the (lower) remaining capacity of all physicians together. The effect is especially strong in the case of 100% workload: The dedicated case increases the capacity reserved for the more profitable and now more abundant open-access patients $(N - N_p)$ relative to the lower workload cases, while the flexible configurations decrease it to allow for more of the now plentiful prescheduled patients and still meet open-access demand through sharing any unused capacity.

In contrast, in the high system workload cases (120% and 140%), there is enough demand for the high revenue open-access appointments to lower the total of the clinic. The flexibility cases have a lower total $N_i^{p*}$ value than the dedicated case, reserving more capacity for open access. This is because there is a higher probability of using the additional capacity when physicians are able to see each others' open-access appointments.

Thus, using the easily computable dedicated case $N_i^{p*}$ as a reference, practices can heuristically determine their capacity allocation to be above or below the dedicated value, depending on their flexibility configuration and overall system workload.

# 6 Future Research Opportunities

In this section, we first outline the decision framework for using flexibility in the dynamic case, report our preliminary findings, and then discuss other directions for future work.

## *6.1 Flexibility in the Dynamic Case*

The discussion of flexibility so far assumed that demands are instantly realized and fulfilled. In practice, however, allocation decisions for open-access and same-day appointments have to be made without full realization of demand. At the beginning of the day, all nonurgent appointments scheduled on physician calendars are known in advance, but calls for same-day appointments come throughout the day and have to be dynamically assigned to available physician slots. As before, the challenge is to balance timely access (minimize the number of denied same-day appointment requests) with continuity (ensuring patients see their own physician as much as possible).

Consider again a clinic with $M$ physicians, each with a panel of patients for which he/she is the primary care physician (PCP). Depending on the flexibility configuration, patients will be allowed to see one or more physicians. The time horizon begins when the clinic opens its open-access appointments and ends when the clinic stops taking further appointments. This will typically mean the entire

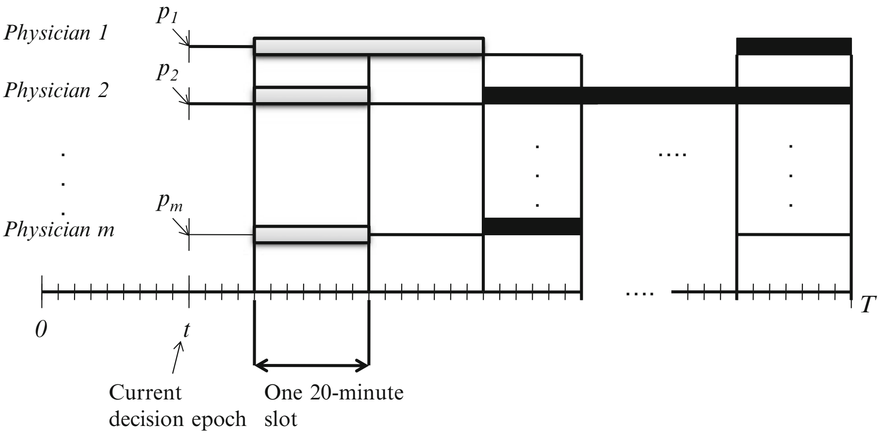


**Fig. 8.10** Visual illustration of the model at decision epoch $t$. The *dark slots* correspond to prescheduled appointments; the *gray slots* to already booked same-day appointments. Slots that are *not shaded* are available to be assigned

duration of a day (7–8 h). Calls for a given physician's slots come with a certain probability ($p_i$ for physician $i$) at every time point during the horizon. Each physician's calendar consists of successive 20-min slots that are booked as calls come in during the day. In the ideal situation, all open-access appointments are contiguous and occur during the same time of the workday. In practice, because of patient preferences for time slots, available same-day slots will be interspersed with prescheduled slots. The situation is shown in Fig. 8.10.

The decision framework can be modeled in a finite-horizon stochastic dynamic program. For the mathematical details, see Hippchen (2009). At each time point $t$, if there is a request from physician panel $i$, the decision facing the clinic is whether this request should be (1) assigned to his/her PCP, (2) to some other physician (as allowed by the flexibility configuration), or (3) denied. If there are no requests at time $t$ for any of the physicians, then the action is to "do nothing." The optimization problem is to choose the best action at each decision epoch to minimize total cost of denied requests and missed continuity (measured by the number of non-PCP diversions) for the day. The state of the system at any decision epoch $t$ is represented by the number of open-access patients booked in the future in each physician's calendar. Denying a request incurs a cost—denied requests are a reflection of the lack of timely access to primary care or the costs needed to provide care to these patients outside the regular hours of clinic operations.

What impact does flexibility have in the operational or dynamic case? Recall that in the tactical case, flexibility was beneficial in hedging against variability of demand. In the dynamic case, there is an additional component of variability, since appointment requests arrive randomly over time. There is therefore greater opportunity for flexibility to meet demand imbalances at different points in time. On the other hand, patient calls require an immediate appointment allocation decision, under only partial demand information available at that point; this decreases the impact of flexibility, since allocation decisions that can be made optimally in the tactical case may not be as effective in the dynamic case. These counteracting effects may be the reason why the benefits of flexibility are mostly identical in both the tactical and dynamic cases. Our computational experience with the stochastic dynamic program (Hippchen 2009) shows that the benefits of full and partial flexibility in the dynamic case produce the same percentage improvements in timely access rate shown in Fig. 8.8.

While the stochastic dynamic program has been used to illustrate the impact of flexibility, primary care offices require easily implementable policies or heuristics that can be put into practice as calls for same-day appointments come in. Consider two contrasting policies in the fully flexible case: primary first (PF) and most slots (MS). PF assigns incoming same-day calls to the patient's PCP first, so long as slots are available. If PCP slots are not available for the day, it assigns the patient to the physician with the most available slots. MS, on the other hand, assigns an incoming patient call to the physician with the most slots available.

PF thus maximizes continuity, while MS utilizes physician slots more effectively and increases the number of patients seen per day. A hybrid approach that balances continuity with timely access would be assigned to PCP so long as the difference the available slots between PCP and other physicians does not exceed a certain predetermined threshold. Our results show that the choice of heuristic also depends on the system workload or utilization. PF is the best choice in an underutilized as well as overutilized setting, while MS and the hybrid approach are better choices in systems where arrival rates and available capacity are relatively balanced.

## *6.2 Other Research Directions*

The dynamic case discussed above considers flexibility at the level of a workday. However, there is one further level that we have not considered in this chapter. This pertains to *patient flow* measures such as in clinic patient waiting, idle time, and overtime for the clinicians. The relevant questions here are how many patients can be seen per day and how the appointments should be spaced and sequenced, and where overbooking is appropriate in the workday to counter no-shows. The decisions—from panel design to using physician flexibility—discussed in this chapter are at the highest planning level. Integrating these decisions and quantifying their impact on patient flow are a natural direction for future work. How do panel size, case-mix and the degree of continuity provided impact the time spent with clinicians? How robust are the tactical level decisions (such as the number of same-day patients clinics should plan for) to changes at the patient flow level? Simulation–optimization approaches can be used to answer these questions.

Another important research direction is capturing patient flow measures empirically and making such data sets (or prototypes of such data sets) more widely available. While most of the operations research literature focuses on new models, few papers examine the validity of these models in practice. Rigorous time studies, reconstructions of patient flow through time-stamp data, and the use of real-time location systems (RTLS) are all methods of capturing the operational aspects of patient flow.

## 7 Summary and Conclusions

In summary, we have discussed capacity allocation for primary care practices at three different levels of the capacity planning hierarchy. The goal in each case has been to maximize timely access and continuity. At the highest level, the design of physician panels, we demonstrated the impact of casemix, or the type of patients in a physician's panel, on the ability to provide timely access and continuity. Casemix can be considered using age and gender as predictors or, when patient clinical data is available, using comorbidity counts. Using casemix, a practice can create overflow profiles for the physicians in the practice as function of daily capacity and determine which physicians are overburdened. This in turn can point to opportunities for redesigning panels so that patients can see their own PCP as much as possible and redirections to unfamiliar physicians are minimized.

Panel redesign involves changing existing patient–physician relationships. A viable alternative to redesign is managing the flexibility of physicians—the ability of physicians to see patients of other physicians. We discussed flexibility at tactical as well as operational levels. The design of flexibility at the tactical level has to consider two types of appointments: (1) prescheduled appointments which are booked in advance and require continuity with the patient's PCP, and (2) same-day or open-access appointments which have to be fulfilled during the course of the day. We proposed a framework—commonly observed in practice—in which the short notice open-access appointments can be flexibly shared between physicians while mandating continuity for the prescheduled appointments. Using a two-stage stochastic integer programming model, we demonstrated the impact of flexibility on the ability to provide timely access to patients, measured by the number of patients seen in a given workday. Specifically, we found that the 2-chain partially flexible practice, which restricts the number of physicians a patient sees to two, performs almost as well as the fully flexible practice with regard to timely access. The impact of flexibility increases as the number of physicians in the practice increases and as the demand loads between physicians are asymmetric or uneven. Our results also show that practices can heuristically determine their capacity allocation for prescheduled appointments depending on their flexibility configuration and overall system workload.

Finally, the implementation of flexibility at the level of a workday has to be made under partial demand information, since calls arrive dynamically over the course of a day. We outline a decision framework to evaluate the impact of flexibility in this dynamic case and discuss heuristics that practices can use to balance timely access and continuity.

**Acknowledgements** This work was funded in part by the grant CMMI 1031550 from the National Science Foundation and grant R03 HS 018795 from the Agency of Healthcare Research and Quality (AHRQ). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF or AHRQ.

## References

American College of Physicians (2006) The impending collapse of primary care and its implications for the state of the nation's healthcare, Technical report. Available via: http://www.acponline.org/advocacy/events/state_of_healthcare/statehc06_1.pdf

Atlas S, Grant R, Ferris T, Chang Y, Barry M (2000) Patient–physician connectedness and quality of primary care. Ann Intern Med 150(5):325–226

Balasubramanian H, Banerjee R, Denton B, Naessens J, Wood D, Stahl J (2010) Improving clinical access and continuity using physician panel redesign. J Gen Intern Med 25(10):1109–1115

Balasubramanian H, Denton B, Lin M (2011) Managing physician panels in primary care. In: Yih Y (ed) Handbook of healthcare delivery systems, 10–1. CRC, West Palm Beach (Taylor and Francis)

Balasubramanian H, Muriel A, Wang L (2012) The impact of provider flexibility and capacity allocation on the performance of primary care practices. Flex Serv Manuf J 24(4):422–447

Bennett KJ, Baxley EG (2009) The effect of a carve-out advanced access scheduling system on no-show rates. Pract Manag Fam Med 41(1):51–56

Chou MC, Chua GA, Teo C (2010) On range and response: dimensions of process flexibility. Eur J Oper Res, Elsevier 207(2):711–724

Gill JM, Mainous A (1999) The role of provider continuity in preventing hospitalizations. Arch Fam Med 7:352–357

Gill JM, Mainous A, Nsereko M (2000) The effect of continuity of care on emergency department use. Arch Fam Med 9:333–338

Graves SC, Tomlin BT (2003) Process flexibility in supply chains. Manag Sci 49(7):907–919

Green LV, Savin S (2008) Reducing delays for medical appointments: a queueing approach. Oper Res 56(6):1526–1538

Green LV, Savin S, Murray M (2007) Providing timely access to care: what is the right patient panel size? Joint Comm J Qual Patient Saf 33:211–218

Gupta D, Wang L (2008) Revenue management for a primary-care clinic in the presence of patient choice. Oper Res 56(3):576–592

Gupta D, Potthoff S, Blowers D, Corlett J (2006) Performance metrics for advanced access. J Healthcare Manag 51(4):246–259

Hippchen J (2009) Flexibility in primary care. Masters thesis. (Advisors: Hari Balasubramanian and Ana Muriel). Accessible at: http://people.umass.edu/hbalasub/FlexibilityThesis.pdf

Hopp W, Tekin E, Van Oyten MP (2004) Benefits of skill chaining in serial production lines with cross-trained workers. Manag Sci 50(1):83–98

Jordan WC, Graves SC (1995) Principles and benefits of manufacturing process flexibility. Manag Sci 41(4):577–594

Kopach R, DeLaurentis P, Lawley M, Muthuraman K, Ozsen L, Rardin R, Wan H, Intrevado P, Qu X, Willis D (2007) Effects of clinical characteristics on successful open access scheduling. Health Care Manag Sci 10:111–124

Liu N, Ziya S, Kulkarni V (2010) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. Manuf Serv Oper Manag 12.2:347–365

Macinko J, Starfield B, Shi L (2003) The contribution of primary care systems to health outcomes within organization for economic cooperation and development (OECD) countries. Health Serv. Res. 38(3):831–865

Macinko J, Starfield B, Shi L (2007) Quantifying the health benefits of primary care physician supply in the United States. Int J Health Serv 37(1):111–126

Muriel A, Somasundaram A, Zhang Y (2006) Impact of partial manufacturing flexibility on production variability. Manuf Serv Oper Manag 8(2):192–205

Murray M, Berwick DM (2003) Advanced access: reducing waiting and delays in primary care. J Am Med Assoc 289(8):1035–1040

Naessens J, Stroebel R, Finnie D, Shah N, Wagie A, Litchy W, Killinger P, O'Byrne T, Wood D, Nesse R (2011) Effect of multiple chronic conditions among working-age adults. Am J Manag Care 17(2):118–122

Nutting P, Goodwin MA, Flocke S, Zyzanski S, Kurt C (2003) Continuity of primary care: to whom does it matter and when? Ann Fam Med 1:149–155

O'Hare CD, Corlett J (2004) The outcomes of open-access scheduling. Family practice management in Available via: http://journals.dev.aafp.org/XML\discretionary-journal-files/fpm/2004/0200/.svn/text\discretionary-base/fpm20040200p35.pdf.svn\discretionary-base. Accessed on Sept 2011

O'Malley AS, Cunningham PJ (2008) Patient experiences with coordination of care: the benefit of continuity and primary care physician as referral resource. J Gen Intern Med 24(2):170–177

Ozen A, Balasubramanian H (2012) The impact of case mix on timely access to appointments for a primary care group practice. Health Care Manag Sci: DOI 10.1007/s10729-012-9214-y. Accessed on sept 2011

Qu X, Rardin R, Williams JAS, Willis D (2007) Matching daily healthcare provider capacity to demand in advanced access scheduling systems. Eur J Oper Res 183(2):812–826

Robinson L, Chen R (2010) A comparison of traditional and open access policies for appointment scheduling. Manuf Serv Oper Manag 12.2:330–347

Rust G, Ye J, Baltrus P, Daniels E, Adesunloye B, Fryer GE (2008) Practical barriers to timely primary care access. Arch Int Med 268(15):1705–1710

Shi L, Starfield B, Macinko J (2005) Contribution of primary care to health systems and health. Milbank Quart 83(3):457–502

Simchi-Levi, D Wei Y (2012) Long chain in process flexibility. Oper Res 60(5):1125–1141

Solak S, Clarke J-P, Johnson E, Barnes E (2010) Optimization of R&D portfolios under endogenous uncertainty. Eur J Oper Res 207(1):420–433

Starfield B, Macinko J, Shi L (2007) Quantifying the health benefits of primary care physician supply in the United States. Int J Health Serv 37(1):111–126

# Chapter 9
# Improving Scheduling and Flow in Complex Outpatient Clinics

**Craig M. Froehle and Michael J. Magazine**

## 1 Outpatient Care: Workhorse of the Healthcare System

By nearly any measure, outpatient care is increasingly critical to healthcare systems' abilities to care for the masses of patients requesting service. In the USA, like many countries, outpatient services are growing the fastest of any mode of care delivery (Fig. 9.1). Three drivers of this trend are (a) the increasing ability of medicine to treat patients without admitting them to the hospital, (b) the financial advantages of providing care via outpatient means, and (c) the ongoing need to reserve inpatient capacity (which is not growing at the same rate as the population and, in some areas, is actually shrinking) for the most severely ill patients (Bazzoli et al. 2003).

One of the greatest concerns of outpatient facility managers is the rapidly rising cost of providing care, both overall and per capita (Peterson and Burton 2007). This trend reinforces the need to operate outpatient clinics in the most efficient ways possible. Human resources, such as physicians and nurses, and physical resources, like exam rooms and testing equipment, often represent significant fixed expenses, so maintaining a reasonable level of utilization is vital to maintaining the financial viability of a clinic or hospital.

C.M. Froehle (✉)
Department of Operations, Business Analytics, and Information Systems, Carl H. Lindner College of Business, University of Cincinnati, 2925 Campus Green Drive, Cincinnati, OH, 45221-0130 USA

James M. Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229-3039, USA
e-mail: craig.froehle@uc.edu

M.J. Magazine
Department of Operations, Business Analytics, and Information Systems, Carl H. Lindner College of Business, University of Cincinnati, 2925 Campus Green Drive, Cincinnati, OH, 45221-0130 USA

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_9, © Springer Science+Business Media New York 2013

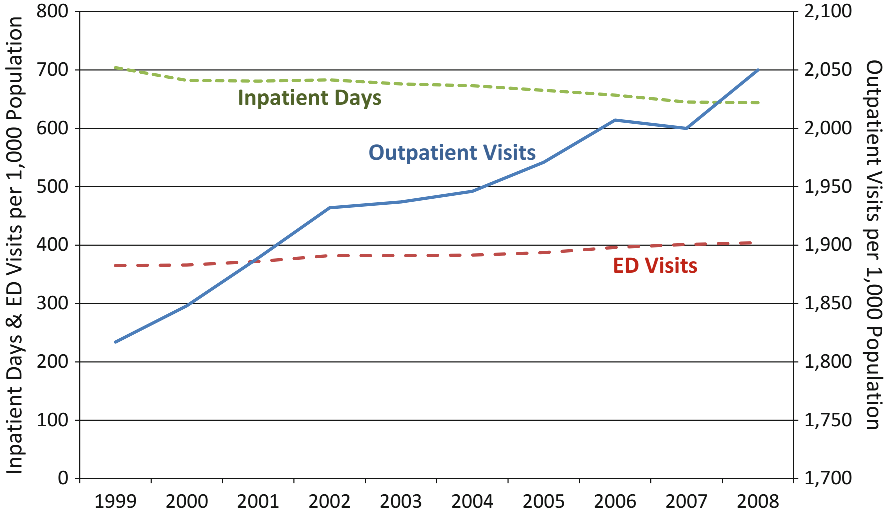


**Fig. 9.1** US hospital visits per 1,000 population, 1999–2008. *Notes*: Data are for community hospitals (85 % of all US hospitals), while federal hospitals, long-term care hospitals, psychiatric hospitals, and other specialty institutions not included; Source data: StateHealthFacts.org (2011)

Another motivation for improving the ability to manage flow, or the ability of a healthcare organization to "serve patients quickly and efficiently as they move through stages of care" (Hall 2008), in outpatient environments is that many of the lessons we learn there are applicable to other healthcare settings. For example, scheduling of non-emergency surgeries has benefited from the extensive work on scheduling in outpatient care, as both involve elective visits and are appointment-driven operations (Cardoen et al. 2010; Gul et al. 2011). Additionally, inpatient and emergency care areas of a hospital often have common patient-contact points or share resources with outpatient functions; staff can provide care in more than one location, and the services can share common physical space, such as parking facilities. In some instances, emergency department (ED) flow can be modeled very similarly to outpatient care clinics, especially for that portion serving less urgent patients. In fact, scheduling of non-emergent ED visits, just as is typical of outpatient visits, is becoming increasingly common (Fiore 2010). Thus, the findings from research on outpatient operations can often be immediately useful in other care contexts, if not also to other service industries.

As the title suggests, the focus of this chapter is on "complex" outpatient clinics. We use that term here to refer primarily to those clinics caring for patients who need access to more than a few different types of care providers during their visits and whose paths through the system may be significantly heterogeneous. For example, Fig. 9.2 shows the actual paths (i.e., the sequence of care providers) through a pediatric, pulmonary clinic taken by a sample of 30 patients. This clinic provides outpatient care to children who suffer from pulmonary diseases, such as asthma and cystic fibrosis. Note that while there is a single path that represents a typical (i.e.,

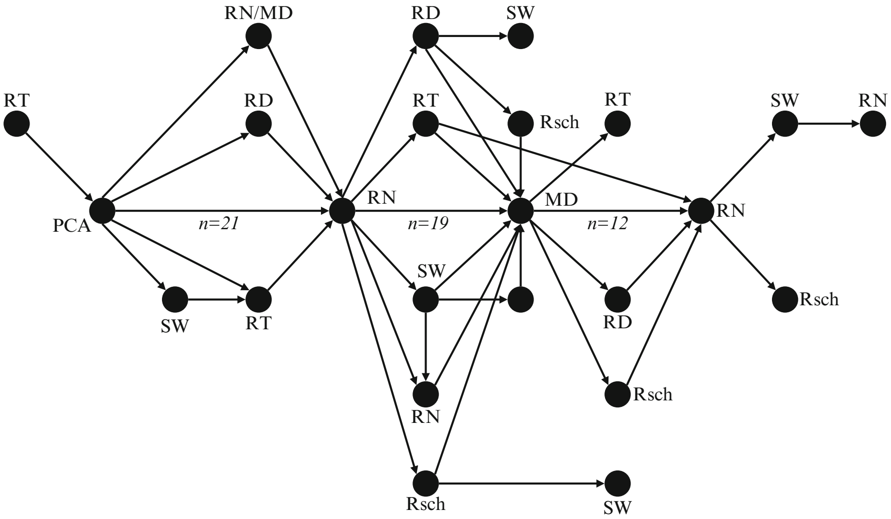


**Fig. 9.2** Realized patient paths through a complex outpatient pulmonary clinic. *Notes*: *RT* respiratory therapist, *PCA* patient care assistant, *RD* registered dietician, *RN* registered nurse, *MD* medical doctor (physician), *SW* social worker, *Rsch* researcher, *n* number of patients on a path; Source: Cincinnati Children's Hospital Medical Center, Division of Pulmonary Medicine

most common) route through the system (i.e., PCA (patient care assistant) → RN (registered nurse) → MD (medical doctor) → RN), less than half the patients in the sample traveled along this path. Also note that this does not include the registration activity, nor does it include initial or mid-process waits, all of which can add complexity to the flow in a clinic.

Operational complexity is often a concern for specialty and multidisciplinary clinics (Gupta and Denton 2008), those caring for patients with comorbidities or who are chronically ill, and teaching institutions that employ fellows and residents. Large patient volume is a contributing factor to complexity, but large patient volume alone does not necessarily make a clinic complex. Similarly, even if different types (medically diverse) of patients are treated, it may not be an operationally complex environment if all patients progress through the same care process. Akin to a typical assembly line, homogeneous flow, even at high volumes, is generally more straightforward to plan for and manage.

Realistically, different outpatient clinics may choose rather different operational objectives. For example, the high-volume, low-acuity pediatric clinic may prefer to minimize patient waiting, while the for-profit, imaging clinic may wish to maximize facility utilization, and the orthopedic clinic may prefer to maximize provider (physician) utilization while keeping average patient waits below some chosen value. Or each of these may choose a different metric (or set of metrics), depending on their priorities, constraints, and stakeholder preferences. Devising desirable performance metrics, and balanced combinations of metrics, is a vital element of improving any clinic's operations (see Table 9.1 for examples).

**Table 9.1** Some common operational metrics used in outpatient clinics

| Metrices | Notes/description |
|---|---|
| *Patient-centered metrics* | |
| Time to access clinic | Also referred to as "indirect waiting" (Gupta and Denton 2008) |
| Flow time | Total duration a patient is in the clinic |
| Waiting time | Also referred to as "direct waiting" (Gupta and Denton 2008) |
| Touch time ratio | Time patient is being served divided by flow time |
| *Provider-centered metrics* | |
| Provider idle time | Total time a provider is in clinic, but not serving patients |
| Provider utilization | Time a provider is serving patients divided by available time |
| Provider touch time ratio | Time serving patients divided by available time |
| *Organization-centered metrics* | |
| Clinic overtime | Amount of time past scheduled end that patients are in clinic |
| Number of patients seen | Less useful in clinics with heterogeneous patient mixes |
| Facility (exam room) utilization | Can be helpful for space allocation decision-making |
| Revenue per clinic | Can depend greatly on reimbursement assumptions |

In this chapter, we broadly examine some of the primary barriers and opportunities for operational performance and improvement in complex outpatient environments. That makes this work different but complementary to those focused on primary care. After reviewing the challenges, we propose a managerial and analytical framework that illustrates how solving various problems can contribute to a comprehensive operational approach to improved clinic flow. While this chapter is neither a comprehensive literature review nor an in-depth methodological treatise, it should help healthcare professionals understand how different studies and topics within operations research may be relevant to this context. It should also aid academics in identifying specific research opportunities for extending our knowledge of, and capability for improving, outpatient clinic operations.

## 2 Pervasive Barriers to Outpatient Clinics Operating Effectively

Two phenomena that greatly inhibit a clinic's ability to manage itself in an efficient and effective manner are *uncertainty* and *complexity*. By uncertainty, we mean the inability to know in advance, with perfect accuracy, some operational aspect of the clinic, such as the duration of a consultation activity or the set of providers with whom a patient will need to interact during his visit. By complexity, again, we refer primarily to the quantity and diversity of medical staff and physical resources involved in providing care to the clinic's patients.

These phenomena affect two fundamental operational activities of a clinic—(a) planning and scheduling and (b) managing clinic flow—in different ways (see Table 9.2). Planning and scheduling activities for an outpatient clinic involve those functions that help ensure that clinics will be as accessible to patients, and executed

**Table 9.2** Some implications of complexity and uncertainty for clinic operations

| | Planning and scheduling (prior to clinic) | Managing clinic flow (during the clinic) |
|---|---|---|
| Complexity | 1. Identifying all resources each patient needs to access takes more effort | 5. Necessary to orchestrate multiple providers' and patients' activities |
| | 2. Requires joint scheduling of multiple providers | 6. Aggregating and analyzing complete and timely operational data difficult to do manually |
| Uncertainty | 3. Patient/provider activity sequencing is generally suboptimal<br>4. Total patient resource needs unclear | 7. Arrival of patients unknown<br>8. Timing of tasks unknown<br>9. Care needs (especially new patients) unknown |

as smoothly, as possible. This encompasses a wide variety of tasks, such as initial scheduling of patient appointments, determining (or estimating) which services and providers each patient will need to see, reserving the necessary equipment and facilities, and so forth (Pinedo 2009). All of these planning and scheduling activities related to a specific clinic session occur before that particular clinic session starts.

In contrast, managing clinic flow is the daily challenge of implementing the clinic plan and directing the activities of staff and patients as the clinic occurs. During the clinic session, as patients arrive (early, on time, late, or not at all) and clinic staff meet with patients (for various amounts of time), deviations from the original schedule for that clinic nearly always occur. Reaction, in a timely manner, to these changes necessitates the constant reformulation of a new plan. That, combined with the responsibilities of guiding people and coordinating resources needed to deliver the appropriate care, contributes to the potential difficulty of managing clinic flow. Below, we discuss some of the particular challenges associated with both planning and scheduling activities and managing patient flow in complex and uncertain clinic environments.

## *2.1 Planning and Scheduling Prior to the Clinic Session*

A great deal of planning for a clinic session generally takes place well in advance of the session actually getting underway. In complex and chronic care clinics, scheduling patients in advance is commonplace. Creating the ideal appointment templates (the arrangement of appointment times and durations reserved for different patient/visit types who need to see specific providers or provider types), and scheduling in general, has been the subject of extensive operations research in healthcare over the years. SeeCayirli and Veral (2003) and Gupta and Denton (2008) for excellent overviews of this area of study and Pinedo (2009) for a general review of scheduling work that potentially could be applied to this environment.

After the schedule has been filled, or is nearly full, a common practice for planning for a clinic is for staff to hold a meeting prior to the clinic session to review the medical needs of the slate of patients scheduled to receive care during the clinic session. This meeting, which some hospitals and clinics refer to as a "chart conference," is typically held to ensure that each patient's visit includes all care activities needed to address that patient's particular condition, and is often additional work for clinicstaff (#1 in Table 9.2). One result of a chart conference is the need to then schedule those care providers who may not already be scheduled to see that particular patient. Coordinating the schedules of the various providers can be challenging, especially if the providers are being drawn from a multitude of specialties or practices (#2 in Table 9.2).

Just as in manufacturing and other production environments, in complex outpatient clinics, uncertainty can make planning and scheduling more difficult. Patient arrivals to the clinic can be earlier than scheduled, on time, later than scheduled, or not at all (no-shows) (Salzarulo et al. 2011). Another source of uncertainty is the duration of the patient–provider consultation. While most clinic schedule templates have predetermined blocks of time roughly corresponding with the expected duration of the patient's consultation with the physician and/or other staff, the actual durations can be highly variable (Harper and Gamlin 2003; Chand et al. 2009; LaGanga 2011).

This combination of uncertain patient arrivals and uncertain consultation durations makes creating the optimal schedule of all clinic activities difficult (#3 in Table 9.2) (Please see our approach to this in the research opportunities section of this chapter). Research has shown that scheduling low-variance patients earlier in the clinic, and high-variance patients later, can generate small reductions in patient waiting and overall clinic duration (White et al. 2011a). However, with few exceptions, these studies have generally considered only simple clinic environments (e.g., a single patient flow through a fixed sequence of providers), and there is as yet little evidence to guide scheduling of patients who need to see multiple providers and with different variability profiles.

Further complicating matters is that the exact resource needs of patients are not often fully known when planning the clinic session. For example, the specific providers a patient will need to see may not be entirely known until he has consulted with one or more care providers (#4 in Table 9.2). This is especially common with new patients and those with complex, chronic diseases. It creates a situation where the clinic must then react to changes in their assumptions about who needs to interact with that patient during his visit. An alternative that few prefer is to ask the patient to come back another day to meet with additional staff; this puts a hardship on the patient and creates another planning and scheduling responsibility for a future clinic. A different form of uncertainty regarding total patient resource needs (#4 in Table 9.2) is the fact that not all patients who will participate in the clinic may be known when planning and scheduling activities occur; add-ons (those added to the schedule after the chart conference has occurred) and walk-ins (those showing up without advance notice or an appointment) are two examples.

## 2.2 Managing the Flow of the Clinic

One of the keys to effective flow is proper planning and scheduling before the clinic session begins. As shown in Table 9.1, both complexity and uncertainty can limit the organization's ability to effectively manage clinic flow. As the complexity—the number and variety of care providers and patients—of the clinic increases, it becomes progressively more difficult to track and coordinate the actions of everyone involved (#5 in Table 9.2). The location and activity of each provider and patient, the time elapsed since they began those activities, and other details are essential to actively managing flow in a clinic. Keeping track of one or a few exam rooms with one or a few staff might be possible, even effective, using purely manual methods; however, any clinic more complex than that will likely require some form of information system to support the human decision-makers (#6 in Table 9.2).

Uncertainty surrounding patient arrivals creates a challenge for staff trying to manage the flow of the clinic (#7 in Table 9.2). Unless a patient contacts the clinic to let it know he will not be showing up on time, the clinic has no reliable information about when, or even if, the patient will ultimately arrive. This can make task sequencing decisions difficult. For example, a patient has arrived early (ahead of his appointment time) and another patient is late (has not yet arrived even though his appointment time has passed). If we place the present patient in an exam room (ahead of his appointment), it will delay serving the late patient should he arrive soon thereafter. While first come, first served (FCFS) may have been honored, FAFS (first appointment, first served) would be violated. See White et al. (2012) for work with an outpatient clinic that considered different queue disciplines.

Clear policies can help ensure consistent decision-making, but left unresolved is the question of whether or not the policy leads to the best decision at this particular moment with respect to the clinic's performance measures (e.g., average patient waiting, overtime, etc.). In the above example, the policy may be to make the early patient wait until his appointment time. Of course, that may be suboptimal if the late patient turns into a no-show; the waiting patient was delayed for no benefit to anyone. Some clinics have specific policies for this and other situations, but many do not, instead leaving it up to whoever may be managing the clinic at the time.

Similarly, if a patient and physician began their consultation 25 min ago and the scheduled duration is 30 min, there is little guarantee they will be done in less than 5 min, let alone exactly 5 min. This type of uncertainty (#8 in Table 9.2) creates a need for constant monitoring of all activities. Then, making decisions based on both actual events (the end of the consultation) and anticipated events (the fact that the physician and patient should both be freed up within the next few minutes) can easily overwhelm manual systems or those relying simply on a clinic manager acting as an "air traffic controller" (a common analogy in outpatient clinics).

Also, in complex clinics, the exact care needs of a particular patient may not be easily determined prior to physically meeting with the patient in the clinic (#9 in Table 9.2). Apart from the planning and scheduling challenges identified above, this complicates managing the flow of the clinic because additional activities may need to be inserted into the schedule while the clinic is happening. These real-time adjustments can pose significant problems for clinic managers for two reasons. First, the schedule was likely based on some estimates of how much staff time each patient would consume; if that time gets increased significantly, many other activities may need to be pushed back, potentially resulting in an overly long clinic, excessive patient waiting, and other issues. Second, if the added care component requires a specific type of medical staff, that resource may not be available at the time she is needed in the clinic. This can even further exacerbate an already chaotic and operationally compromised clinic.

In order to move patients through their clinic visits with minimal delay while keeping staff and facilities adequately utilized, a comprehensive approach is needed that integrates decision-making associated with planning and scheduling and managing flow in real time. A proposed framework for such a system is described in the next section. The framework represents an assembly of existing policies and practices that can be found in many hospitals and clinics today as well as opportunities for overcoming the inadequacies of unassisted, manual decision-making in these dynamic, complex, and uncertain environments. Those opportunities are then further discussed in the final section of this chapter.

## 3 COMS: A Proposed Approach for Improving Flow in Outpatient Clinics

We refer to the Clinic Operations Management System (COMS) as a set of policies, practices, and analytical methods that may be embedded in a clinic's information system. It is a conceptual framework, used here to frame the discussion about the broader problem of planning and scheduling outpatient clinics in a comprehensive manner. To date, the literature has largely focused on planning and scheduling scenarios that consider only simplified models (e.g., a single server or no recourse over the clinic's duration) or address only subsets of the overall process. In contrast, COMS details one possible vision of an integrated system, with many opportunities for applying advanced analytical methods to improve the clinic's operating performance. A functional schematic is shown in Fig. 9.3, where COMS can be seen to assume four main responsibilities:

(a) It serves as the central clearinghouse for clinic schedule data.
(b) It assists planning by providing information about past patient visits and by recording decisions made about upcoming patient visits (e.g., which providers a patient must see during his visit).
(c) It tracks patients and providers as they move through the clinic (or works in conjunction with a separate real-time locating system, or RTLS).
(d) It supports real-time operational decision-making for coordinating the activities of staff and patients as the clinic progresses.

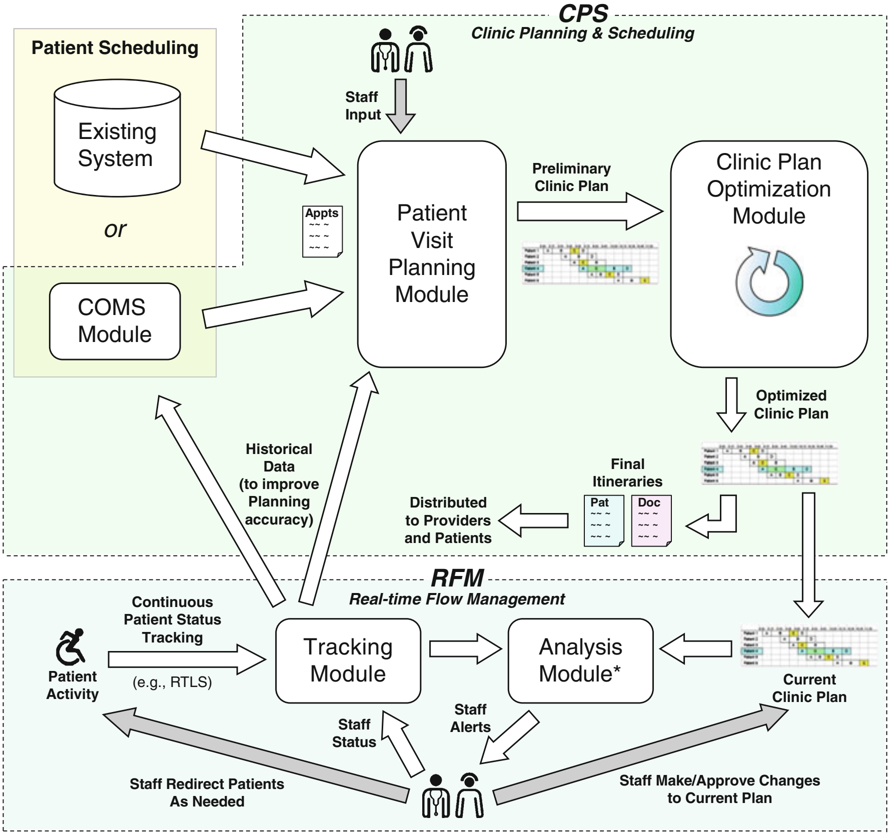

CPS
Clinic Planning & Scheduling
Patient Scheduling
Existing System
or
COMS Module
Staff Input
Appts
Patient Visit Planning Module
Preliminary Clinic Plan
Clinic Plan Optimization Module
Optimized Clinic Plan
Final Itineraries
Pat
Doc
Distributed to Providers and Patients
Historical Data (to improve Planning accuracy)
RFM
Real-time Flow Management
Continuous Patient Status Tracking
(e.g., RTLS)
Patient Activity
Tracking Module
Analysis Module*
Current Clinic Plan
Staff Status
Staff Alerts
Staff Redirect Patients As Needed
Staff Make/Approve Changes to Current Plan

**Fig. 9.3** COMS functional architecture. The Clinic Operations Management System (COMS) represents the major information flows, analytical components, and coordination mechanisms of a complete outpatient clinic planning, scheduling, and flow management system. *Asterisk* could employ analytical approaches similar to the Clinic Plan Optimization Module (see Sect. 4 for more details)

Responsibilities (a) and (b) are associated with improving the clinic's planning and scheduling activities, whereas responsibilities (c) and (d) have to do with improving the clinic's ability to do real-time flow management. Planning and scheduling activities occur prior to a clinic session, while flow management occurs during a specific clinic session.

## 3.1 *Clinic Planning and Scheduling Functions*

Clinic Planning and Scheduling (CPS) incorporates all functions needed by clinicians to collect, structure, and organize the operational information involved in planning a particular clinic session (e.g., morning clinic, Wednesday, August 14th), including:

1. Patient appointment requests—Ideally, when a patient contacts the clinic (or call center) for an appointment, the system (COMS or a separate scheduling system with information fields populated by COMS) would suggest appointments of appropriate durations based on the type of appointment being requested (e.g., new vs. follow-up) that permit that patient to see all appropriate providers in as compact a schedule (minimal patient waiting and provider idle time) as possible.
2. Patient visit planning—Before the patient arrives at the clinic, a list (or sequence, if task precedents are important) of service activities (providers, tests, etc.) for that patient's visit must be provided by clinic staff, such as, through the aforementioned chart conferences. These lists can be, and will likely initially start out as, generic templates, perhaps differentiated by patient category. However, the more accurate and detailed these patient visit plans are, the better able the system will be to optimize the operating plan for that clinic. This combination of patient arrivals (appointments) and the slate of providers each patient should see represents a preliminary *clinic plan* for that clinic session and is fed into the next step: optimization.
3. Clinic plan optimization—Once a clinic's available appointment slots have been filled, or the start of the clinic is a minimum planning period away (e.g., two days from now), an "optimization"[1] module will arrange patients, providers, and/or tasks so as to achieve one, or a combination, of several objectives (e.g., minimize patient waiting, minimize staff idle time, minimize total patient flow time, and maximize robustness to disruption). This can be achieved in several ways, which are discussed in the last section of this chapter.

The result of these three tasks is a comprehensive plan for a specific clinic session, including the roster of patients to be served, the (desired) order of their arrival, the sequence of service activities for each patient, and the planned start and end times for each activity. These sequences, plus their associated start and end times, create *itineraries* for both patients and providers as they travel through the clinic's overall operating plan. Figure 9.4 shows a small clinic plan covering one physician, a total of four different care providers/activities (e.g., medical assistant, nurse, physician, and phlebotomy), and six patients.

Patient 4's itinerary (shaded row) would consist of:

```
a. Arrive and see Provider A at 9:00
b. See Provider C at 9:15
c. See Provider B at 9:45
d. See Provider D at 10:15
e. Exit at 10:30
```

[1] The term "optimization" is used here to include both models/systems that produce truly optimal solutions as well as those that may rely on other methods, such as heuristics, to generate optimal or near-optimal solutions.

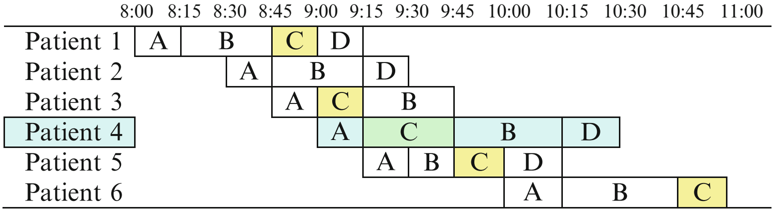


**Fig. 9.4** Example of a clinic plan

Provider C's itinerary would consist of:

1. See Patient 1 at 8:45 for 15 minutes
2. See Patient 3 at 9:00 for 15 minutes
3. See Patient 4 at 9:15 for 30 minutes
4. See Patient 5 at 9:45 for 15 minutes
5. Break at 10:00 for 45 minutes
6. See Patient 6 at 10:45 for 15 minutes

The planned duration of each scheduled patient–provider task on the itinerary (e.g., a physician consulting with the patient) should be based on a combination of historical performance (i.e., data), ideal consultation times depending on the particular patient and reason for his visit, and the clinic's operational objectives. The system should know these estimated/planned task durations, which could be determined by clinic management, calculated from historical data, or some combination thereof.

## *3.2 Real-Time Flow Management Functions*

All planning and scheduling activities described immediately above take place prior to the start of the clinic session. Then, with the optimized clinic plan in place, and after all patients and providers have been informed as to their individual itineraries, the clinic session begins. Ideally, all patients arrive at their designated start times and do not require significantly more time from any provider or in any task than was originally scheduled in the plan. If these conditions are met, it should be possible to come close to achieving the clinic's operational performance goals.

However, it will be rare that all patients show up when they are asked, that all providers spend the planned amount of time with each patient, and that there are no delays due to missed communication, misplaced files, or the myriad other sources of disruption typical of an outpatient clinic. When such disruptions happen, COMS must be able to advise the clinic manager as to the best recourse depending on the nature and magnitude of the deviation(s) from the plan, the current state of the system, and other policies and objectives that may influence the clinic's operations.

For example, patient D shows up 20 min late for his 10 a.m. appointment (10 a.m. being the time he was scheduled to first sign in with registration, *not* when he was to see the physician, as is the customary use of the "appointment time" concept). Should the clinic serve this patient (i.e., move him to the next part of his itinerary) next, should it prioritize another waiting patient who was not late, or should it take some other action? Few individuals running a clinic larger than a single physician and a few exam rooms will not have the awareness and real-time analytical capacity necessary to ensure that optimal decisions are consistently known. The system should assess the state of the clinic (e.g., patient A is in task 3 of his itinerary, patient B is in task 2, and patient C is in task 1; each has been in that task for a number of minutes and, therefore, is estimated to have a certain number of minutes remaining before moving onto the next task); consider patient D's itinerary, as well as any patients scheduled to show up soon; and determine what course of action will most likely achieve the clinic's desired operational results.

Rerunning the decision model every few minutes (or on demand, perhaps motivated by a significant event) would ensure that the clinic staff constantly have access to the best (or approximate best, if heuristics are relied on) course of action based on the latest information about the current state of the clinic. While powerful, this rescheduling capability, which also exists in many MRP systems (Ho 1989), causes disruptions to the system that may actually make things worse. These frequent shocks to the system, called "system nervousness," may not actually move the solution to a better place and can lead to staff and patient unrest. The challenge is to find the proper level of nervousness that balances the value of acting on new information with the cost of frequent adjustment.

One common scenario that such a decision model should be ready to address is the arrival of an unscheduled, or add-on, patient who was not part of the original plan developed for the clinic during the planning and scheduling process. Upon being informed that a patient will be arriving, or has arrived, for which the clinic plan was not prepared, COMS would ideally provide clinic management with the best course of action (e.g., serve immediately and at which task, serve at some later point in the clinic, or schedule for another day). It would also be necessary to input a new preliminary plan for the add-on patient. COMS would then generate an itinerary for him as well as update the staff itineraries for all those providers who will see the add-on patient during that clinic session.

### *3.3 Other Functions*

In order to maximize its usefulness and overcome the barriers to clinic operating effectiveness identified earlier, COMS should handle several other responsibilities. Tracking of individuals and other resources will be requisite in order for the system to know the status of each patient, provider, and exam room at all times. Tracking should not rely on manual entry of one's status into the system, as that introduces data errors, delay, and unreliability, all of which reduce COMS' ability to make accurate recommendations. Therefore, the tracking system should require no action on part of providers or patients for it to know where they are at all times during the clinic. Some existing, commercially available RTLS solutions have this functionality, relying on RFID, ultrasound, and other technologies, but a constant data feed from the RTLS to COMS would be necessary.

An equally important function is the ability to record operating data as it is generated during the clinic session and synthesize those data into useful inputs for the planning and scheduling process. For example, if every consultation by a physician is accurately measured, then the system will be much better equipped to estimate the duration of a specific consultation involving that physician and a patient of a certain type (if not the specific patient). Or if patient arrivals are tracked accurately, then the clinic can have more confidence in certain decisions, such as determining which appointment slots should be overbooked with a second patient. Without these data, the clinic must resort to relying on averages and rough estimates at best, or simple guesses at worst. Unfortunately, in practice, the quality of operational data is often less than ideal, possibly because hospitals and clinics must often rely on manual data collection methods, which can be costly (or frustrating for staff and patients) to adopt as a standard practice. Automating the collection of highly detailed operational data, which many healthcare information systems do not do, represents a significant opportunity for improvement.

A system like COMS could also enhance the clinic staff's ability to participate during the planning and scheduling process. A browser-based interface for collaborative patient-visit planning by multiple providers would be helpful when face-to-face chart conferences are impossible. Automating the distribution of patient and provider itineraries (e.g., via email or post) would also be a natural function for the system.

The chaotic environment of a busy clinic can make coordinating, and communicating with, the patients, staff, rooms, etc., during the clinic session a daunting task. COMS should facilitate these activities by providing real-time status updates via a variety of media and means. For example, real-time display of the current clinic plan on large displays (e.g., wall-mounted displays) as well as portable devices (e.g., smartphones) would be invaluable to staff who are constantly circulating around the clinic. Also, flexible, multimode delivery of staff alerts, such as via text messaging, paging, email, and perhaps push notifications (such as through custom smartphone apps), would improve the timeliness of staff being alerted to changes to the clinic plan without being overly intrusive and disrupting patient–provider consultations. Or there may be a need to disrupt those meetings, such as in the case of a provider being unaware of the time (and patients queueing up in the waiting room), where an automated nudge would be helpful to indicate that it is time to wrap up the consultation.

COMS is a general concept, comprised of many elements, some of which are well established in practice, whereas others have yet to be developed. The next section explores in greater detail opportunities represented by some of the functions included in the COMS model, both in terms of innovation for practice and research for academics.

# 4 Optimizing the Clinic Plan

Perhaps, the primary opportunity for innovative improvement to outpatient clinic management is the clinic plan optimization component required by COMS. Multiple approaches could be taken in constructing a model (or system of models) to optimize the clinic plan. While solving a flow shop-style scheduling problem has been widely addressed in the literature (e.g., Minella et al. 2008; Li and Ierapetritou 2008; Pinedo 2009; Baker and Trietsch 2009), the fact that some tasks may have appointment times (versus complete jobs having due dates, as is typical of job-shop scenarios) makes this variation on the traditional machine-scheduling problem more complicated.

One approach could be to treat the set of patient arrival times—the order and approximate times of patient arrivals—as exogenous and fixed, unable to be modified by clinic plan optimization. Taking that approach would limit the model to manipulating only the sequence of activities for each patient within his visit. This can become a challenging problem as more activities, other than arrival times, are associated with appointment times: If there is only one scheduled activity (e.g., a physician's consultation) for each patient, there may be greater opportunity to resequence all of the patient's other activities in order to avoid provider conflicts, reduce patient idle time, etc. However, some care providers in addition to the primary physician may also have their appointments scheduled. This is common for hospital-based providers who may see patients in several different specialties' clinics throughout the day. If several of a patients' tasks are associated with appointment times (e.g., the physician consultation, the dietician consultation, and the physical therapist consultation), the opportunities for improving the patient's itinerary may be very limited. If meeting all of these patient–provider appointment times is considered desirable, but not mandatory (e.g., deviation from the desired appointment time is included in the objective function instead of making task start times constrained to be equal to appointment times), the model may produce better results while still staying within reasonable windows around the original appointment times. This approach also lends itself very well to application both as the initial planning optimizer and the real-time clinic management *re*-optimizer, since reasonably short solution times should be possible, especially if deterministic consultation times are assumed (Froehle et al. 2011).

An alternative approach to planning the clinic could be to treat the patient scheduled arrival times as flexible, which should permit better solutions through manipulation of patient arrival order in addition to the provider/task sequence within each patient described above. In the absence of a scheduling policy where all patients arrive at the beginning of the clinic, taking this approach may require a two-stage patient appointment-setting process; the patient would first be assigned to a specific clinic (e.g., the morning of Wednesday, August 14th), and then, once the roster is full and the planning problem is solved, the patient would be contacted with a specific arrival time (e.g., 10:30 a.m.). A similar two-stage approach is relatively common in scheduling elective surgeries, where "first, advance scheduling gives patients some future date for surgery [and] second, allocation scheduling determines the sequence and resource assignment of the cases in a given day" (Pham and Klinkert 2008). Of course, in an outpatient setting, patients' unwillingness to learn the specific time of their appointment so close to the visit date may limit the utility of this approach.

Another method commonly used is "open access" (also known as "advanced access") scheduling, where some capacity is reserved for patients who need appointments on short notice (Murray and Berwick 2003). While not typically used for the complex and/or multidisciplinary clinics discussed in this chapter, investigating its use in those contexts may be fruitful. See Dobson et al. (2011)and for discussion of this topic.

Many, if not most, clinic scheduling optimization models take a single-period approach; they determine the best plan now and assume future disruptions, which could drive the clinic away from the optimal plan, will not happen. Immediately below is an example of a mixed-integer programming model that provides an initial plan for a clinic using predetermined patient arrival times and focusing on arranging the sequence of patient–provider tasks within each visit.

### *4.1 Example: Sequencing of Tasks in a Clinic*

Considering the clinical environment, an optimization model that would determine the sequence of tasks to be performed on a set of patients with scheduled arrivals could be useful. Here, a mixed-integer programming model has the aim of sequencing all the given tasks for all patients such that the total time in the system for the patients, weighted by their respective processing time, would be minimized. As a first step, we assume processing time for each task to be deterministic and therefore minimizing total time in system amounts to minimizing patient wait time. Weighting by total processing time results in taking a relative perspective on patient wait time, where patients who necessitate longer processing time can wait longer than patients who only need short services.

Consistent with clinical workflow, the model uses the following inputs: a set of patients scheduled for the given block schedule; a set of tasks to be performed on each patient; a set of resources that will perform the tasks; and a duration for each task. In addition, the tasks are assigned to one of four categories (not counting the registration task, which must occur first): (a) tasks that must be performed before the patient sees the physician; (b) tasks that must be performed after the patient sees the physician; or (c) tasks that can be performed either before or after the patient sees the physician. For example, taking the patient's vitals may be a task that has to occur before the patient sees the physician. Similarly, a patient might have to get a blood test after seeing the physician. Other tasks, such as seeing a dietician, could take place either before or after the encounter between the patient and physician. The fourth category of tasks is (d) the encounter between the patient and the physician. This categorization constitutes a constraint on the sequencing and has to be captured in the optimization model. The physician is typically viewed as the bottleneck resource, and thus, the objective of the model is to minimize physician idleness. In summary, each task is uniquely identified, involving a patient, a resource, a duration, and a sequencing category. We now define the elements needed to formulate the optimization model and then express the model in mathematical terms.

Decision variables:

- $s_u =$ the start time of task $u$
- $w_l =$ the amount of time in the system for patient $l =$ completion time—appointment time for patient $l$
- $x_{u,v}^k = 1$ if task $u$ occurs before $v$ for resource $k$, zero otherwise
- $y_{u,v}^l = 1$ if task $u$ occurs before $v$ for patient $l$, zero otherwise

Sets:

- $T = \{$Tasks to be sequenced during the time horizon (e.g., one clinic day)$\}$
- $P = \{$Patients scheduled to visit the clinic during the time horizon$\}$
- $R = \{$Resources involved in the clinic (the nurses, the physician, the dietician, etc.)$\}$
- $R_k = \{$Tasks to be performed on resource $k\}$
- $P_l = \{$Tasks to be performed on patient $l\}$
- $O_l = \{$Registration task(s) for patient $l\}$
- $A_l = \{$Tasks of category A to be performed on patient $l$, where A tasks have to occur before the physician sees the patient$\}$
- $B_l = \{$Tasks of category B to be performed on patient $l$, where B tasks involve the physician$\}$
- $C_l = \{$Tasks of category C to be performed on patient $l$, where C tasks have to occur after the physician sees the patient$\}$
- $D_l = \{$Tasks of category D to be performed on patient $l$, where D tasks can occur either before or after the physician sees the patient$\}$

Model parameters:

- $t_u =$ the duration of task $u$
- $a_l =$ the appointment time of patient $l$
- $p_l = \sum\limits_{u \in P_l} t_u$
- $\alpha_l = \frac{1}{p_l}$
- $M$ is a nonrestrictive large value in the context of this environment (e.g., the sum of all the processing times)

The objective function can then be defined as:

$$\min \sum_{l}^{P} \alpha_l w_l \tag{9.1}$$

The objective is fulfilled by determining the starting time of each task, thus providing a sequence of tasks. To ensure that the sequence of tasks is valid, we insert the following constraints, where each equation below represents a set of constraints:

$$s_u \geqslant a_l \quad \forall l \in P,\ \forall u \in P_l \tag{9.2}$$

$$M(1-x^k_{u,v}) + s_v \geqslant s_u + t_u \quad \forall k \in R,\ \forall u,v \in R_k \quad \text{such that} \quad u < v \tag{9.3}$$

$$Mx^k_{u,v} + s_u \geqslant s_v + t_v \quad \forall k \in R,\ \forall u,v \in R_k \quad \text{such that} \quad u < v \tag{9.4}$$

$$M(1-y^l_{u,v}) + s_v \geqslant s_u + t_u \quad \forall k \in P,\ \forall u,v \in P_l \quad \text{such that} \quad u < v \tag{9.5}$$

$$My^l_{u,v} + s_u \geqslant s_v + t_v \quad \forall k \in P,\ \forall u,v \in P_l \quad \text{such that} \quad u < v \tag{9.6}$$

$$y^l_{u,v} = 1 \quad \forall l \in P,\ \forall u \in A_l,\ \forall v \in B_l \quad \text{such that} \quad u \neq v \tag{9.7}$$

$$y^l_{u,v} = 1 \quad \forall l \in P,\ \forall u \in B_l,\ \forall v \in C_l \quad \text{such that} \quad u \neq v \tag{9.8}$$

$$y^l_{u,v} = 1 \quad \forall l \in P,\ \forall u \in O_l,\ \forall v \in A_l \quad \text{such that} \quad u \neq v \tag{9.9}$$

$$y^l_{u,v} = 1 \quad \forall l \in P,\ \forall u \in O_l,\ \forall v \in D_l \quad \text{such that} \quad u \neq v \tag{9.10}$$

$$w_l \geqslant s_u + t_u - a_l \quad \forall l \in P,\ \forall u \in P_l \tag{9.11}$$

$$s_u, w_l \geqslant 0 \quad \forall u,l \quad x^k_{u,v}, y^l_{u,v} \in \{0,1\} \quad \forall u,v,k,l \tag{9.12}$$

Equation (9.2) ensures that a patient's tasks cannot start before the patient's appointment time. Equations (9.3) and (9.4) capture the fact that a given resource can only process one task at a time and that any task is either before or after another task. These two equations capture these requirements as they are the integer program formulation of a disjunctive pair, meaning that either one or the other constraint, but not both, must hold for a solution to be feasible. Similarly, Eqs. (9.5) and (9.6) capture the fact that for a given patient, only one task can be performed at a time and that any task is either before or after another task. With Eqs. (9.7) through (9.10) we account for the categorization constraints on the order in which tasks have to be performed for a given patient, where any task in $A_l$ should be before any task in $B_l$, any task in $B_l$ should be before any task in $C_l$, and registration tasks should be performed before any other task. Finally we use Eq. (9.11) to capture the total time in system for each patient.

To test this example, we coded the model in the AMPL language and solved several small instances using CPLEX (ILOG). Based on an actual pediatric clinic as a convenient example, in each instance solved, we kept the number of resources constant, capturing registration, two nurses, a physician, a dietician, and a social worker, totaling six different caregiver resources. We assumed one unit of each resource (one person), so that the resource can only handle one patient at a time. Also, for convenience, we assumed the patients arrived every 15 min, and, given the short registration task, we assumed this created no initial patient waiting. We were able to solve these problems in a reasonable amount of time, but further work refining the model and developing solution techniques (or heuristics) will likely be necessary for clinics having more patients, providers, and provider types.

## 5 Opportunities for Future Research and Practice

The improvement of outpatient clinics' operational effectiveness is a broad topic area with many avenues to explore further. Below, we discuss three: (1) the effects of disruptions on efforts to optimize the clinic plan; (2) multidisciplinary coordination; (3) the use of operational data to improve future planning and scheduling activities; and (4) the challenges of managing human behavior in these complex operational environments.

### *5.1 Disruptions*

Since optimizing the clinic plan, such as is accomplished by the model in Sect. 4, is a planning problem, typically solved one or more days prior to the start of the clinic session, we can afford the luxury of finding optimal solutions. But, further research on reformulations (see, e.g., Camm et al. 2008) may improve computation times and enable real-time use of this model. Naturally, this MIP uses deterministic data. However, given the almost certainty of disruptions (e.g., tardy or no-show patients or longer-than-expected consultation times) occurring during clinic operation, we suggest taking one of two approaches. First, the solution process could incorporate an iteration loop that takes the above model's solution and simulates the clinic with realistic and stochastic processing times. Research is needed to develop appropriate measures of robustness (ability to remain a "good" plan even in the face of ordinary disruptions) for the clinic and, using these metrics, revise the optimization model to develop a new clinic plan. This could be repeated until there is an acceptable and reasonably robust plan. Other approaches that might be appropriate here could be the use of safe scheduling, where buffers are added to task times to take into account potential task time changes (Baker and Trietsch 2009).

A second alternative (or perhaps complimentary) approach to handling disruptions would involve regularly re-solving the optimization model during the clinic session. Once we enter the real-time flow management (RFM) phase (i.e., at the start of the clinic), disruptions will occur, making the current plan potentially difficult to adhere to, if even still feasible. With each event epoch (e.g., patient arrivals, task completion, patient departure) or on some predetermined schedule (e.g., every 10 min), we could attempt to re-optimize the clinic plan to get the patient flow back on track to try to meet the clinic's operational goals. Unfortunately, these disruptions can happen often, and solving a mixed-integer program to make adjustments may take too long. Plus, system nervousness might dictate less frequent changes. So, fast heuristics, such as swapping upcoming tasks, may be needed to guide rapid updates to the current clinic's operating plan. Research into practical algorithms, both from a run-time point-of-view and also from realistic clinic change possibilities, is needed. For instance, a new solution that tells the next arriving patient (due to arrive in 5 min) to come 1 h later is likely not a reasonable response.

## 5.2 Coordination

A related area of opportunity concerns the growing use of multi-specialty clinics, where patients with very complex conditions may end up seeing physicians and specialists from several different areas of medicine (e.g., cardiology, immunology, and pulmonology) in a single visit. Since these specialists will be coming from different parts of the organization, if not from different organizations entirely, scheduling them in a way that works with their other commitments as well as produces a reasonably compact schedule for the patient is increasingly desirable. Analytically, as the scope of the scheduling problem grows, dimensionality can become a significant barrier to producing optimal schedules. Therefore, again, heuristic-based approaches will likely need to be developed.

And since these heuristics will be of little use by themselves, embedding them within pan-organizational, or even interorganizational, information systems, will be necessary to add the most value to practice. This problem of coordination across multiple divisions of a hospital or other large healthcare provider has been around for decades. Mayo Clinic first dealt with this issue shortly after World War II by establishing a "central appointment desk," which coordinated appointment scheduling for multispecialty patient visits (Berry and Seltman 2008). Today, fast, responsive, and automated information systems offer significant opportunities for improving the effectiveness of these scheduling coordinationactivities.

## 5.3 Using Operational Data Better

One important function of COMS is its use of historical clinic data to forecast consultation durations involving different combinations of care providers and patients. This is helpful in determining a clinic plan that will be less likely to disruption, as the plan will already reflect the best possible estimate of the task duration (and likely far better than a standard schedule template, which generally attempts only crude differentiation among patients). Mining those historical data in order to make better forecasts about upcoming clinic visits represents a significant opportunity for both research and practice. While standard regression models may provide more accurate predictions than taking global means, using methods to identify and disaggregate influential subpopulations may further improve the accuracy of these estimates.

Another opportunity for future research is to develop ways to draw from the advances made in goods-producing industries, such as by seeking multiple sources of data and using multiple forecasting techniques, including quantitative as well as qualitative methods (Makridakis and Winkler 1983; Fisher and Raman 1996). While historical data will likely be informative, there may be some special cause or new influence that may create deviation from historical patterns. Seeking estimates from all the nurses and physicians that have cared for this patient, or patients like him, could be a useful supplement to the historical data when developing the best prediction of how long an upcoming consultation will take.

### *5.4 Managing Behavior*

Finally, there is the issue of modifying the behavior of patients or providers, if not both. To this point, we have largely focused on accommodating sources of variability (which can lead to uncertainty if the variability is not easily predicted) by improving both our ability to plan for it and our ability to react to it in a more timely and rational manner. A complementary strategy that exists is reducing the variability so that less uncertainty has to be accommodated. For example, while much of the discussion above has involved approaches to predicting and compensating for consultations of uncertain durations, we could combine that with managerial tools that reward more consistent adherence to the scheduled durations (or, alternately, that discourage gross deviations). While there will always be legitimate reasons to extend a consultation (or to end it earlier than planned), not all activities during a consultation might add value and could be safely eliminated to help keep the clinic on schedule. If awareness of the time elapsed during a consultation is a barrier to providers keeping true to their itineraries, a system of automated reminders, such as pages or texts sent to cellular phones, might be implemented (an excellent addition to the functionality of COMS, which will already be trackingactivities).

A better understanding of the sources of this variability will require further research but will be invaluable in helping outpatient clinics plan and manage their flow. This is especially true in complex clinics, as the variability from multiple consultations can accumulate to produce wildly varying operational performance from clinic session to clinic session. It is important to understand that variability cannot be eliminated completely. It is possible, however, to distinguish that variability created by provider input, or even the schedule itself, from "natural" (exogenous) variation, such as that arising from different patients' needs. It is desirable to minimize this first kind of variability, as suggested by Litvak (2005).

## 6 Conclusions

This chapter explored some of the challenges of achieving consistently excellent operational performance in complex outpatient clinics, offered a comprehensive model of how scheduling and flow might be improved, and discussed related opportunities for innovation in practice and in research. As complex outpatient visits becomes increasingly common, the problem of poor clinic flow will continue to undermine the financial viability of healthcare providers while simultaneously motivating patients to avoid care (or seek it through alternate means). If complex care clinics are to improve their operations, they must begin to consider how tools like COMS can be developed and implemented. Researchers must also contribute to this objective by developing new analytical methods and technologies. With the increasing investment in electronic health record (EHR) systems, and the evolution of new technologies, such as RFID and other sensors, there is tremendous opportunity for making dramatic gains in the operational effectiveness of complex outpatient clinics.

## References

Baker KR, Trietsch D (2009) Safe scheduling: setting due dates in single-machine problems. Eur J Oper Res 196(1):69–77

Baker KR, Trietsch D (2009) Principles of sequencing and scheduling. Wiley, New York

Bazzoli GJ, Brewster LR, Liu G, Kuo S (2003) Does U.S. hospital capacity need to be expanded? Health Aff 22(6):40–54

Berry LL, Seltman KD (2008) Management lessons from Mayo Clinic. McGraw-Hill, New York

Camm J, Magazine M, Polak G, Zaric G (2008) Scheduling parallel assembly lines to minimize a shared pool of labor. IIE Trans 40(8):749–758

Cardoen B, Demeulemeester E, Belien J (2010) Operating room planning and scheduling: a literature review. Eur J Oper Res 201(3):921–932

Cayirli T, Veral E (2003) Outpatient scheduling in healthcare: a review of the literature. Prod Oper Manag 12(4):519–549

Chand S, Moskowitz H, Norris JB, Shade S, Willis DR (2009) Improving patient flow at an outpatient clinic: study of sources of variability and improvement factors. Health Care Manag Sci 12(3):325–340

Dobson G, Hasija S, Pinker EJ (2011) Reserving capacity for urgent patients in primary care. Prod Oper Manag 20(3):456–473

Fiore K (2010) Living room is new ED waiting room. MedPage Today. http://www.medpagetoday.com/EmergencyMedicine/EmergencyMedicine/23315. Accessed 12 May 2011

Fisher ML, Raman A (1996) Reducing the cost of demand uncertainty through accurate response to early sales. Oper Res 44(1):87–99

Froehle CM, Platt M, Magazine MJ (2011) Optimal outpatient clinic scheduling and rescheduling. University of Cincinnati working paper

Gul S, Denton BT, Fowler JW, Huschka T (2011) Bi-criteria scheduling of surgical services for an outpatient procedure center. Prod Oper Manag 20(3):406–417

Gupta D, Denton BT (2008) Appointment scheduling in health care: challenges and opportunities. IIE Trans 40:800–819

Hall RW (2008) Patient flow: the new queueing theory for healthcare. OR/MS Today 33:36–40

Harper PR, Gamlin HM (2003) Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. OR Spectr 25:207–222

Ho C-J (1989) Evaluating the impact of operating environments on MRP system nervousness. Int J Prod Res 27(7):1115–1135

LaGanga L (2011) Lean service operations: reflections and new directions for capacity expansion in outpatient clinics. J Oper Manag 29(5):422–433

Li Z, Ierapetritou M (2008) Process scheduling under uncertainty: review and challenges. Comput Chem Eng 32(4–5):715–727

Litvak E (2005) Optimizing patient flow by managing its variability. In: Berman S (ed) Front office to front line: essential issues for health care leaders. Joint Commission Resources, Oakbrook Terrace, pp 91–111

Makridakis S, Winkler RL (1983) Averages of forecasts: some empirical results. Manag Sci 29(9):987–996

Minella G, Ruiz R, Ciavotta M (2008) A review and evaluation of multiobjective algorithms for the flowshop scheduling problem. INFORMS J Comput 20(3):451–471

Murray M, Berwick DM (2003) Advanced access: reducing waiting and delays in primary care. JAMA 289(8):1035–1040

Peterson CL, Burton R (2007) U.S. health care spending: comparison with other OECD countries. CRS Reports for Congress, Congressional Research Service, Washington, DC. http://assets.opencrs.com/rpts/RL34175_20070917.pdf. Accessed 25 Aug 2011

Pham D-N, Klinkert A (2008) Surgical case scheduling as a generalized job shop scheduling problem. Eur J Oper Res 185(3):1011–1025

Pinedo M (2009) Planning and scheduling in manufacturing and services. Springer, New York

Salzarulo PA, Bretthauer KM, Côté MJ, Schultz KL (2011) The impact of variability and patient information on health care system performance. Prod Oper Manag 20(6):848–859

StateHealthFacts.org (2011) www.statehealthfacts.org/comparetrend.jsp?cat=8&sort=a&sub=94&typ=1&yr=63&ind=392&srgn=1. Accessed 18 Jan 2011

White D, Froehle CM, Klassen KJ (2011a) The effect of integrated scheduling and capacity policies on clinical efficiency. Prod Oper Manag 20(3):442–455

White D, Magazine M, Morelli S (2012) Who's next: using patient analysis and operations management to address non-punctual and no-show patients. University of Cincinnati working paper

# Chapter 10
# No-Show Modeling for Adult Ambulatory Clinics

**Ayten Turkcan, Lynn Nuti, Po-Ching DeLaurentis, Zhiyi Tian, Joanne Daggy, Lingsong Zhang, Mark Lawley, and Laura Sands**

## 1 Introduction

Access to health services is usually controlled by appointment scheduling practices. Over the past few years, there has been resurgence in appointment scheduling research, largely fueled by healthcare applications where access is vital and challenges

A. Turkcan
Department of Mechanical and Industrial Engineering, Northeastern University,
Boston, MA 02115, USA
e-mail: a.turkcan@neu.edu

L. Nuti • L. Sands
School of Nursing, Purdue University, West Lafayette, IN 47907, USA
e-mail: lnuti@purdue.edu; lsands@purdue.edu

P.-C. DeLaurentis
Oncological Sciences Center, Purdue University, West Lafayette, IN 47907, USA
e-mail: poching@purdue.edu

Z. Tian
Regenstrief Center for Healthcare Engineering, Purdue University,
West Lafayette, IN 47907, USA
e-mail: tianz@purdue.edu

J. Daggy
Department of Biostatistics, Indiana University School of Medicine, Indianapolis,
IN 46202, USA
e-mail: jdaggy2@iupui.edu

L. Zhang
School of Statistics, Purdue University, West Lafayette, IN 47907, USA
e-mail: lingsong@purdue.edu

M. Lawley (✉)
Weldon School of Biomedical Engineering, Purdue University,
West Lafayette, IN 47907, USA
e-mail: malawley@purdue.edu

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_10, © Springer Science+Business Media New York 2013

are unique. Perhaps the most visible and difficult problem is the prevalence of nonattendance or no-show among scheduled patients. No-show rates of 20% are typical in ambulatory settings and can exceed 50% in some environments. Further, no-show behavior is often worst among those most in need of care due to socio-economic challenges, and it has far-reaching effects on clinic efficiency, patient outcomes, and healthcare costs. Clinic efficiency suffers when scheduled patients do not attend and resources are underutilized or when overbooked schedules result in clinic congestion, provider overtime and burnout, and patient dissatisfaction. Health outcomes suffer when no-show patients miss the care they need or when patients cannot get timely appointments because part of the schedule is filled with patients who will not attend. Healthcare costs rise when patients who miss appointments seek emergency care or experience hospital admissions that could have been avoided with proper ambulatory care. In fact, no-show behavior is much more than a scheduling annoyance; it negatively impacts all major healthcare policy areas. Thus, understanding no-show behavior is fundamental for delivering effective care.

Our objective in this chapter is to provide a blueprint for modeling no-show behavior. We believe that the most basic ability is to accurately estimate the probability that a given patient will keep a scheduled appointment. If this is well done, many possibilities emerge, ranging from patient-specific interventions such as identifying and targeting patients at risk for bad outcomes, to system and software design efforts such as developing and implementing effective appointment scheduling systems, to answering policy questions such as estimating the expected yearly national cost of no-shows.

In this chapter, we first provide a structured, representative literature review that will serve as a guide and provide entry points into the no-show research literature. We will then discuss practical data issues such as the types of data required and how data should be cleaned, maintained, and managed. We follow this with a discussion of the model building process, the most appropriate statistical techniques, and how to perform validation studies. We then present an example no-show model from our own research and discuss it in detail. Finally, we conclude by discussing practical implementation issues, summarizing thoughts, and directions for future research.

## 2 Literature Review

This section provides a structured and representative review of no-show literature up to 2011. Our aim is not to provide a comprehensive review, but to provide coverage sufficient to reveal the structure of the literature, that is, to illustrate the variety of perspectives, contexts, methods, and results that the literature holds. Further, we believe our review will serve as a guide for the interested reader who wants to explore the literature in more detail.

Throughout the literature, the term "no-show" indicates scheduled appointments that patients unexpectedly miss. This typically does not include canceled appointments. Terms from the literature that are equivalent to "no-show" include *visit non-adherence* (Alafaireet et al. 2010), *appointment breaking* (Bean and Talaga 1992), *nonattendance* (Cohen et al. 2008), *dropping out* (Deyo and Inui 1980), *missed appointment* (Karter et al. 2004), and *appointment failures* (Lindauer et al. 2009).

We searched the databases "PubMed" and "Academic Search Premier" using keywords "missed appointment," "no-show and appointment," "visit nonadherence," and "nonattendance and appointment." We considered only full-text articles that we could find electronically. We excluded the articles that were not written in English and that considered pediatric patients. We also excluded studies that consider mitigating the effect of no-show (e.g., overbooking). Based on our search, we identified more than 100 papers in total.

Table 10.1 provides the detailed listing and categorization. The literature is organized by ambulatory setting, which includes *Mental Health*, *HIV*, *Dialysis*, *Primary Care*, *Chronic Care/Diabetes*, and others. We further classify each paper with respect to its specific no-show-related topic. These topics include *Health Outcomes*, *Statistical Prediction Models*, *Interventions*, and *Self-Reported Reasons*, and they provide the basis for discussion in the following subsections. Also, we indicate those papers for which an overall no-show rate is reported. Thus, a researcher interested in no-show prediction models and overall no-show rates in dentistry could access the row labeled *Dental* and the columns labeled *Statistical Prediction Models* and *No-show rate* to find the most relevant sets of papers.

As part of our review, we identified several literature review papers. These are presented in Table 10.2. The first paper, (Deyo and Inui 1980), reviewed studies published in 1953–1979. At that time (1980), most of the published literature had focused on psychiatric and pediatric populations. The no-show rates ranged from 15% to 33% in adult clinics. The review identified the determinants of no-show, self-reported reasons for no-show, and interventions to reduce no-show. Determinants of no-show were classified according to patient features (both demographic and sociobehavioral), medical provider, disease or reason for appointment, patient–provider interaction, therapeutic regimen, medical facility, and administrative process. Most studies emphasized the demographic features, while other factors were less studied (Deyo and Inui 1980). The most frequently self-reported reasons were forgetting the appointment, not knowing about the appointment or misunderstanding. Other self-reported reasons were family problems, lack of transportation, time or work conflicts, and patient's health status (feeling better or feeling too sick). Among several interventions related to patient education and follow-up, appointment reminders were the most successful.

Bean and Talaga (1992) reviewed the literature published in 1977–1990. During those years, the studies used more sophisticated methods (multivariate statistical techniques) compared to simple descriptive methods of earlier studies, and most of the research focused on interventions to reduce no-show. The predictors of no-shows were classified as demographic, patient behavior (previous appointment keeping, psychosocial problems, and health belief), situational characteristics (seriousness of problem, referral source, day and time of appointment, and weather), and marketing

Table 10.1 Classification according to clinic/primary diagnosis

| Setting | Health outcomes | Statistical prediction models | Interventions | Self-reported reasons | No-show rate |
|---|---|---|---|---|---|
| Mental health | Murphy et al. 2011; Pang et al. 1996 | Alafaireet et al. 2010; Centorrino et al. 2001; Charupanit 2009; Compton et al. 2006; Killaspy et al. 1999; Kruse et al. 2002; Livianos-Aldana et al. 1999; Weinerman et al. 2003 | Blank et al. 1996; Hochstadt and Trybula 1980; Jayaram et al. 2008; Kluger and Karras 1983; LaGanga 2011; Lowe 1982; Rowett et al. 2010; Swenson and Pekarik 1988; Williams et al. 2008 | Defife et al. 2010; Killaspy et al. 1999; Neal et al. 2005; Peeters and Bayer 1999; Sparr et al. 1993 | Kruse et al. 2002; Leigh et al. 2009; Livianos-Aldana et al. 1999; Matas et al. 1992; Mitchell and Selmes 2007; Pang et al. 1996; Sparr et al. 1993 |
| Primary care | Bigby et al. 1984 | Bennett and Baxley 2009; Cashman et al. 2004; Daggy et al. 2010; Hamilton et al. 2002; Hurtado et al. 1973; Kopach et al. 2007; Lehmann et al. 2007; Moore et al. 2001 | Belardi et al. 2004; Bennett and Baxley 2009; Bundy et al. 2005; Cameron et al. 2010; Fischman 2010; George and Rubin 2003; Guse et al. 2003; Hershey et al. 1987; Johnson et al. 2007; Kopach et al. 2007; Kros et al. 2009; Leong et al. 2006; Mehrotra et al. 2008 | Lacy et al. 2004; Martin et al. 2005 | Cayirli et al. 2006, 2008; Jonas 1973; Liu et al. 2010; Moore et al. 2001; Xakellis and Bennett 2001 |

| | | | | | |
|---|---|---|---|---|---|
| Chronic care/diabetes | Davidson et al. 2000; Karter et al. 2001, 2004; Rhee et al. 2005; Samuels et al. 2008; Schectman et al. 2008 | Bowser et al. 2010; Ciechanowski et al. 2006; Griffin 1998; Karter et al. 2004 | Bowser et al. 2010; Griffin 1998; Hardy et al. 2001; Liew et al. 2009; Smith et al. 1986 | Mirotznik et al. 1998 | |
| HIV | Mugavero et al. 2009a,b; Murphy et al. 2011 | Catz et al. 1999; Kunutsor et al. 2010a; Mugavero et al. 2007, 2009a,b | Andersen et al. 2007; Cavaleri et al. 2010; Kunutsor et al. 2010b | Bakken et al. 2000; Park et al. 2008; Sarnquist et al. 2011; Tuller et al. 2010 | Yehia et al. 2008 |
| Colorectal | | | | Corfield et al. 2008 | |
| Dental | | Lindauer et al. 2009; Mandall et al. 2008 | Can et al. 2003 | Herrick et al. 1994; Lin 2009 | Can et al. 2003 |
| Dermatology | | Cohen et al. 2008 | | | |
| Dialysis | Denhaerynck et al. 2007; Obialo et al. 2008; Saran et al. 2003 | Obialo et al. 2008; Saran et al. 2003 | | | Rocco and Burkart 1993 |
| Endocrinology | | Tseng 2010 | | | |
| Gastroenterology | | | | Murdock et al. 2002 | |
| Multispecialty | | Dove and Schneider 1981; Parikh et al. 2010 | Bech 2005; Milne et al. 2006 | | Bech 2005; Rockart and Hofmann 1969 |
| Neurology or orthopaedic | | Collins et al. 2003 | | Collins et al. 2003 | |
| Obstetric | | | | Campbell et al. 2000 | |
| Oncology | | | | Gany et al. 2011 | |
| Ophthalmology | | | | Potamitis et al. 1994 | |
| Tuberculosis | | | Tanke and Leirer 1994 | | |
| Urology | | | Snow et al. 2009 | | |
| Vascular lab | | | Satiani et al. 2009 | | |

**Table 10.2** Literature review papers

| Year | Reference | Setting | Focus | Number of papers reviewed |
|---|---|---|---|---|
| 1980 | Deyo and Inui (1980) | General | No-show predictors, self-reported reasons, interventions | 87 references |
| 1992 | Bean and Talaga (1992) | General | No-show predictors, self-reported reasons, interventions | 51 references |
| 1992 | Macharia et al. (1992) | General | Interventions | 38 references |
| 1998 | Garuda et al. (1998) | General | No-show predictors, self-reported reasons, interventions | 50 references |
| 1998 | Griffin (1998) | Diabetes | No-show predictors, self-reported reasons, interventions | 136 references (59 reviewed in full) |
| 2003 | George and Rubin (2003) | Primary Care | No-show predictors, self-reported reasons, interventions | 55 references (31 reviewed in full) |
| 2007 | Denhaerynck et al. (2007) | Dialysis | Health outcomes | 88 references (17 reviewed in full, 8 on attendance) |
| 2010 | Bowser et al. (2010) | Diabetes Mental Health | No-show predictors, self-reported reasons, health outcomes | 64 references (50 reviewed in full, 9 on attendance) |
| 2010 | Rowett et al. (2010) | Mental Health | Interventions | 88 references (4 reviewed in full) |

mix (service, communication, price, and transportation). The interventions included appointment reminders, health belief interventions (educating the patient about the disease), and incentives. Appointment reminders were the most commonly studied and most effective interventions. The authors proposed strategies to reduce no-shows such as minimizing the time to appointment, changing health beliefs of the patient, maintaining patient satisfaction, using reminders, reducing financial costs of an appointment, reducing transportation costs, and conducting follow-up appointments by phone. They also proposed overbooking, estimation of no-show probabilities, block booking of high-probability no-show patients, and physician activity scheduling to mitigate the effect of no-shows.

Macharia et al. (1992) reviewed the interventions used to improve patient compliance with screening, referrals, and appointments for counseling or administration of medications. The studies that consider appointments for ongoing care were not included. The interventions were classified as cuing (appointment reminders by phone or mail), reducing perceived barriers (orientation statement, automatic appointment, and clerical assistance), and increasing patient motivation (enhanced information and health belief model card). The most commonly tested and successful interventions were reminders.

Garuda et al. (1998) provided background information on the predictors of no-show (waiting time, payer type, number of visits, previous no-show behavior, referral source, day and time of appointments, transportation, education and socioeconomic status, age, gender, race, and personal illness) and interventions that had been empirically tested (reminders, education, incentives, orientation, automatic rescheduling, self-appointment scheduling, and overbooking). Once again, appointment reminders were the most studied and successful intervention. The authors also proposed a marketing approach to choose the best strategies that address the particular needs and characteristics of a patient population. The marketing approach included segmentation of the population into distinct groups, understanding the most common reasons for no-show in each group, developing a plan to use the most effective interventions for targeted patient groups, and testing the plan on a sample to evaluate its effect.

Griffin (1998) focused on a diabetic patient population and reviewed the predictors of no-shows and interventions to reduce no-shows. The author provided several factors (patient sociodemographic features, patient clinical features, appointment characteristics, environmental factors, and provider–patient relations) that show positive or negative association with no-shows. The interventions related to changing patient appointment keeping behavior such as reminders, changes in organization and delivery of care, and patient–provider relationship are discussed. The author mentioned that interventions targeting the delivery of healthcare and the patient–provider relationship were more effective because they were stronger predictors of no-shows. This study identified only eleven papers about diabetes and emphasized the need for further research on chronic care patients due to increasing number of patients and the strong association of no-show with the adverse patient outcomes.

George and Rubin (2003) performed a systematic review to identify the characteristics of no-show patients, self-reported reasons for no-show, and impact of appointment systems on no-show. They also discussed the effectiveness of interventions such as appointment reminders, reminding patients to cancel, and orientation statements in the existing studies. Even though the included studies were shown to reduce no-show, the authors discussed the issue of having accurate records of phone numbers to make reminder calls, the patient's access to phones to make cancellation calls, and the practice's ability to handle cancellation calls. They presented open-access scheduling and demand management as new approaches to address the no-show problem.

Denhaerynck et al. (2007) reviewed the literature on nonadherence (fluid, dietary, medication, and appointment nonadherence) to hemodialysis treatment and effect of nonadherence on health outcomes. Nonadherence was shown to be associated with higher morbidity and mortality in four studies that consider appointment nonadherence. The authors concluded that patient behavior should be considered to achieve adequate treatment outcomes.

Bowser et al. (2010) reviewed the literature to determine the relationship between the diagnosis of diabetes and depression and missed appointments in a low-income, uninsured adult population. The existing studies identified depression as an important predictor of no-shows and showed the effect of no-show on diabetes patient health outcomes (poor glycemic control).

Rowett et al. (2010) reviewed the literature to identify the effect of prompts (phone calls and orientation letters) on attendance for severe mental illness patients. They found only four studies that showed the effectiveness of prompts for severe mental illness patients. However, there were no statistical differences between the intervention groups and control groups based on the reported confidence intervals. The authors mentioned that there should be more studies with larger sample sizes. They also proposed using automated systems for contacting people and using other means of reminders such as text messages and emails.

The following sections provide more detailed information about additional papers that identify self-reported reasons for no-shows using interviews and surveys (*Self-Reported Reasons*), studies that use interventions to reduce no-shows (*Interventions*), studies that develop statistical prediction models to identify predictors of no-show (*Statistical Prediction Models*), and studies that identify the relationship between no-show and health outcomes (*Health Outcomes*).

## *2.1 No-Show Rates Reported in the Literature*

This section presents the variety of no-show rates reported in our literature sample. Our purpose is to give the reader some idea of the pervasiveness and variability of this problem.

No-show rates were taken from 62 journal articles. The mean across all studies was 23.8%, with rates of 24.3% for Asian studies, 14.9% for European studies, and 27.1% for North American studies (see Figs. 10.1–10.3 for histograms). Further, Fig. 10.4 provides no-show rates for a variety of populations, including mental health and primary care. It is remarkable that some North American clinics suffer no-show rates of 48–64%.

## *2.2 Self-reported Reasons for No-Show*

We identified eighteen papers that investigate self-reported reasons for no-show. Eight papers covered studies in US clinics (Blankson et al. 1994; Bowser et al. 2010; Campbell et al. 2000; Defife et al. 2010; Gany et al. 2011; Lacy et al. 2004; Sarnquist et al. 2011; Sparr et al. 1993), seven papers in European clinics (Corfield et al. 2008; Herrick et al. 1994; Killaspy et al. 1999; Martin et al. 2005; Neal et al. 2005; Peeters and Bayer 1999; Potamitis et al. 1994), and three papers in other countries (Collins et al. 2003; Park et al. 2008; Tuller et al. 2010). Further, mental health was represented most often with five papers, followed by HIV, which had three (see Table 10.1). These papers used either written surveys or interviews to identify the

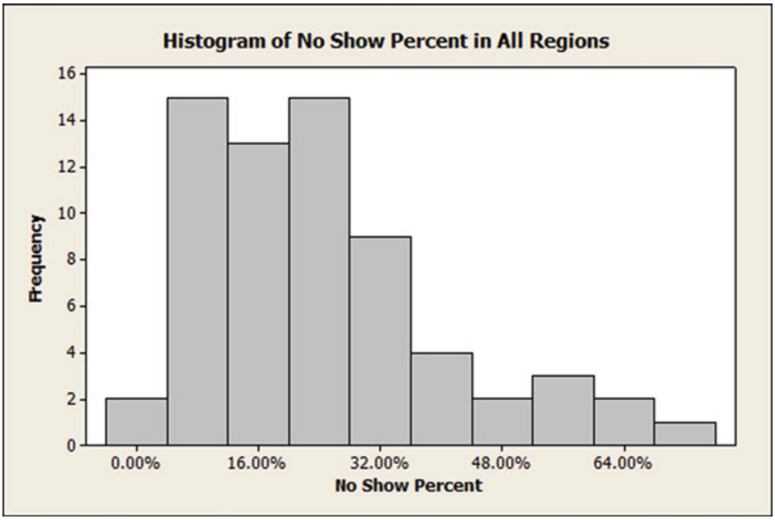


**Fig. 10.1** Histogram of reported no-show percentages in all regions based on results from 62 articles

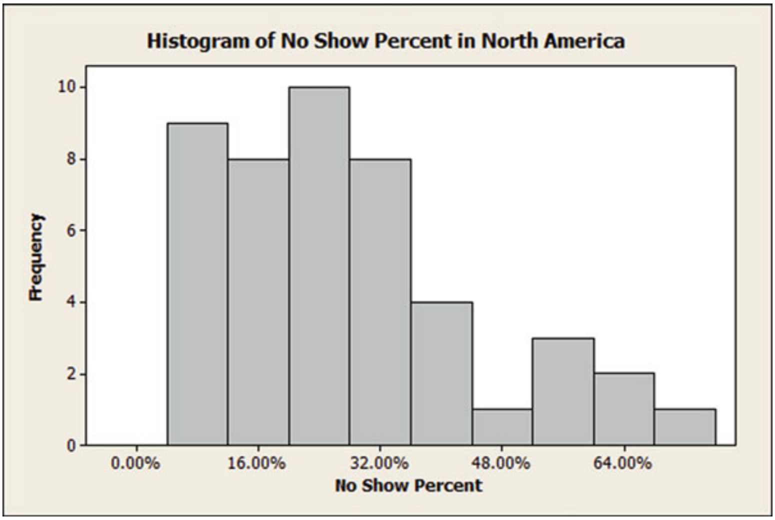


**Fig. 10.2** Histogram of reported no-show percentages in North America based on results from 43 articles

reasons for no-show. Special populations studied included rural HIV in both the USA and Africa (Sarnquist et al. 2011; Tuller et al. 2010), Chinese immigrants in the USA (Gany et al. 2011), and women's high-risk obstetrics (Blankson et al. 1994; Campbell et al. 2000). Self-reported reasons include patient-related factors, scheduling system problems, and environmental and financial factors.

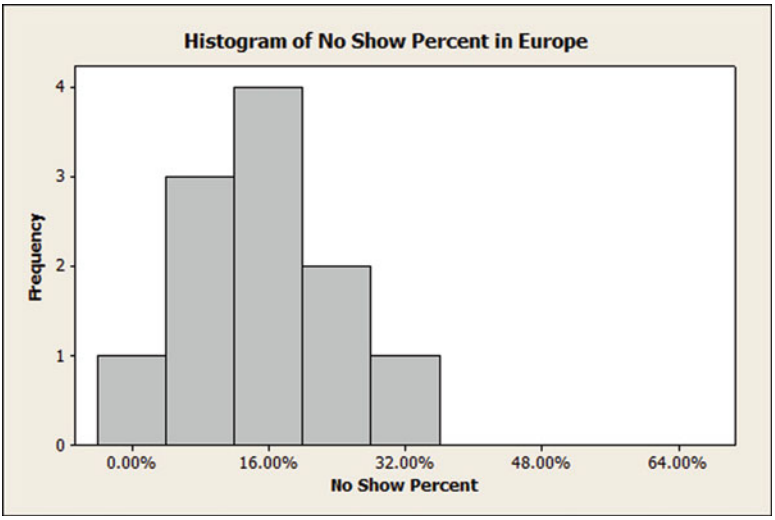


**Fig. 10.3** Histogram of reported no-show percentages in Europe based on results from 10 articles

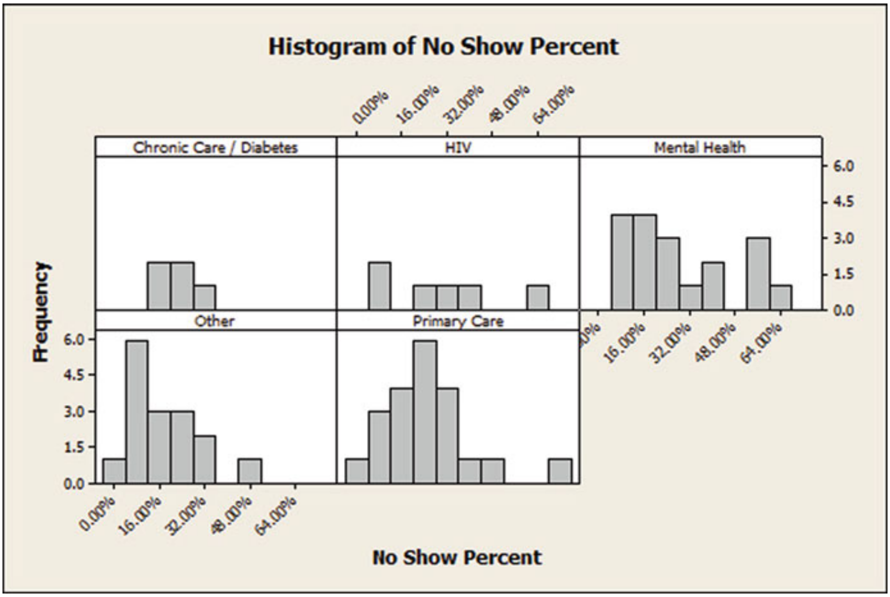


**Fig. 10.4** Histogram of reported no-show percentages by population: chronic care/diabetes (5 articles), HIV (5 articles), mental health (17 articles), other (14 articles), and primary care (21 articles). Patient populations studied in less than 5 articles are grouped as "other," and it includes dental, dermatology, dialysis, endocrinology, multispecialty, neurology, obstetric, orthopedic, tuberculosis, urological, and vascular

Patient-related reasons include forgetting the appointment, other competing priorities or conflicts, and the patient's health status. The most commonly reported reason for no-show is "forgetting the appointment" (Blankson et al. 1994; Campbell et al. 2000; Collins et al. 2003; Herrick et al. 1994; Killaspy et al. 1999; Martin et al. 2005; Neal et al. 2005; Park et al. 2008; Potamitis et al. 1994; Sparr et al. 1993). Competing priorities include work schedules and taking care of a family member (Blankson et al. 1994; Campbell et al. 2000; Defife et al. 2010; Gany et al. 2011; Martin et al. 2005; Neal et al. 2005). Reasons related to patient health status include being physically or mentally ill (Bowser et al. 2010; Collins et al. 2003; Defife et al. 2010; Gany et al. 2011; Lacy et al. 2004; Peeters and Bayer 1999; Potamitis et al. 1994; Sarnquist et al. 2011) and feeling better and not needing the appointment any more (Corfield et al. 2008; Killaspy et al. 1999; Murdock et al. 2002; Potamitis et al. 1994).

Scheduling problems were also reported as reasons for nonattendance in eight studies (Collins et al. 2003; Corfield et al. 2008; Gany et al. 2011; Herrick et al. 1994; Lacy et al. 2004; Martin et al. 2005; Peeters and Bayer 1999; Potamitis et al. 1994). The problems related to scheduling system include difficulty in getting an appointment with the primary care provider, difficulty in canceling the appointment, not receiving an appointment card, losing the appointment card, and wrong information about the date and time of the appointment. As George and Rubin (2003) mention, appointment systems can become barriers to healthcare when waiting times to appointment day are long. As the waiting time increases, no-show rates increase.

Transportation is another commonly reported reasons for no-show (Blankson et al. 1994; Campbell et al. 2000; Gany et al. 2011; Neal et al. 2005; Park et al. 2008; Peeters and Bayer 1999; Sarnquist et al. 2011; Sparr et al. 1993; Tuller et al. 2010). Even though financial problems are reported only in two studies (Gany et al. 2011; Tuller et al. 2010), other performance measures such as unavailability of transportation can be thought as proxy measures for socioeconomic status. As noted by Campbell et al. (2000), patients are constantly challenged with very real unfilled needs related to housing, food, employment, education, and transportation. Inadequate social support systems and difficult societal issues adversely affect patient motivation and attitudes toward care.

Table 10.3 provides a summary of all studies that identified most commonly reported reasons for no-shows. While many reasons were given for missing appointments, the most common self-reported reasons were "forgetting the appointment" (9 papers), "transportation" (8 papers), "mentally or physically unwell" (8 papers), and "scheduling system problems" (8 papers). Limitations cited for these studies include the fact that reasons for no-show were self-reported (not observed) so that independent verification was not possible. Further, there were concerns about sampling or "no-response" bias since not all selected patients could be reached for interview and the response rates for surveys were sometimes low.

Some researchers believe that objective reasons, such as lack of transportation, are cited in lieu of more difficult problems such as depression, resentment, and fear. Defife et al. (2010) claim that keeping a clinic appointment is the result of

**Table 10.3** Most common self-reported reasons

| Self reported reasons | Number of papers | Papers |
|---|---|---|
| Forgot appointment | 10 | Blankson et al. 1994; Campbell et al. 2000; Collins et al. 2003; Herrick et al. 1994; Killaspy et al. 1999; Martin et al. 2005; Neal et al. 2005; Park et al. 2008; Potamitis et al. 1994; Sparr et al. 1993 |
| Conflicts with appointment | 6 | Blankson et al. 1994; Campbell et al. 2000; Defife et al. 2010; Gany et al. 2011; Martin et al. 2005; Neal et al. 2005 |
| Transportation | 9 | Blankson et al. 1994; Campbell et al. 2000; Gany et al. 2011; Neal et al. 2005; Park et al. 2008; Peeters and Bayer 1999; Sarnquist et al. 2011; Sparr et al. 1993; Tuller et al. 2010 |
| Physically or mentally unwell | 8 | Bowser et al. 2010; Collins et al. 2003; Defife et al. 2010; Gany et al. 2011; Lacy et al. 2004; Peeters and Bayer 1999; Potamitis et al. 1994; Sarnquist et al. 2011 |
| Scheduling system problems | 8 | Collins et al. 2003; Corfield et al. 2008; Gany et al. 2011; Herrick et al. 1994; Lacy et al. 2004; Martin et al. 2005; Peeters and Bayer 1999; Potamitis et al. 1994 |
| Perceived disrespect | 2 | Lacy et al. 2004; Martin et al. 2005 |
| Bad weather | 2 | Killaspy et al. 1999; Neal et al. 2005 |
| Financial problems | 2 | Gany et al. 2011; Tuller et al. 2010 |

a combination of several factors that include personal meanings of the symptoms, resulting emotions, anticipated consequences, and the perceived relationship of the patient with the provider and the clinic staff. As Blankson et al. (1994) point out, the patient might not perceive the benefits of the appointment and thus places no emphasis on keeping it. Lacy et al. (2004) urge providers to understand the patient's fears and educate the patient about what to expect and why treatment is important. Campbell et al. (2000) emphasize the need for providers to develop a deeper understanding of a patient's daily priorities.

## *2.3 Interventions to Reduce No-Shows*

Interventions to reduce no-shows include appointment reminders, patient education, follow-up after a no-showed appointment, open-access scheduling, and lean process improvement methods. Tables 10.4 and 10.5 show the no-show rates before and after implementation of several interventions in 26 studies.

**Table 10.4** Appointment reminders, orientation/information packages, ancillary services, and follow-up

| | | No-show rate (sample size) | |
|---|---|---|---|
| Study | Intervention | No intervention | Intervention |
| Can et al. (2003) | Mail (15 days before), patient returned a confirmation | 23.3% ($n = 116$) | 12.2% ($n = 82$) |
| | Mail (15 days before), patient did not return a confirmation | | 33.3% ($n = 33$) |
| Hardy et al. (2001)[a] | Information package (2 weeks before) | 15% ($n = 1{,}336$) | 7.3% ($n = 178$) |
| | Information package (2 weeks before) plus phone (1 week before) | | 1.4% ($n = 147$) |
| Hochstadt and Trybula (1980)[a] | Phone (1 day before) | 55% ($n = 22$) | 9% ($n = 22$) |
| | Mail (3 days before) | | 32% ($n = 22$) |
| | Phone (3 days before) | | 32% ($n = 22$) |
| Jayaram et al. (2008) | Letter | 27% ($n = 1{,}074$) | 17% ($n = 1{,}433$) |
| Kluger and Karras (1983)[a] | Orientation (after appt. is made) | 56% ($n = 25$) | 28% ($n = 25$) |
| | Phone (1 day before) | | 54% ($n = 50$) |
| | Orientation plus phone | | 39% ($n = 41$) |
| Leong et al. (2006) | Text message | 51.9% ($n = 335$) | 41% ($n = 329$) |
| | Mobile phone (1–2 days before) | | 40.4% ($n = 329$) |
| Liew et al. (2009) | Text message (1–2 days before) | 23% ($n = 309$) | 13.7% ($n = 314$) |
| | Phone (1–2 days before) | | 15.6% ($n = 308$) |
| Milne et al. (2006) | SMS, new patient, before partial booking | 15.3% ($n =1{,}770$) | 9.8% ($n = 192$) |
| | SMS, new patient, after partial booking | 5.2% ($n = 1{,}971$) | 3.4% ($n = 816$) |
| | SMS, return patient, before partial booking | 17.4% ($n = 10{,}009$) | 16.4% ($n = 1{,}642$) |
| Satiani et al. (2009) | Automated call | 8.9% ($n = 4{,}118$) | 5.9% ($n = 4{,}648$) |
| Smith et al. (1986) | Information package, appointment reminders, intense follow-up of no-show for rescheduling | 24.4% ($n = 430$) | 21.5% ($n = 429$) |

(continued)

**Table 10.4** (continued)

| Study | Intervention | No-show rate (sample size) | |
|---|---|---|---|
| | | No intervention | Intervention |
| Swenson and Pekarik (1988)[a] | Mail (1 day before) | 43% ($n = 30$) | 33% ($n = 30$) |
| | Mail (3 days before) | | 37% ($n = 30$) |
| | Orientation (1 day before) | | 17% ($n = 30$) |
| | Orientation (3 days before) | | 27% ($n = 30$) |
| Tanke and Leirer (1994) | Automated call (Evening before) | 47% ($n = 456$) | 38% ($n = 1{,}552$) |
| Andersen et al. (2007)[b] | Transportation | 63.9% missed appointments | 28.6% missed appointments ($n = 61$) |
| | Transportation and case management | 60.8% missed appointments | 34.2% missed appointments ($n = 51$) |
| Blank et al. (1996) | Mail follow-up | 32.9% ($n = 85$) | 25.0% ($n = 32$) |
| | Phone follow-up | | 28.6% ($n = 28$) |
| | Home visit | | 0% ($n = 5$) |
| Guse et al. (2003) | Exit interview/education | 21.7% ($n = 297$) | 16.5% ($n = 146$) |
| Kunutsor et al. (2010b) | Follow-up by mobile phone (calls or text messages) | 11% ($n = 560$) | 2.3% (did not return) |
| Lowe (1982) | Follow-up letter (asking to reschedule) | | No difference |
| | Follow-up letter (automatically rescheduled appointment) | | Higher no-show |
| | No follow-up | | No difference |

[a] First appointment
[b] Self-reported no-show

**Table 10.5** Open-access scheduling and lean methodology to reduce waiting times

| | | No-show rate (sample size) | |
|---|---|---|---|
| Study | Intervention | No intervention | Intervention |
| Belardi et al. (2004) | Traditional access vs. advanced access | 9.23% | 6.67% |
| Bennett and Baxley (2009) | Advanced access | 20–25% | 17.6–23.7% |
| Bundy et al. (2005) | Open access | 16% ($n = 1{,}633$) | 11% ($n = 3{,}248$) |
| Cameron et al. (2010) | Open access | 3.33% ($n = 21{,}838$) | 1.89% ($n = 21{,}819$) |
| Fischman (2010) | Lean methodology to improve efficiency and reduce in-clinic waiting times | 30.8% | 26.0% |
| LaGanga (2011) | Lean process improvement | 13.85% ($n = 816$) | 2.20% ($n = 910$) |
| Mehrotra et al. (2008) | Open access (5 practices) | 3–18% ($n = 49{,}603$) | 3–17% ($n = 115{,}167$) |
| Snow et al. (2009) | Improved access | 5% | 4–5% |
| Williams et al. (2008)[a] | Advanced access | 52% | 18% |

[a] First appointment

One of the most commonly self-reported reasons for no-show is forgetting the appointment, as discussed in Sect. 2.2. Different methods such as mailing letters or postcards (Can et al. 2003; Hochstadt and Trybula 1980; Jayaram et al. 2008; Swenson and Pekarik 1988), phone calls (Hardy et al. 2001; Hochstadt and Trybula 1980; Kluger and Karras 1983; Leong et al. 2006; Liew et al. 2009; Satiani et al. 2009; Tanke and Leirer 1994), and text messages (Leong et al. 2006; Liew et al. 2009; Milne et al. 2006) are used as appointment reminders. According to the studies that compare letters and phone calls, phone calls are more effective in reducing no-show rates (Hardy et al. 2001; Hochstadt and Trybula 1980). However, the timing of the intervention is also an important factor. As the time of intervention gets closer to the appointment time, the reduction in no-show rate increases (Hochstadt and Trybula 1980; Swenson and Pekarik 1988). The applicability of text messages is limited because not all patients have cell phones (Milne et al. 2006).

Orientation and information packages give detailed information about when and where to come, where to park, what to bring, who the patient will see, and what will occur during the appointment. There are a few studies that show the importance of providing information and orientation packages before the appointment on reducing no-show rates (Hardy et al. 2001; Kluger and Karras 1983; Swenson and Pekarik 1988). This method is mainly used for the initial appointments to reduce the uncertainty and possible fear of patients. In Guse et al. (2003), patient education is performed in the form of an exit interview after the appointment. The interview included visit debriefing, written patient information where appropriate, and review of clinic policies. Guse et al. (2003) showed that the exit interview improved attendance in subsequent visits.

Follow-up after a no-show appointment is used to reduce no-show rates for subsequent visits. Letters, phone calls, home visits, and text messages are used as follow-up strategies (Blank et al. 1996; Guse et al. 2003; Kunutsor et al. 2010b; Lowe 1982). Lowe (1982) reported that the follow-up letter asking to reschedule does not reduce no-show rates, which shows the importance of patient agreement on scheduling an appointment at a convenient time.

Waiting time for an appointment and in-clinic waiting times increase patient dissatisfaction and might lead to perceived disrespect. Open-access scheduling and lean methods are used to minimize waiting times. "Open access" (also known as "same-day scheduling" or "advanced access") aims to reduce waiting time by providing appointments within a day or two (Murray and Tantau 2000). Even though there are studies that show the effectiveness of open-access scheduling in reducing no-shows (Bundy et al. 2005; Cameron et al. 2010; Williams et al. 2008), implementation of open-access scheduling does not always reduce no-show rates significantly (Belardi et al. 2004; Bennett and Baxley 2009; Mehrotra et al. 2008; Snow et al. 2009). In a recent survey on open-access scheduling by Rose et al. (2011), it was shown that open-access scheduling was not effective in reducing no-show rates significantly for practices with lower baseline no-show rates.

Lean methodology is implemented to improve efficiency and reduce waiting time (Fischman 2010; LaGanga 2011). Fischman (2010) implemented lean methodology to improve workflow processes, and LaGanga (2011) implemented several changes such as providing intake assessment on the same day and realigning appointments to better-attended days. Both studies show the effect of lean methodology in reducing no-show rates.

Although several studies illustrate the effectiveness of interventions in reducing no-show, patient no-show remains a widespread problem. Garuda et al. (1998) mention that one of the most important reasons for current failure of institutions to sufficiently resolve the no-show issue is the sparse number of studies that match specific causes of no-show behavior with appropriate solutions to these problems. For example, transportation was one of the most commonly reported reasons for no-show. However, Andersen et al. (2007) is the only study that evaluates the effectiveness of providing transportation in reducing no-shows. Another problem is limited resources to implement a full range of appropriate interventions. Research is needed to determine the optimally effective mix of no-show interventions for a given clinic environment and patient population.

## *2.4 No-Show Predictors*

This section discusses patient, provider, and scheduling factors that are associated with no-show behavior. Factors identified below were from studies with adequate sample sizes and robust experimental and statistical methods.

Patients who are middle aged or older are less likely than younger patients to miss appointments with their providers (Cashman et al. 2004; Cohen et al. 2008; Daggy et al. 2010; Hurtado et al. 1973; Lehmann et al. 2007; Parikh et al. 2010; Tseng 2010). Some studies reported that patients younger than middle age are twice as likely to no-show compared to older subjects (Cashman et al. 2004; Daggy et al. 2010). For example, one study of more than 3,000 patients in a primary care clinic reported that 31% of patients younger than 50 years of age did not show up to their medical appointments compared to 17% of patients aged 51–60, and 9% of subjects older than 70. Two studies distinguished no-show rates for younger (aged 18–25) patients and found that younger patients were significantly less likely to no-show than patients aged 26–50 (Cashman et al. 2004; Kruse et al. 2002).

Most studies that considered whether males and females differed in no-show rates did not find a significant difference in rates between the genders (Cashman et al. 2004; Catz et al. 1999; Cohen et al. 2008; Lehmann et al. 2007; Livianos-Aldana et al. 1999; Moore et al. 2001; Mugavero et al. 2009b; Weinerman et al. 2003). Of the few studies that did find statistically significant differences, most found that females were less likely to no-show (Hurtado et al. 1973; Lindauer et al. 2009), although the differences in rates were small. For example, in a study of over 5,000 health maintenance organization patients, 17.2% of males did not show up to their medical appointment compared to 15.7% of females (Hurtado et al. 1973).

A few studies considered whether no-show rates are associated with having family members or being married (Alafaireet et al. 2010; Daggy et al. 2010; Hurtado et al. 1973). Typically, having family members or being married is associated with lower no-show rates. For example, a study of veterans at a Midwestern outpatient facility reported that the no-show rate for married patients was 9.5% compared to 22.7% for non-married patients (Daggy et al. 2010). Large families, however, may worsen no-show rates. A study of over 5,000 patients enrolled in a health maintenance organization found that appointment failures were greater for larger families. For example, no-show rates were 11.3% for a family of size two, but they were 22.6% for a family size of six or more members (Hurtado et al. 1973).

In the USA, patients from minority racial or ethnic groups were more likely to no-show than whites (Alafaireet et al. 2010; Bennett and Baxley 2009; Catz et al. 1999; Kruse et al. 2002; Moore et al. 2001; Mugavero et al. 2009b). For example, a study of attendance to the first appointment at a psychiatric clinic revealed that Hispanics were less likely to attend their first appointment compared whites (OR = 0.48; 95% CI = 0.23–0.99) (Kruse et al. 2002). An odds ratio (OR) of 0.48 indicates that Hispanics are 48% more likely to miss their first appointment than are whites. If the 95% confidence interval (CI) about the odds ratio includes the value one, then the odds of the outcome is not significantly different between the two categories being compared. Similar trends are found among other patient populations including HIV patients (Catz et al. 1999; Mugavero et al. 2009b) and primary care patients (Bennett and Baxley 2009; Moore et al. 2001). The finding that minority race and ethnic status is associated with greater likelihood of no-showing was also consistent across studies that statistically controlled for financial and insurance status (Bennett and Baxley 2009; Kruse et al. 2002; Mugavero et al. 2009b).

Patients without health insurance are more likely to no-show than patients with health insurance (Alafaireet et al. 2010; Bennett and Baxley 2009; Kruse et al. 2002). For example, one study of primary care patients found that self-pay patients were nearly 40% more likely to no-show (OR = 1.38; 95% CI = 1.23–1.54) (Bennett and Baxley 2009). Similarly, a high patient co-payment is associated with higher no-show rate. A study of outpatient diabetics found that patient co-payments exceeding ten dollars were associated with a 30% increase in no-show rate (OR = 1.20; 95% CI = 1.2–1.4) (Karter et al. 2004). Not surprisingly, a study of orthodontic patients found that those patients whose payments were overdue or with a collection agency were much more likely to miss their appointments (Lindauer et al. 2009).

The greater the length of the interval between the day the appointment is made and the actual appointment day, the greater the risk for no-show (Alafaireet et al. 2010; Cohen et al. 2008; Compton et al. 2006; Daggy et al. 2010; Hamilton et al. 2002; Livianos-Aldana et al. 1999). For example, no-show probability for the first appointment after psychiatric hospitalization increased by 4% (OR = 1.04; 95% CI = 1.01–1.07) for each day between time the appointment was made and the actual day of the appointment (Compton et al. 2006). A similar result was found for primary care patients. A study of veteran outpatients revealed that patients scheduled more than 2 weeks in advance were more than twice as likely to no-show (OR = 2.68; 95% CI = 1.90–3.85) (Daggy et al. 2010). Other appointment characteristics such as time of day and day of the week were inconsistently associated with no-show status (Alafaireet et al. 2010; Cohen et al. 2008; Livianos-Aldana et al. 1999; Tseng 2010). A few studies concluded no-shows were more likely to occur for winter appointments than appointments scheduled during other seasons (Bennett and Baxley 2009; Daggy et al. 2010). For example, Daggy et al. found that among veteran outpatients, those with winter appointments were twice as likely to miss an appointment than those whose appointments were scheduled during other seasons (OR = 2.14; 95% CI = 1.67–2.73) (Daggy et al. 2010).

Patients with diseases that were likely to require hospitalization were less likely to no-show than patients with less illness severity (Daggy et al. 2010; Hurtado et al. 1973). In contrast, psychiatric diagnoses are associated with high no-show rates (Cashman et al. 2004; Charupanit 2009; Ciechanowski et al. 2006; Compton et al. 2006; Weinerman et al. 2003). Among patients with psychiatric diagnoses, those who had the greatest severity of psychiatric illness (Axis 4 diagnoses) were nearly twice as likely to no-show compared to other psychiatric patients (OR = 1.83; 95% CI = 1.00–3.34) (Compton et al. 2006). Similarly, those who needed psychiatric medications were significantly more likely to miss medical appointments (R = 4.32; 95% CI = 1.32–14.22) than those who did not need medications (Kruse et al. 2002).

Few studies considered whether provider characteristics are associated with no-show status. Of those who considered provider characteristics, several trends were evident. Clinicians with greater expertise (e.g., faculty versus resident) are less likely to have patients no-show (Bennett and Baxley 2009; Tseng 2010) as are clinicians whose practice focuses on continuity of care (Bennett and Baxley 2009; Compton et al. 2006).

## *2.5 No-Show and Health Outcomes*

While ample literature is available discussing predictors of no-show and interventions for no-show, there is little that describes the relationship between no-show and health outcomes. Research studying the association of no-show and health outcomes can be found across the following clinical populations: diabetic (Davidson et al. 2000; Karter et al. 2001, 2004; Rhee et al. 2005; Samuels et al. 2008; Schectman et al. 2008), dialysis (Denhaerynck et al. 2007; Obialo et al. 2008; Saran et al. 2003), human immunodeficiency virus (HIV) (Mugavero et al. 2009a,b; Murphy et al. 2011), primary care (Bigby et al. 1984), and psychiatric (Murphy et al. 2011; Pang et al. 1996).

To date, the no-show literature studying health outcomes consistently reveals that diabetic patients with greater rates of missed appointments were associated with higher glycosylated hemoglobin (A1C) levels indicating poorer glycemic control (Davidson et al. 2000; Karter et al. 2001, 2004; Rhee et al. 2005; Samuels et al. 2008; Schectman et al. 2008). Poor glycemic control is associated with increased risk for long-term vascular complications such as coronary disease, heart attack, stroke, heart failure, kidney failure, blindness, and neuropathy (American Diabetes Association 2010). In addition to elevated A1C levels, Samuels et al. (2008) found those diabetic patients least adherent with their medical appointments had significantly higher mean systolic blood pressure.

Of the research studying no-shows in the dialysis population, two articles discussed that missing at least one dialysis session a month was associated with a 25–30% increased risk of mortality (Denhaerynck et al. 2007; Saran et al. 2003). In conjunction with increased mortality, Saran et al. (2003) showed that skipping one or more dialysis sessions a month was associated with 13% increased risk of hospitalization. One article illustrates that patients dialyzed on a Tuesday, Thursday, Saturday schedule missed more appointments and had higher morbidity than patients dialyzed on a Monday, Wednesday, Friday schedule; however, these findings were found to be not statistically significant.

Mugavero et al. (2009a) in discussing no-show and associated health outcomes in the HIV population revealed that appointment nonadherence was significantly associated with antiretroviral therapy failure. Mugavero et al. (2009b) showed that mortality rates were more than two times higher for patients with missed appointments compared to patients who attended all scheduled appointments during the first year of care.

One article describes no-show and health outcomes in the primary care population. Bigby et al. (1984) examined 200 primary care patients finding that no hospitalizations or deaths could be directly attributed to a missed appointment. This article was published in 1984, and more current literature regarding no-shows and health outcomes in the primary care population is needed.

Pang et al. (1996) examined the association of missed appointments in a psychiatric outpatient setting in Hong Kong and found that nonattendance was significantly associated with mortality.

In summary, the literature shows that missed appointments are associated with poorer glycemic control in diabetic patients, increased risk for mortality and hospitalization in dialysis patients, increased antiretroviral therapy failure and higher mortality rates in HIV patients and increased mortality in psychiatric patients. It is apparent from the scarcity of literature examining the relationship between no-show and health outcomes that more research is needed with diverse patient populations in a variety of healthcare settings.

## 3 Data and Methodology

We now turn our attention to no-show modeling. Our intention is to provide the reader with a practical guide for obtaining and managing data and building and validating models.

Modeling no-show behavior of patients allows us to determine the associated patient and environment factors. This information is needed to identify which patients are at highest risk for not showing up to their medical appointments and to determine whether there are modifiable factors that can be targeted for interventions to reduce no-show rates. No-show modeling has the potential for informing the design of clinic interventions (e.g., advanced scheduling considering patient no-show probability) to mitigate the impact of patient no-show and improve clinic performance and patient outcomes.

### *3.1 Data Requirements*

It is becoming common for healthcare organizations to store data related to clinic operations. For example, scheduling is typically conducted using electronic scheduling systems that store information about appointments including the timing of the appointment, the type of appointment, the clinicians involved in the appointment, and whether the patient kept the appointment. Similarly, electronic billing systems store information about the services a patient received during an appointment, the diagnoses associated with the appointment, and basic demographic and insurance information about the appointment. Electronic medical records store data about the patient's medical history, current diagnoses, laboratory values, prescriptions ordered, and treatment plans. These sources include many of the variables that have been shown in prior research to be predictive of no-show behavior.

Data use should be guided by federal, state, local, and institutional guidelines and principles of fair use. It is outside of the scope of this chapter to discuss regulatory guidelines; however, we will state that it is the responsibility of each individual who uses clinic data to understand and adhere to existing guidelines (e.g., HIPAA, institution specific guidelines, internal review board). Fair use of data refers to using data only for purposes of improving clinic operations including improving performance to optimize patient outcomes. In the case of extracting data for no-show modeling, fair use indicates that the data will be used only for the purposes of devising and validating no-show models.

Selection of electronic data that are relevant to developing and validating no-show models first requires a basic understanding of reasons why patients are likely to no-show. We recommend beginning with a framework that describes the factors that should theoretically contribute to no-show status. For example, a commonly used conceptual framework for studying factors related to healthcare utilization is Andersen's behavioral model which has been used by many to predict future use of health services (Andersen and Newman 2005). That model describes predisposing, enabling, and need characteristics associated with healthcare utilization. This is one of many possible models to guide derivation of hypotheses about which patient, provider, clinic, environmental, and policy factors explain no-show behavior. We will not recommend a specific model to guide the derivation of hypotheses because the choice of a model should be influenced by characteristics of the population and the healthcare setting to be studied. Existing scientific literature should be systematically reviewed to provide validation of a priori hypotheses about factors associated with no-show behavior. Beginning with a strong conceptual framework that is validated (at least in part) by existing research avoids extracting data unrelated to the objective of creating an accurate and parsimonious no-show model.

Our prior work and the existing literature describe important factors for no-show modeling including predisposing characteristics (e.g., age, gender, race, marital status), enabling characteristics (e.g., insurance status), encounter type (e.g., new patient or follow-up), need characteristics (e.g., comorbidities, patient no-show history), and environmental characteristics (e.g., types of provider, appointment reminder system in place). Data representing these characteristics may be collected from clinic scheduling systems, patient medical records, and billing information systems. The data used to develop a no-show model should accurately represent the hypothesized patient and system characteristics related to no-show status. A thorough understanding of the data is developed from learning how the data are collected and coded. Typically, the data are collected by clinic personnel and entered into an electronic system. Data are usually coded according to clinic specified criteria. For example, appointment attendance status is usually coded as canceled, arrived, or no-show. Which code is applied to the patient's visit is determined by criteria set by clinic personnel (e.g., the designation of no-show is used when the patient did not show up within an hour of the scheduled appointment). Some electronic data systems include a coding manual in which each variable within the electronic system is specified and defined. The manual may also specify and define all possible values for each variable. Coding manuals are invaluable for understanding of the meaning of the data. Whether or not a coding manual is available, it is critical that the personnel involved in coding and entering data into electronic systems be interviewed about the processes and guidelines used to code each of the variables that will be considered in the no-show model.

Currently, most clinics do not have integrated scheduling, billing, and medical records systems. To link data across systems, an encrypted patient identification system (often referred to as a "crosswalk") must be created that allows each patient's scheduling, billing, and medical records information to be combined. Typically this crosswalk is created by clinic information technology personnel who use that crosswalk to combine data from all electronic resources. Extracted data should be in the form of a limited data set (U.S. Department of Health and Human Services, National Institute of Health 2007). A limited data set excludes sixteen identifiers that could be used to identify the individual associated with the data. Accurate no-show modeling does not depend on knowing the identity of the individuals contributing the information. Rather, it is the combination of data from all clinic patients that is needed to create a model that can be applied across patients. It is recommended that the data are stored on a secured server for secure data management. A secured server is typically password protected and housed within a fire-walled network environment with a limited number of users allowed to access it.

## *3.2 Data Analysis for No-Show Modeling*

Most no-show models consider whether or not the patient showed up to a specified appointment (e.g., the most recent appointment). Most models do not include cancellations because cancellations reflect a different type of patient behavior than no-shows and likely would not have the same predictors. Further, patient and clinic operations outcomes associated with cancellations are different than for no-shows. Logistic regression is a natural choice to model the association between multiple predictors on no-shows, since the response is a dichotomized variable (e.g., Hosmer and Lemeshow 2000). The remainder of this section reviews steps associated with constructing a logistic regression equation to model no-show probability.

### 3.2.1 Variable Selection

Examination of candidate variables for inclusion in a no-show model is the first analytic step in building a predictive model. All variables that are considered conceptually and empirically relevant to predicting no-show status should be examined for distributional characteristics. Distributional characteristics relevant to modeling no-show status include determining whether there are missing values for a variable. Although missing values can occur at random, typically they do not.

Clinic personnel involved in data entry may know reasons for missing values which will inform how missing values should be treated. Options include excluding the variable from analysis, recoding the missing values using agreed upon criteria or imputation techniques, or deleting cases with the missing value. We caution readers that many statistical packages use a default listwise deletion technique that excludes a subject from the analysis if there are missing values for one or more variables for the subject. This might result in biased model estimates. However, treated, missing values can affect results. Thus, interpretation of results must be in the context of how missing values may have biased findings.

We recognize that categorizing variables may reduce resolution of the association between a variable and the outcome. A reduced number of categories increases interpretability of the results and increases sample size within a category which in turn improves statistical power to detect differences between categories.

Graphing or listing the frequency for each possible value for a variable is critical for understanding the data. Extreme values should be examined to determine whether they reflect error in measurement or data entry. Clinic personnel can be extremely helpful in interpreting unusual values of variables. Some variables have infrequent values. When this occurs, variable distributions may be transformed. For example, age is a variable with a continuous set of response values. Few patients are aged 90 or greater, so it would be logical to categorize those aged 90 or above into one category. Typically categories are constructed so that a variable distribution is represented by the fewest categories needed to describe which individuals are at increased risk for the outcome. Determining how many categories to use and the numerical cutoffs for the categories should be guided by clinical relevance and the distribution of responses. For example, most no-show models categorize insurance type into private payer (e.g., Blue Cross), state payer (e.g., Medicaid), federal payer (e.g., Medicare), or self-pay. The clinical relevance of these categories reflects the patient's out of pocket cost of attending the appointment. The distributional relevance of these categories is that typically there are too few subjects represented by specific types of health insurance to have adequate statistical power to detect risk for each type of health insurance. We recognize that categorizing variables may reduce resolution of the association between a variable and the outcome. A reduced number of categories increases interpretability of the results and increases sample size within a category which in turn improves statistical power to detect differences between categories.

Selection of independent (predictor) variables is influenced by relationships between independent variables. Inclusion of highly related independent variables can result in unstable estimates of the regression coefficients and their significance and therefore affect interpretation of model results. To avoid this, we suggest considering both conceptually and statistically whether independent variables are highly associated. It is logical to begin by considering whether independent variables are conceptually related. For example, income and insurance status are conceptually related. If one were to categorize income according to poverty status, there would be both a conceptual and statistical association between poverty status and Medicaid status. When two predictor variables are highly related, that variable that is conceptually better related to the outcome should be selected. It is also important to statistically assess associations between variables. It is reasonable to begin these assessments by computing correlations between continuous variables and associations between nominal variables. However, bivariate associations may not reveal that groups of variables may be related. Many statistical packages provide regression diagnostic statistics by which to assess associations between independent variables. It is outside of the scope of this chapter to discuss these techniques; we refer the readers to standard statistical references to learn more about these techniques (Hosmer and Lemeshow 2000; SAS Institute 2002).

Once a set of candidate predictor variables is determined, each candidate variable should be evaluated, it is recommended that the association between each predictor variable and the outcome be assessed. Typically, variables that do not have an association with the outcome at a level of significance of 0.25 or less are not considered in the regression model because it is unlikely that they would significantly contribute to the prediction of the outcome (Hosmer and Lemeshow 2000).

### 3.2.2 Logistic Regression

Logistic regression with multiple independent variables allows simultaneous assessment of the contribution of multiple predictors of no-show. The outcome in the logistic model is a dichotomous variable indicating whether the $i$th patient does not show up to its next medical appointment with probability $p_i$ or shows up to the medical appointment with probability of $1 - p_i$. The equation that describes the relationship between the predictor variables and the outcome follows:

$$\text{logit}(p_i(\mathbf{x}_i)) = log\left(\frac{p_i(\mathbf{x}_i)}{1 - p_i(\mathbf{x}_i)}\right) = \alpha + \beta_1 x_{i_1} + \cdots + \beta_n x_{i_d}$$

where $\alpha$ is the constant of the equation, the $\beta$'s are the coefficients of the predictor variables, and $\mathbf{x}_i$ is the vector of predictor variables for patient $i$, that is, $\mathbf{x}_i = (x_{i_1}, \ldots, x_{i_d})^{\mathrm{T}}$. Statistical packages include a variety of variable selection techniques (e.g., forward, backward, stepwise) to select from the set of candidate variables the most parsimonious set of predictors that most accurately predicts the outcome. We refer readers to standard statistical references to determine the type of variable selection technique that is most appropriate for their work.

There are several ways to evaluate the fit of the model. The coefficients in the model should be examined to make sure the directions of the regression coefficients make sense. The Wald chi-square statistic will inform whether a variable significantly contributes to the prediction equation (Hosmer and Lemeshow 2000). One should also consider the fit of the model as a whole. Most statistical packages present a chi-square goodness of fit test that compares whether the number of persons predicted to have the outcome is similar to the actual number of persons with the outcome for categories of the independent variable. A nonsignificant chi-square test indicates that the number predicted to have the outcome is similar to the number who actually experienced the outcome. The Hosmer–Lemeshow test is similar to chi-square goodness of fit test, except that the categories of the independent variables that are compared have approximately equal numbers of observations. The c-statistic is also used to assess the prediction performance (Hosmer and Lemeshow 2000).

## *3.3 Validating the No-Show Model*

The next important step is to validate the no-show predictive model. Validation can be done by internal and external methods (Harrell 2001a; Steyerberg 2009). This involves determining whether an estimated no-show probability accurately reflects an individual's no-show behavior. Since the number of appointments for any given person in the data set is likely to be small (assuming that the researcher has obtained 2–3 years worth of data and the typical patient schedules 2–3 appointments per year), it is better to assess differences between expected and observed no-show behaviors for groups of patients. For example, if we randomly select a "sufficiently large" group of patient appointments from the validation cohort and compute the no-show probability for each patient appointment, we can easily compute the expected number of no-shows for the group. This expectation can then be compared with the actual no-show results for that randomly selected group since, as noted earlier, missed appointments will be coded as no-show. Once this is repeated for many randomly selected groups, goodness of fit tests can be used to access differences between expected and observed group no-shows. Further refinement of the model may be required if the comparison indicates large differences in the expected and observed no-shows.

Another approach to model validation uses the receiver operating characteristic (ROC) curve, which plots the model sensitivity versus the model specificity. We note that sensitivity is the probability that the model correctly predicts success (in our case, no-show), and specificity is the probability that the model correctly predicts failure (in our case, attendance). The logistic regression model will predict the probabilities of success. A decision rule is formed by selecting a threshold, and thus the pair (specificity, sensitivity) can be calculated. Different thresholds lead to different pairs of (specificity, sensitivity). The ROC curve plots all pairs of (1-specificity, sensitivity) by different thresholds. It is expected that a good predictive model should be very close to both the $Y$-axis and $Y = 1$ line, that is, has both high sensitivity and high specificity, and thus, the area under the ROC curve (typically referred to as c-statistic) will be large as well. Usually, a model with an ROC area exceeding 0.7 is considered to be valid in practice.

# 4 Example Application

This section presents an example of the no-show model building process from our research. We begin by discussing the data set and then present the bivariate associations from which we select candidate variables. Then we present the results of the logistic regression for estimating no-show probabilities. Finally, we validate no-show probabilities using the approach discussed in the previous section.

## *4.1 Methods*

Data for no-show modeling was obtained from 20 outpatient clinics of a midwestern hospital. The data used for analysis included information from 130,819 patients obtained over a 3-year period. SAS Version 9.1 (SAS Institute Inc., Cary, NC) was used for all statistical computations. Demographic variables and scheduling information were used to determine a patient's probability of no-showing to their most recent scheduled appointment using logistic regression.

To begin model building, all bivariate associations with no-show status were examined for the development cohort using likelihood ratio chi-square tests or *t* tests. Variables were considered as candidates for inclusion in the multivariate logistic regression model if the significance level from their univariate test was less than 0.25 as recommended by Hosmer and Lemeshow (2000). Tables 10.6 and 10.7 present patient demographics and appointment characteristics of the data. The associations between the outcome "no-show for last appointment" and categorical and continuous predictor variables are provided in Tables 10.6 and 10.7. These tables reveal that all predictor variables have an association with the outcome variable at a significance level of $p \leq 0.25$.

A full multivariate logistic regression model predicting the probability a patient was a no-show to their appointment included all candidate variables along with prior total number of scheduled visits, prior cumulative no-show rate, and the interaction of prior total number of scheduled visits and cumulative no-show rate. Both backward and forward model selection with a threshold of $p = 0.05$ resulted in the same final model. The final logistic regression model for predicting no-show probability included the following variables: race, learning site, provider type, insurance type, whether the prior scheduled visit was bumped, the time and the date of the appointment, whether the appointment was for a new visit, whether the patient received a telephone reminder via Televox, whether the appointment was the same day, and whether the appointment was overbooked.

Continuous variables included in the model were age, days of lead time for the appointment, and days since the last provider visit. Restricted cubic spline functions were used to allow a nonlinear relationship between the continuous variables and no-show probability. The interaction of number of previous scheduled visits and prior no-show probability was also included. The odds ratios for the logistic regression

Table 10.6 Associations for categorical variables

| | | Last appointment no-show | | |
|---|---|---|---|---|
| | | Yes | No | |
| $N = 130{,}819$ | | N (%) | N (%) | $p$ value |
| Gender | Male | 12,705 (24.4%) | 39,296 (75.6%) | 0.1663 |
| | Female | 19,521 (24.8%) | 59,291 (75.2%) | |
| Race | White | 9,118 (18.2%) | 40,887 (81.8%) | <0.0001 |
| | Black | 15,554 (31.6%) | 33,753 (68.4%) | |
| | Other/unknown | 7,557 (24.0%) | 23,950 (76.0%) | |
| Learning site | Yes | 25,811 (30.5%) | 58,798 (69.5%) | <0.0001 |
| | No | 6,418 (13.9%) | 39,792 (86.1%) | |
| Provider type | MD staff | 27,925 (23.7%) | 89,880 (76.3%) | <0.0001 |
| | MD resident | 1,942 (37.4%) | 3,251 (62.6%) | |
| | NP/PA | 2,362 (30.2%) | 5,459 (69.8%) | |
| Insurance type (last known) | Medicaid | 8,782 (25.9%) | 25,145 (74.1%) | <0.0001 |
| | Medicare | 1,468 (13.7%) | 9,269 (86.3%) | |
| | Self-pay | 2,673 (34.5%) | 5,065 (65.5%) | |
| | County tax-funded program | 5,469 (22.9%) | 18,405 (77.1%) | |
| | Other | 3,891 (12.1%) | 28,284 (87.9%) | |
| | Unknown | 9,946 (44.5%) | 12,422 (55.5%) | |
| Prior scheduled visit bumped | Yes | 2,413 (34.0%) | 4,674 (66.0%) | <0.0001 |
| | No | 29,816 (24.1%) | 93,916 (75.9%) | |
| AM appointment | Yes | 16,096 (24.5%) | 49,641 (75.5%) | 0.2031 |
| | No | 16,133 (24.8%) | 48,949 (75.2%) | |
| Appointment day | Monday | 7,045 (25.8%) | 20,274 (74.2%) | <0.0001 |
| | Tuesday | 7,266 (24.8%) | 21,997 (75.2%) | |
| | Wednesday | 5,053 (21.8%) | 18,154 (78.2%) | |
| | Thursday | 7,345 (25.9%) | 20,994 (74.1%) | |
| | Friday/Saturday | 5,520 (24.3%) | 17,171 (75.7%) | |
| New visit | Yes | 3,847 (47.7%) | 4,224 (52.3%) | <0.0001 |
| | No | 28,382 (23.1%) | 94,366 (76.9%) | |
| Televox appointment reminder | Yes | 22,150 (21.4%) | 81,147 (78.6%) | <0.0001 |
| | No | 10,079 (36.6%) | 17,443 (63.4%) | |
| Same-day appointment | Yes | 1,030 (4.1%) | 24,296 (95.9%) | <0.0001 |
| | No | 31,199 (29.6%) | 74,294 (70.4%) | |
| Overbooked visit | Yes | 251 (2.7%) | 9,178 (97.3%) | <0.0001 |
| | No | 31,978 (26.3%) | 89,412 (73.7%) | |

**Table 10.7** Associations for continuous variables

| | Last appointment no-show | | |
|---|---|---|---|
| | Yes | No | |
| | Mean (SD) | Mean (SD) | |
| $N = 130{,}819$ | $N = 32{,}229$ | $N = 98{,}590$ | $p$ value |
| Age (at last visit) | 26.5 (20.3) | 32.7 (23.6) | <0.0001 |
| Lead time (days) | 46.4 (46.1) | 24.1 (37.0) | <0.0001 |
| Days since prior scheduled provider visit | 103.8 (124.2) | 119.8 (135.8) | <0.0001 |
| Prior scheduled visits | 6.4 (9.4) | 7.6 (7.1) | <0.0001 |
| Prior no-show rate | 34.6 (33.5) | 16.9 (25.1) | <0.0001 |

are provided in Table 10.8. In the regression equation, which is used to estimate no-show probability, the coefficient for each variable is the natural logarithm of the associated odds ratio, that is, $\beta_{\text{Televox}} = \ln(0.51)$. For continuous variables, the restricted cubic spline components are calculated based on the restricted cubic spline function (see Harrell 2001b for more information). Based on restricted cubic splines for continuous variables, younger patients are more likely to no-show than older patients, the patients with lower lead times between the time appointment is made and the actual appointment time are less likely to no-show compared to the ones with higher lead times, and as the time since the last appointment increases, the patients become less likely to no-show.

## *4.2 Model Validation*

Model validation was assessed by calculating the area under the ROC curve which may be characterized by the c-statistic for dichotomous outcomes. The c-statistic that measures the model's ability to discriminate between patients who no-showed versus those who attended was high, at 0.8236 (Fig. 10.5). This shows that the logistic regression model has excellent discrimination between no-show patients and patients who attended.

In addition, we considered the performance of the model in the context of provider workload. We devised 100 groups of 30 patients from a randomly selected set of approximately 12,000 patients. Note that 30 patients is the approximate size of one provider's daily schedule at our collaborating clinics. For each patient in a group, we calculated the no-show probability of the patient using the no-show prediction model. We also noted the patients actual no show status. Then, we determined the expected number of no-shows for the group by summing the no-show probabilities of all patients in the group. We computed the actual number of no-shows by summing the number of appointments for which patients did not attend. We then calculated the difference between these two values. The overall average is $-0.14$ with a standard deviation of 1.77. Approximately 40% of all groups fall in the

**Table 10.8** Odds ratios and confidence limits for candidate variables

| Effect | | Odds ratio | 95% C.I. | *p* value |
|---|---|---|---|---|
| Age | | 1.05 | [1.04, 1.05] | <0.0001 |
| Age restricted cubic spline component 1 | | 0.82 | [0.81, 0.83] | <0.0001 |
| Age restricted cubic spline component 2 | | 1.42 | [1.37,1.47] | <0.0001 |
| Race | White vs. black | 0.71 | [0.69, 0.74] | <0.0001 |
| | Other vs. black | 0.67 | [0.65, 0.70] | |
| Learning site | Yes vs. no | 1.19 | [1.13, 1.24] | <0.0001 |
| Provider type | MD resident vs. MD staff | 1.32 | [1.24, 1.41] | <0.0001 |
| | NP/PA vs. MD staff | 1.13 | [1.06, 1.20] | 0.0001 |
| Insurance | Medicaid vs. unknown | 0.61 | [0.58, 0.64] | <0.0001 |
| | Medicare vs. unknown | 0.65 | [0.60, 0.71 ] | <0.0001 |
| | Self-pay vs. unknown | 1.01 | [0.95, 1.08] | 0.7114 |
| | County tax-funded program vs. unknown | 0.63 | [0.60, 0.66] | <0.0001 |
| | Other vs. unknown | 0.4 | [0.38, 0.43] | <0.0001 |
| Prior scheduled visit bumped | Yes vs. no | 1.37 | [1.30, 1.46] | <0.0001 |
| Lead time (days) | | 1.07 | [1.07, 1.08] | <0.0001 |
| Lead time (days) restricted cubic spline component 1 | | 0.3 | [0.24, 0.38] | <0.0001 |
| Lead time (days) restricted cubic spline component 2 | | 4.23 | [3.17,5.64] | <0.0001 |
| Days since prior scheduled provider visit | | 0.99 | [0.99,1.00] | <0.0001 |
| Days since prior scheduled provider visit restricted cubic spline component 1 | | 1.04 | [1.01,1.07] | 0.0095 |
| Days since prior scheduled provider visit restricted cubic spline component 2 | | 0.95 | [0.91,1.00] | 0.0339 |
| AM appointment | Yes vs. no | 0.95 | [0.93,0.98] | 0.0015 |
| Appointment day | Monday vs. Friday/Saturday | 1.06 | [1.01,1.11] | 0.0243 |
| | Tuesday vs. Friday/Saturday | 0.99 | [0.95,1.04] | 0.7421 |
| | Wednesday vs. Friday/Saturday | 0.93 | [0.89, 0.98] | 0.0072 |
| | Thursday vs. Friday/Saturday | 1 | [0.95,1.05] | 0.8998 |
| New visit type | New visit vs. Return | 1.11 | [1.05,1.17] | 0.0004 |
| Televox | Yes vs. no | 0.51 | [0.49,0.52] | <0.0001 |
| Same-day appointment | Yes vs. no | 0.33 | [0.30,0.35] | <0.0001 |
| Overbooked appointment | Yes vs. no | 0.07 | [0.06,0.08] | <0.0001 |

The regression model includes an interaction between the prior cumulative no-show rate and cumulative number of visits. Main effects of prior cumulative no-show rate and cumulative number of visits have *p* values of <0.0001, and their interaction has *p* value of <0.0001. We do not present the regression estimate for this interaction because it is not easily interpreted

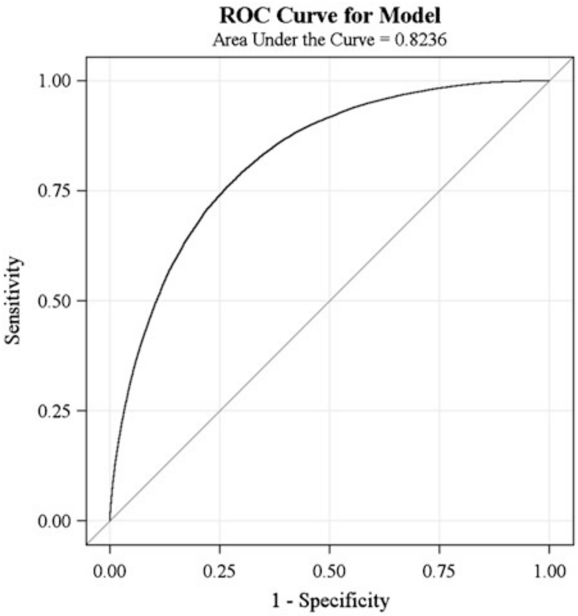


**Fig. 10.5** Receiver operating characteristic (ROC) curve for the logistic regression model

range [−1,1], indicating that observed no-show differs from the expected no-show by no more than 1. Similarly, 71% fall in the range [−2,2], and 91% fall in the range [−3,3]. Further, the chi-square test indicates good agreement between observed and expected, with the test statistic of $\sum \frac{(\text{expected}-\text{observed})^2}{\text{expected}} = 61.23$, which is much less than the critical value of 123.24 for a test with 99 degrees of freedom. Thus, we conclude that the no-show model provides important information for determining provider workload.

## 5 Discussion

It seems clear that the true reasons for no-show are context dependent and difficult to accurately identify. Consider, for example, the following set of reasons that a clinic might encounter in surveying no-show patients:

- *I forgot.*
- *I did not have a ride.*
- *I had to take care of my kids.*
- *I was broke.*
- *I felt too bad.*
- *They don't respect me.*
- *I don't like my doctor.*
- *I was afraid.*
- *I just don't care.*

The drivers of no-show cover the spectrum of human frailty. Patients who have missed appointments in the past are likely to continue missing them in the future. This lends credence to the point that the true reasons for no-show are often hidden, and thus it is unclear that straightforward interventions targeted toward objective reasons will reduce no-show below some benchmark level. Clearly, interventions targeting emotional, entrenched reasons such as fearfulness, perceived disrespect, or lack of motivation need to be much more sophisticated (and perhaps more costly) than those targeting forgetfulness. The clinic needs to understand which reasons are most prevalent for its patient panel, what interventions can be selected for addressing them, and what the cost/benefit implications of implementation are.

Statistical no-show modeling typically involves some type of statistical regression on a few years worth of historical data to identify important "predictors." Published studies tend to end with the reporting of these predictors. But, little has been done with these types of models to operationally mitigate the no-show problem. What shortcomings would emerge if these statistical models were used in practice over time? This is unclear. One can imagine that dynamic changes in a clinic's patient panel might soon render a no-show model obsolete. How should such a model be kept up-to-date? No such studies or recommendations have been published. As for mitigating the effect of no-show, the most commonly suggested method is overbooking, which has gotten much attention in the appointment scheduling literature of late. To our knowledge, there are few implementations of the "optimal" overbooking models that have been proposed, only computer simulations.

Many authors note that their no-show models might not be generalizable because studies were always conducted in specific clinical populations. Therefore, it appears that every clinic needs its own unique model. What is the most effective way to develop such a model? Does every clinic need to follow the methods spelled out in this chapter or can methods that require no initial data by taking advantage published work be developed? Intuition tells us that it should be possible to let new models develop and "evolve" with clinic operation, but the precise means of doing this are not currently known.

Finally, there is only limited literature addressing the health outcomes of no-show behavior, and almost none appears to address the costs to society. We believe the fragmented nature of health information record keeping has made this line of research very difficult in the past. Based on very rough and unreported estimates from our own work, no-show among diabetics could be costing as much as $10 billion per year in increased medical costs in the USA. But no one really knows, and the full implications of the no-show problem remain hidden.

## 6 Future Research Opportunities

Based on our reading and work, we believe the following statements capture the essence of the patient no-show problem:

1. No-show behavior is harmful and costly.
2. No-show behavior has its root causes in the human condition.

3. No-show behavior can be reduced through interventions, but it cannot be eliminated.
4. No-show behavior can be statistically "predicted," but statistical models are not generalizable.

These statements drive what we believe are the greatest research needs. The first need is to understand the full impact of no-show. More research needs to address the health outcomes and systemic impact of no-show for a larger set of conditions. We hope that this research will become more broadly feasible in the future with emerging electronic medical record systems and health information exchange technology.

The second need is to have a better understanding of the underlying drivers and how to devise optimal intervention programs to deal with them. More research is needed to devise better interventions, to map interventions onto underlying no-show drivers, to identify costs and benefits, to develop acceptable methods of economic analysis, and to formulate models for identifying a clinic's optimal mix of interventions.

The third need is to develop more effective methods for predicting no-shows and mitigating their impact. Since each clinic needs its unique no-show model, "meta-methods" that use the published literature to devise a first-pass model that can continuously evolve during clinic operations should be developed to reduce the time spent for model development.

## 7 Conclusion

This chapter provides a review of existing no-show studies and develops an example of a statistical no-show prediction model. The studies in the literature are discussed in four classes: self-reported reasons for no-shows, interventions to reduce no-shows, statistical models to predict no-shows, and impact of no-shows on health outcomes. The most commonly self-reported reasons include patient-related factors, scheduling system problems, and environmental and financial factors. Interventions to reduce no-shows are patient-related interventions such as appointment reminders, patient education, follow-up after a no-showed appointment, and system-related interventions to reduce waiting times such as open-access scheduling, and lean process improvement methods. The predictors of no-shows identified using statistical methods include age, sex, family status, race/ethnicity, insurance status, appointment characteristics, patient diagnoses, and type of provider. The effect of no-shows on health outcomes is analyzed for different patient populations such as diabetes, dialysis, HIV, primary care, and psychiatric patients. The no-show percentages reported in these studies are presented according to the patient population (mental health, primary care, HIV, chronic care, etc.).

The second part of the chapter explains how statistical no-show prediction models can be developed. The data requirements, determination of significant factors, development of logistic regression models, and validation of prediction models are explained. A regression model is developed using the scheduling and billing data from 20 outpatient clinics of a midwestern hospital. The variables identified as predictors of no-show include age, race, learning site, provider type, insurance type, whether the prior scheduled visit was bumped, the time and the date of the appointment, whether the appointment was for a new visit, whether the patient received a telephone reminder via Televox, whether the appointment was on the same day and whether the appointment was overbooked, lead time for the appointment, and days since the last provider visit. The regression model includes variables that are not included in earlier studies such as appointment reminders, and the type of the clinic (learning site for residents or not).

## References

Alafaireet P, Houghton H, Petroski G, Gong Y, Savage GT (2010) Toward determining the structure of psychiatric visit nonadherence. J Ambul Care Manag 33(2):108–116

American Diabetes Association (2010) Standards of medical care in diabetes—2010. Diabetes Care 33(Supplement 1):S11–S61

Andersen M, Hockman E, Smereck G, Tinsley J, Milfort D, Wilcox R, Smith T, Connelly C, Adams L, Thomas R (2007) Retaining women in HIV medical care. J Assoc Nurses AIDS Care 18(3):33–41

Andersen R, Newman J (2005) Societal and individual determinants of medical care utilization in the United States. Milbank Quart 83(4), Online–only

Bakken S, Holzemer WL, Brown MA, Powell-Cope GM, Turner JG, Inouye J, Nokes KM, Corless IB (2000) Relationships between perception of engagement with health care provider and demographic characteristics, health status, and adherence to therapeutic regimen in persons with HIV/AIDS. AIDS Patient Care STDs 14(4):189–197

Bean AG, Talaga J (1992) Appointment breaking: causes and solutions. J Health Care Market 12(4):14–25

Bech M (2005) The economics of non-attendance and the expected effect of charging a fine on non-attendees. Health Policy 74(2):181–191

Belardi F, Weir S, Craig F (2004) A controlled trial of an advanced access appointment system in a residency family medicine center. Fam Med 36(5):341–345

Bennett KJ, Baxley EG (2009) The effect of a carve-out advanced access scheduling system on no-show rates. Fam Med 41(1):51–56

Bigby JA, Pappius E, Cook EF, Goldman L (1984) Medical consequences of missed appointments. Arch Intern Med 144(6):1163–1166

Blank MB, Chang MY, Fox JC, Lawson CA, Modlinski J (1996) Case manager follow-up to failed appointments and subsequent service utilization. Commun Ment Health J 32(1):23–31

Blankson M, Goldenberg R, Keith B (1994) Noncompliance of high-risk pregnant women in keeping appointments at an obstetric complications clinic. South Med J 87(6):634

Bowser DM, Utz S, Glick D, Harmon R (2010) A systematic review of the relationship of diabetes mellitus, depression, and missed appointments in a low-income uninsured population. Arch Psychiatr Nurs 24(5):317–329

Bundy D, Randolph G, Murray M, Anderson J, Margolis P (2005) Open access in primary care: results of a North Carolina Pilot Project. Pediatrics 116(1):82–87

Cameron S, Sadler L, Lawson B (2010) Adoption of open-access scheduling in an academic family practice. Can Fam Physician 56(9):906–911

Campbell J, Chez R, Queen T, Barcelo A, Patron E (2000) The no-show rate in a high-risk obstetric clinic. J Wom Health Gend Bas Med 9(8):891–895

Can S, Macfarlane T, O'Brien KD (2003) The use of postal reminders to reduce non-attendance at an orthodontic clinic: a randomised controlled trial. Br Dent J 195(4):199–201

Cashman SB, Savageau JA, Lemay CA, Ferguson W (2004) Patient health status and appointment keeping in an urban community health center. J Health Care Poor Underserved 15(3):474–488

Catz SL, McClure JB, Jones GN, Brantley PJ (1999) Predictors of outpatient medical appointment attendance among persons with HIV. AIDS Care 11(3):361–373

Cavaleri MA, Kalogerogiannis K, McKay MM, Vitale L, Levi E, Jones S, Wallach F, Flynn E (2010) Barriers to HIV care: an exploration of the complexities that influence engagement in and utilization of treatment. Soc Work Health Care 49(10):934–945

Cayirli T, Veral E, Rosen H (2006) Designing appointment scheduling systems for ambulatory care services. Health Care Manag Sci 9(1):47–58

Cayirli T, Veral E, Rosen H (2008) Assessment of patient classification in appointment system design. Prod Oper Manag 17(3):338–353

Centorrino F, Hernán MA, Drago-Ferrante G, Rendall M, Apicella A, Längar G, Baldessarini RJ (2001) Factors associated with noncompliance with psychiatric outpatient visits. Psychiatr Serv 52(3):378–380

Charupanit W (2009) Factors related to missed appointment at psychiatric clinic in Songklanagarind Hospital. J Med Assoc Thai 92(10):1367–1369

Ciechanowski P, Russo J, Katon W, Simon G, Ludman E, Von Korff M, Young B, Lin E (2006) Where is the patient? The association of psychosocial factors and missed primary care appointments in patients with diabetes. Gen Hosp Psychiatr 28(1):9–17

Cohen AD, Dreiher J, Vardy DA, Weitzman D (2008) Nonattendance in a dermatology clinic – a large sample analysis. J Eur Acad Dermatol Veneorol 22(10):1178–1183

Collins J, Santamaria N, Clayton L (2003) Why outpatients fail to attend their scheduled appointments: a prospective comparison of differences between attenders and non-attenders. Aust Health Rev 26(1):52–63

Compton MT, Rudisch BE, Craw J, Thompson T, Owens DA (2006) Predictors of missed first appointments at community mental health centers after psychiatric hospitalization. Psychiatr Serv 57(4):531–537

Corfield L, Schizas A, Noorani A, Williams A (2008) Non-attendance at the colorectal clinic: a prospective audit. Ann R Coll Surg Engl 90(5):377–380

Daggy J, Lawley M, Willis D, Thayer D, Suelzer C, DeLaurentis P-C, Turkcan A, Chakraborty S, Sands L (2010) Using no-show modeling to improve clinic performance. Health Informat J 16(4):246–259

Davidson MB, Karlan VJ, Hair TL (2000) Effect of a pharmacist-managed diabetes care program in a free medical clinic. Am J Med Qual 15(4):137–142

Defife JA, Conklin CZ, Smith JM, Poole J (2010) Psychotherapy appointment no-shows: Rates and reasons. Psychotherapy 47(3):413–417

Denhaerynck K, Manhaeve D, Dobbels F, Garzoni D, Nolte C, De Geest S (2007) Prevalence and consequences of nonadherence to hemodialysis regimens. Am J Crit Care 16(3):222–235

Deyo RA, Inui TS (1980) Dropouts and broken appointments: a literature review and agenda for future research. Med Care 18(11):1146–1157

Dove HG, Schneider KC (1981) The usefulness of patients' individual characteristics in predicting no-shows in outpatient clinics. Med Care 19(7):734–740

Fischman D (2010) Applying Lean Six Sigma methodologies to improve efficiency, timeliness of care, and quality of care in an internal medicine residency clinic. Qual Manag Health Care 19(3):201–210

Gany F, Ramirez J, Chen S, Leng JCF (2011) Targeting social and economic correlates of cancer treatment appointment keeping among immigrant Chinese patients. J Urban Health 88(1): 98–103

Garuda SR, Javalgi RG, Talluri V (1998) Tackling no-show behavior: a market-driven approach. Health Market Quart 15(4)

George A, Rubin G (2003) Non-attendance in general practice: a systematic review and its implications for access to primary health care. Fam Pract 20(2):178–184

Griffin SJ (1998) Lost to follow-up: the problem of defaulters from diabetes clinics. Diabet Med 15(Suppl 3):S14–S24

Guse C, Richardson L, Carle M, Schmidt K (2003) The effect of exit-interview patient education on no-show rates at a family practice residency clinic. J Am Board Fam Pract 16(5):399–404

Hamilton W, Round A, Sharp D (2002) Patient, hospital, and general practitioner characteristics associated with non-attendance: a cohort study. Br J Gen Pract 52(477):317–319

Hardy KJ, O'Brien SV, Furlong NJ (2001) Information given to patients before appointments and its effect on non-attendance rate. BMJ 323:1298–1300

Harrell F (2001a) Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer, Berlin

Harrell FE (2001b) Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. Springer, Berlin

Herrick J, Gilhooly ML, Geddes DA (1994) Non-attendance at periodontal clinics: forgetting and administrative failure. J Dent 22(5):307–309

Hershey CO, Cohen DI, Goldberg HI, McLaren CE, Dawson NV, Siciliano C, Porter DK, Breslau D (1987) Effect of an academic group practice on patient show rates: a randomized controlled trial. Med Care 25(1):72–77

Hochstadt NJ, Trybula J Jr (1980) Reducing missed initial appointments in a community mental health center. J Commun Psychol 8(3):261–265

Hosmer DW, Lemeshow S (2000) Applied logistic regression. Wiley, New York

Hurtado AV, Greenlick MR, Colombo TJ (1973) Determinants of medical care utilization: failure to keep appointments. Med Care 11(3):189–198

Jayaram M, Rattehalli RD, Kader I (2008) Prompt letters to reduce non-attendance: applying evidence based practice. BMC Psychiatr 8:90

Johnson BJ, Mold JW, Pontious JM (2007) Reduction and management of no-shows by family medicine residency practice exemplars. Ann Fam Med 5(6):534–539

Jonas S (1973) Influence of the weather on appointment-breaking in a general medical clinic. Med Care 11(1):72–74

Karter A, Ackerson L, Darbinian J, D'Agostino R, Ferrara A, Liu J, Selby J (2001) Self-monitoring of blood glucose levels and glycemic control: the Northern California kaiser permanente diabetes registry. Am J Med 111(1):1–9

Karter AJ, Parker MM, Moffet HH, Ahmed AT, Ferrara A, Liu JY, Selby JV (2004) Missed appointments and poor glycemic control: an opportunity to identify high-risk diabetic patients. Med Care 42(2):110–115

Killaspy H, Banerjee S, King M, Lloyd M (1999) Non-attendance at psychiatric outpatient clinics: Communication and implications for primary care. Br J Gen Pract 49(448):880–883

Kluger MP, Karras A (1983) Strategies for reducing missed initial appointments in a community mental health center. Commun Ment Health J 19(2):137–143

Kopach R, DeLaurentis P-C, Lawley M, Muthuraman K, Ozsen L, Rardin R, Wan H, Intrevado P, Qu X, Willis D (2007) Effects of clinical characteristics on successful open access scheduling. Health Care Manag Sci 10(2):111–124

Kros J, Dellana S, West D (2009) Overbooking increases patient access at East Carolina University's student health services clinic. Interfaces 39(3):271–287

Kruse GR, Rohland BM, Wu X (2002) Factors associated with missed first appointments at a psychiatric clinic. Psychiatr Serv 53(9):1173–1176

Kunutsor S, Walley J, Katabira E, Muchuro S, Balidawa H, Namagala E, Ikoona E (2010a) Clinic attendance for medication refills and medication adherence amongst an antiretroviral treatment cohort in Uganda: a prospective study. AIDS Res Treat 2010:872396

Kunutsor S, Walley J, Katabira E, Muchuro S, Balidawa H, Namagala E, Ikoona E (2010b) Using mobile phones to improve clinic attendance amongst an antiretroviral treatment cohort in rural Uganda: a cross-sectional and prospective study. AIDS Behav 14(6):1347–1352

Lacy NL, Paulman A, Reuter MD, Lovejoy B (2004) Why we don't come: Patient perceptions on no-shows. Ann Fam Med 2(6):541–545

LaGanga LR (2011) Lean service operations: reflections and new directions for capacity expansion in outpatient clinics. J Oper Manag 29(5):422–433

Lehmann TNO, Aebi A, Lehmann D, Balandraux Olivet M, Stalder H (2007) Missed appointments at a Swiss university outpatient clinic. Publ Health 121(10):790–799

Leigh H, Cruz H, Mallios R (2009) Telepsychiatry appointments in a continuing care setting: kept, cancelled and no-shows. J Telemed Telecare 15(6):286–289

Leong KC, Chen WS, Leong KW, Mastura I, Mimi O, Sheikh MA, Zailinawati AH, Ng CJ, Phua KL, Teng CL (2006) The use of text messaging to improve attendance in primary care: a randomized controlled trial. Fam Pract 23(6):699–705

Liew S-M, Tong SF, Lee VKM, Ng CJ, Leong KC, Teng CL (2009) Text messaging reminders to reduce non-attendance in chronic disease follow-up: a clinical trial. Br J Gen Pract 59(569):916–920

Lin KC (2009) Behavior-associated self-report items in patient charts as predictors of dental appointment avoidance. J Dent Educ 73(2):218–224

Lindauer SJ, Powell JA, Leypoldt BC, Tufekci E, Shroff B (2009) Influence of patient financial account status on orthodontic appointment attendance. Angle Orthod 79(4):755–758

Liu N, Ziya S, Kulkarni VG (2010) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. Manuf Serv Oper Manag 12(2):347–364

Livianos-Aldana L, Vila-Gómez M, Rojo-Moreno L, Luengo-López MA (1999) Patients who miss initial appointments in community psychiatry? A Spanish community analysis. Int J Soc Psychiatr 45(3):198–206

Lowe RH (1982) Responding to "no-shows": some effects of follow-up method on community mental health center attendance patterns. J Consult Clin Psychol 50(4):602–603

Macharia WM, Leon G, Rowe BH, Stephenson BJ, Haynes RB (1992) An overview of interventions to improve compliance with appointment keeping for medical services. JAMA 267(13):1813–1817

Mandall NA, Matthew S, Fox D, Wright J, Conboy FM, O'Brien KD (2008) Prediction of compliance and completion of orthodontic treatment: are quality of life measures important? Eur J Orthod 30(1):40–45

Martin C, Perfect T, Mantle G (2005) Non-attendance in primary care: the views of patients and practices on its causes, impact and solutions. Fam Pract 22(6):638–643

Matas M, Staley D, Griffin W (1992) A profile of the noncompliant patient: a thirty-month review of outpatient psychiatry referrals. Gen Hosp Psychiatr 14(2):124–130

Mehrotra A, Keehl-Markowitz L, Ayanian JZ (2008) Implementing open-access scheduling of visits in primary care practices: a cautionary tale. Ann Intern Med 148(12):915+

Milne R, Horne M, Torsney B (2006) SMS reminders in the UK national health service: an evaluation of its impact on "no-shows" at hospital out-patient clinics. Health Care Manag Rev 31(2):130–136

Mirotznik J, Ginzler E, Zagon G, Baptiste A (1998) Using the health belief model to explain clinic appointment-keeping for the management of a chronic disease condition. J Commun Health 23(3):195–210

Mitchell AJ, Selmes T (2007) A comparative survey of missed initial and follow-up appointments to psychiatric specialties in the United Kingdom. Psychiatr Serv 58(6):868–871

Moore C, Wilson-Witherspoon P, Probst J (2001) Time and money: effects of no-shows at a family practice residency clinic. Fam Med 33(7):522–527

Mugavero MJ, Lin H-Y, Allison JJ, Willig JH, Chang P-W, Marler M, Raper JL, Schumacher JE, Pisu M, Saag MS (2007) Failure to establish HIV care: characterizing the "no show" phenomenon. Clin Infect Dis 45(1):127–130

Mugavero MJ, Lin H-Y, Willig JH, Westfall AO, Ulett KB, Routman JS, Abroms S, Raper JL, Saag MS, Allison JJ (2009a) Missed visits and mortality among patients establishing initial outpatient HIV treatment. Clin Infect Dis 48(2):248–256

Mugavero MJ, Lin H-Y, Allison JJ, Giordano TP, Willig JH, Raper JL, Wray NP, Cole SR, Schumacher JE, Davies S, Saag MS (2009b) Racial disparities in HIV virologic failure: do missed visits matter? J Acquir Immune Defic Syndr 50(1):100–108

Murdock A, Rodgers C, Lindsay H, Tham TCK (2002) Why do patients not keep their appointments? Prospective study in a gastroenterology outpatient clinic. J R Soc Med 95(6):284–286

Murphy K, Edelstein H, Smith L, Clanon K, Schweitzer B, Reynolds L, Wheeler P (2011) Treatment of HIV in outpatients with schizophrenia, schizoaffective disorder and bipolar disease at two county clinics. Community Ment Health J 47(6):668–671

Murray M, Tantau C (2000) Same-day appointments: exploding the access paradigm. Fam Pract Manag 7(8):45–50

Neal RD, Hussain-Gambles M, Allgar VL, Lawlor DA, Dempsey O (2005) Reasons for and consequences of missed appointments in general practice in the UK: questionnaire survey and prospective review of medical records. BMC Fam Pract 6:47

Obialo CI, Bashir K, Goring S, Robinson B, Quarshie A, Al-Mahmoud A, Alexander-Squires J (2008) Dialysis "no-shows" on Saturdays: implications of the weekly hemodialysis schedules on nonadherence and outcomes. J Natl Med Assoc 100(4):412–419

Pang AH, Lum FC, Ungvari GS, Wong CK, Leung YS (1996) A prospective outcome study of patients missing regular psychiatric outpatient appointments. Soc Psychiatr Psychiatr Epidemiol 31(5):299–302

Parikh A, Gupta K, Wilson AC, Fields K, Cosgrove NM, Kostis JB (2010) The effectiveness of outpatient appointment reminder systems in reducing no-show rates. Am J Med 123(6): 542–548

Park WB, Kim JY, Kim S-H, Kim HB, Kim NJ, Oh M-D, Choe KW (2008) Self-reported reasons among HIV-infected patients for missing clinic appointments. Int J STD AIDS 19(2):125–126

Peeters FP, Bayer H (1999) 'No-show' for initial screening at a community mental health centre: Rate, reasons and further help-seeking. Soc Psychiatr Psychiatr Epidemiol 34(6):323–327

Potamitis T, Chell PB, Jones HS, Murray PI (1994) Non-attendance at ophthalmology outpatient clinics. J R Soc Med 87(10):591–593

Rhee MK, Slocum W, Ziemer DC, Culler SD, Cook CB, El-Kebbi IM, Gallina DL, Barnes C, Phillips LS (2005) Patient adherence improves glycemic control. Diabetes Educ 31(2): 240–250

Rocco MV, Burkart JM (1993) Prevalence of missed treatments and early sign-offs in hemodialysis patients. J Am Soc Nephrol 4(5):1178–1183

Rockart JF, Hofmann PB (1969) Physician and patient behavior under different scheduling systems in a hospital outpatient department. Med Care 7(6):463–470

Rose KD, Ross JS, Horwitz LI (2011) Advanced access scheduling outcomes: a systematic review. Arch Intern Med 171(13):1150–1159

Rowett M, Reda S, Makhoul S (2010) Prompts to encourage appointment attendance for people with serious mental illness. Schizophr Bull 36(5):910–911

Samuels TA, Bolen S, Yeh HC, Abuid M, Marinopoulos SS, Weiner JP, McGuire M, Brancati FL (2008) Missed opportunities in diabetes management: a longitudinal assessment of factors associated with sub-optimal quality. J Gen Intern Med 23(11):1770–1777

Saran R, Bragg-Gresham JL, Rayner HC, Goodkin DA, Keen ML, Van Dijk PC, Kurokawa K, Piera L, Saito A, Fukuhara S, Young EW, Held PJ, Port FK (2003) Nonadherence in hemodialysis: associations with mortality, hospitalization, and practice patterns in the DOPPS. Kidney Int 64(1):254–262

Sarnquist CC, Soni S, Hwang H, Topol BB, Mutima S, Maldonado YA (2011) Rural HIV-infected women's access to medical care: ongoing needs in California. AIDS Care 23(7):792–796

SAS Institute (2002) SAS/STAT 9 User's guide, vol 1, 2 and 3. SAS institute Inc. Cary, NC

Satiani B, Miller S, Patel D (2009) No-show rates in the vascular laboratory: analysis and possible solutions. J Vasc Intervent Radiol 20(1):87–91

Schectman JM, Schorling JB, Voss JD (2008) Appointment adherence and disparities in outcomes among patients with diabetes. J General Intern Med 23(10):1685–1687

Smith DM, Norton JA, Weinberger M, McDonald CJ, Katz BP (1986) Increasing prescribed office visits: a controlled trial in patients with diabetes mellitus. Med Care 24(3):189–199

Snow BW, Cartwright PC, Everitt S, Ekins M, Maudsley W, Aloi S (2009) A method to improve patient access in urological practice. J Urol 182(2):663–667

Sparr LF, Moffitt MC, Ward MF (1993) Missed psychiatric appointments: who returns and who stays away. Am J Psychiatr 150(5):801–805

Steyerberg E (2009) Clinical prediction models: a practical approach to development, validation, and updating. Springer, Berlin

Swenson TR, Pekarik G (1988) Interventions for reducing missed initial appointments at a community mental health center. Commun Ment Health J 24(3):205–218

Tanke ED, Leirer VO (1994) Automated telephone reminders in tuberculosis care. Med Care 32(4):380–389

Tseng F-Y (2010) Non-attendance in endocrinology and metabolism patients. J Formos Med Assoc 109(12):895–900

Tuller DM, Bangsberg DR, Senkungu J, Ware NC, Emenyonu N, Weiser SD (2010) Transportation costs impede sustained adherence and access to HAART in a clinic population in southwestern Uganda: a qualitative study. AIDS Behav 14(4):778–784

U.S. Department of Health and Human Services, National Institute of Health (2007) HIPAA privacy rule. http://privacyruleandresearch.nih.gov/pr_08.asp. Accessed 1 June 2012

Weinerman R, Glossop V, Wong R, Robinson L, White K, Kamil R (2003) Time of day influences nonattendance at urgent short-term mental health unit in Victoria, British Columbia. Can J Psychiatr 48(5):342–344

Williams ME, Latta J, Conversano P (2008) Eliminating the wait for mental health services. J Behav Health Serv Res 35(1):107–114

Xakellis G, Bennett A (2001) Improving clinic efficiency of a family medicine teaching clinic. Fam Med 33(7):533–538

Yehia BR, Gebo KA, Hicks PB, Korthuis PT, Moore RD, Ridore M, Mathews WC, HIV Res Network (2008) Structures of care in the clinics of the HIV research network. AIDS Patient Care STDs 22(12):1007–1013

# Chapter 11
# Simulation and Real-Time Optimised Relocation for Improving Ambulance Operations

**Andrew James Mason**

## 1 Introduction

For many years, the Ambulance Logistics Group (Mason 2011) at the Department of Engineering Science (University of Auckland, New Zealand) has been collaborating with a university-spin-off company, The Optima Corporation (2011a), on the development of software for ambulance operators. Systems developed by Optima are now used to improve emergency response times for over 15 million people in New Zealand, Australia, Denmark, the United Kingdom, Canada and the USA. This chapter discusses the operations research models and methods that have been developed as a result of this successful university and industry collaboration

Work in the ambulance logistics area began when BartSim (Henderson and Mason 1999, 2004), an ambulance simulation system, was created to address staff scheduling questions being raised by St John Ambulance in Auckland, New Zealand. As part of this staff scheduling project, it became clear that there was a need to quantify the staffing required by St John. The number of staff required by an ambulance provider such as St John depends in a complex way on the number and location of calls and the contractual response targets the organisation must meet. To determine the staffing level required for St John, an initial queuing theory model was developed. When this proved to be inadequate, a discrete event simulation system was developed instead that could better model the complexity of the operations. This system allowed response-time performance to be simulated and thus estimated under different scenarios.

In 2001, an opportunity arose to develop and implement an improved ambulance simulation system for the Metropolitan Ambulance Service in Melbourne, Australia. This software, originally known as Siren, has been developed further to create the

A.J. Mason (✉)
Department of Engineering Science, University of Auckland,
Private Bag 92019, Auckland, New Zealand
e-mail: a.mason@auckland.ac.nz

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_11, © Springer Science+Business Media New York 2013

Optima Predict software system. This proved to be the first step in an ongoing process to develop and deliver operations research tools for assisting ambulance operators. These tools now include a real-time software system, used in a number of cities, that solves an integer-programming model to determine how best to move idle ambulances to improve the future performance of the ambulance operation.

Healthcare is an area of growing importance and cost, around the world, and thus an important area for operations research. By documenting our experiences in delivering operations research solutions to ambulance operators, we hope to encourage both practitioners and researchers to become more involved in this area. Simulation is now a well-established technique but has typically being applied to ambulance operations on a case-by-case approach with varying degrees of realism. Our contribution here is the development and documentation of a generic simulation system that has proven its ability to accurately model the full range of problems we have encountered during a decade of implementation experience. We describe the use of this simulation as part of an optimisation algorithm for base location; this is a novel contribution within the simulation/optimisation area. The widespread availability of real-time vehicle location and status information means that the problem of optimising idle ambulance locations in real-time has become an important practical problem for ambulance operators seeking better performance from their current resources. While there have been exciting new research developments in this area, much work is still required; we hope our analysis and discussion of this problem will encourage both practitioners and researchers to contribute to this research effort.

The remainder of this chapter is organised as follows. In Sects. 2 and 3, we introduce a typical ambulance operation and describe the associated ambulance response process. Section 4 describes the simulation software we developed to model this response process. Some practical applications and comments on this are given in Sect. 5. A new simulation–optimisation algorithm for determining improved base locations is detailed in Sect. 6. Sections 7 and 8 conclude the chapter with a detailed examination of the real-time ambulance repositioning problem. These sections start with a literature review, describe the operations research model we have developed for repositioning and then comment on our experience in implementing this. This chapter then finishes with some concluding remarks.

## 2 A Typical Ambulance Operation

The Metropolitan Ambulance Service (MAS) in Melbourne, Australia, now part of the larger Ambulance Victoria organisation, is a good example of a typical ambulance operator. MAS provides services to almost 4 million residents over an approximately 9,000 km$^2$ area (Wikipedia 2011). In the 2009–2010 financial year, ambulances were dispatched to 330,000 incidents within the MAS response area (Ambulance Victoria 2010a).

Like most ambulance services, Ambulance Victoria operates under contracts that stipulate minimum levels of service as specified by certain performance targets. These targets relate to response time, which is defined as the time interval between receiving a phone call requesting an ambulance and an ambulance arriving at the scene. Ambulance Victoria's performance is measured by their achievement of these various performance targets. These include responding to at least 85% of emergency (code 1) incidents within 15 min, and, for centres with more than 7,500 people, achieving at least 90% for this measure (Ambulance Victoria 2010a). The percentage of calls meeting these targets is a key figure of interest when simulating a possible process change.

Ambulance Victoria staff their vehicles with various combinations of officers with different levels of training. Staff training designations include MICA (mobile intensive care ambulance), qualified ambulance paramedics and ACO (ambulance community officer) (Ambulance Victoria 2011). Both MICA paramedics and qualified ambulance paramedics work full-time, with MICA paramedics having the more advanced training. ACOs typically have less training and work on a casual basis to provide first-aid skills to rural communities. Each ambulance officer operates either as part of a two-person crew or works alone as a single responder. The resulting vehicle types, which include single- and double-crewed MICA, ALS (advanced life support) and ACO vehicles, provide varying capabilities to treat different degrees of injury. Ambulance Victoria publicly reports the types, locations and times of availability of all their vehicles (Ambulance Victoria 2010b).

Ambulance Victoria is facing growing demand for its services. Since 1999–2000, ambulance caseload in Victoria has grown by an average of 5.7%p.a. (Ambulance Victoria 2009). An example of how Ambulance Victoria might improve the system's performance under this increasing demand can be found in the Ambulance Victoria 2009–2010 Annual Report (Ambulance Victoria 2010a). This report lists amongst its statement of priorities the following goal: *"Restructure of metropolitan MICA road response, converting the existing 16 MICA units and 4 single responders to 8 MICA units, 4 MICA Peak Period Units and 14 MICA single responders."* Changes such as these are typical of those that can be tested using simulation systems, such as Predict.

## 3 Overview of the Response Process

The ambulance dispatch process used by Ambulance Victoria, shown in Fig. 11.1, is typical of that found in most ambulance organisations. When a new emergency call is received by the ambulance-dispatch call centre, the call taker assesses the new call and determines its urgency. Different ambulance organisations classify calls in different ways. A common scheme is the commercial ProQA scheme (Priority Dispatch Corporation 2005) in which a call is classified into one of several hundred different types. These are then typically grouped into a smaller number of *call priorities*. Each call classification then determines the *dispatch logic* that should

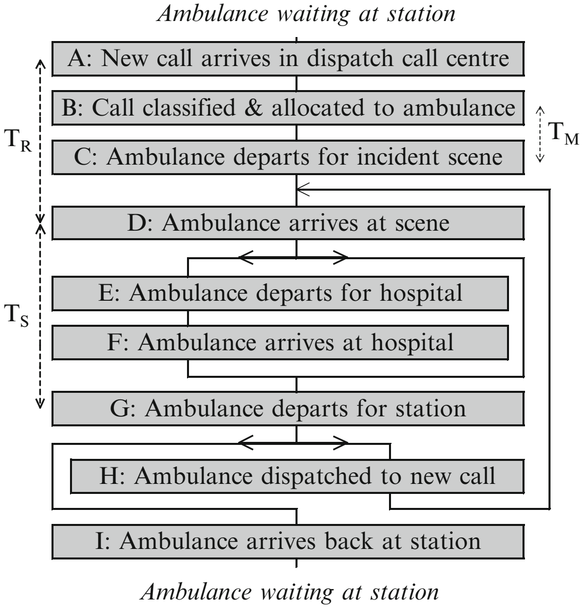

Ambulance waiting at station
A: New call arrives in dispatch call centre
B: Call classified & allocated to ambulance
C: Ambulance departs for incident scene
D: Ambulance arrives at scene
E: Ambulance departs for hospital
F: Ambulance arrives at hospital
G: Ambulance departs for station
H: Ambulance dispatched to new call
I: Ambulance arrives back at station
Ambulance waiting at station
$T_R$
$T_S$
$T_M$

**Fig. 11.1** The ambulance response process, showing the response time $T_R$, the remaining service time $T_S$ until the vehicle becomes free and the mobilisation delay $T_M$

be used to determine which vehicle or vehicles are sent to the accident scene. For a typical ambulance organisation, the dispatch logic specifies that, for all but low-priority calls, the closest vehicle is sent to ensure some form of on-scene treatment is provided as quickly as possible. For high-priority calls, a second vehicle with higher skilled officer may also be dispatched to ensure the correct level of care is provided at the scene.

Once an ambulance vehicle has been given the details of a new call, there is typically a short *mobilisation delay* (also termed *chute time*) $T_M$ before the vehicle begins driving to the accident scene. This is typically zero if the vehicle is already on the road but can be several minutes, for example, during the night when the ambulance officers might be asleep at an ambulance base. The vehicle may then travel to the scene either with lights and sirens operating or without, with this *dispatch priority* being determined by the dispatch logic. Vehicles responding to high-priority calls typically travel at the higher lights-and-sirens speeds, while lower priority calls are responded to at standard traffic speeds. Occasionally the call will be cancelled before the vehicle reaches the scene, in which case the vehicle becomes free and returns to its base.

The most important statistic summarising this response process is the call response time, $T_R$, which is the elapsed time between the call being received at the dispatch centre and a vehicle arriving at the scene. This duration comprises three parts: the call-taking duration, the mobilisation delay and the travel time.

Once at the scene, the ambulance officers assess the patient and determine whether they can be treated at the scene or need to be transported to a hospital. (This at-scene assessment can change the classification initially assigned to the call and can occasionally result in further vehicles being dispatched to the scene.) If no transport is required, then the vehicle becomes free at the scene and typically returns to its designated home base to await its next call. If transport is required, then the patient is loaded into the vehicle and taken to a hospital. Depending on the call classification, this transport operation can be either at lights-and-sirens or normal travel speeds. The vehicle then becomes free after the patient handover process has been completed. Note that once a vehicle has become free, it is typically available to be dispatched either while on the road returning to its home base or once it has reached its home base. The time between the call arriving at the scene and the vehicle becoming available for re-dispatch is termed the service time, shown as $T_S$ in Fig. 11.1.

The dispatch process is complicated by *diversions*. The most common form of diversion—so-called en-route diversion—occurs when a vehicle that is en-route to a low-priority call is sent instead to a higher priority call. The original low-priority call then needs to be processed again to re-dispatch one (or more) vehicles to this call. A much less common procedure is on-free diversion under which the vehicles being sent to a call can be changed whenever a vehicle becomes free. On-free diversion can involve a significant overhead in communicating new destinations to multiple vehicles and, in our experience, does not appear to be commonly used. However, it can occur informally when a vehicle closer to an accident scene overhears a dispatch instruction and requests that they be sent instead.

Each ambulance organisation uses their own rules to determine the dispatch and at-scene behaviour. The various combinations of vehicle types and call priorities can lead to complex sequences of actions. For example, if two single-crewed vehicles are dispatched to a call, then during any transport to hospital, one of the vehicles is typically left parked at the scene while the two officers travel in the second vehicle to the hospital. Both these officers then need to travel back to the scene to reunite the parked vehicle with its officer. Particularly serious incidents can require three officers in the vehicle during the patient transport, with one officer driving and two officers attending to the patient. Typically, two double-crewed vehicles are dispatched to such a call, in which case the fourth officer follows the transporting ambulance in his or her own vehicle. The officers then regroup in their own vehicles at the hospital. This situation becomes more complex when one double-crewed vehicle and two single-crewed vehicles are dispatched to such a call, and thus, a vehicle must be left at the scene to be collected once the transport operation becomes complete. More complicated scenarios can also occur, such as the initial dispatch of a rapid response vehicle to the scene which, after assessing the patient, may then request that additional vehicles be dispatched to provide a higher level of care for the patient. Although these cases may be rare in practice, they exist to better handle the highest-priority calls which are often of the greatest interest to an ambulance organisation. Thus, it is important that any simulation model handles these complexities and that working examples of these cases can be demonstrated to the organisation during the implementation process.

We note that the dispatch of multiple vehicles to a single scene complicates the definition of response time. The first, perhaps more common, definition is *first-vehicle response time* which is computed using the time at which the first vehicle arrives at the accident scene. However, another possibility used by some organisations is to define a *required-skills response time*, which instead uses the time at which sufficient officers of the correct skills have arrived at the scene. It can be useful to report both these times in any simulation results.

The performance targets for response times typically vary by the location of the call (often whether the call is in a metropolitan or in a rural area) and the priority of the call. Metropolitan calls that are designated as life threatening typically have the smallest target response time, while low-priority rural calls have larger target times.

## 4 Simulation Design

The Optima Predict software has been developed using C++ to provide a realistic simulation environment for modelling the ambulance operations described above. It includes a customised simulation engine written in C++, a travel model, a call generator and graphical data analysis capabilities.

The simulation system works by using historical data to run a trace-driven simulation, that is, the calls used in the simulation are real calls (typically a year's worth) that are read in from a stored file. The data used from each call includes the call arrival time, call-taking duration, initial call classification, call location, any updated call classification made at the scene, time spent by an ambulance at the scene, destination hospital to which the patient was transported (if any), the time spent at the hospital, and several other minor fields. A call-cancelled time can also be given, in which case many of the other fields are left empty. The use of this historical data is beneficial for testing and validation of the simulation model and also removes the need to develop and validate statistical models for generating calls with their complex temporal and spatial structures. (Such statistical models are typically required, however, for simulating call growth scenarios; we consider this later in this section.)

A vital component of the simulation is the travel model that computes driving routes and travel times between any pair of locations. This model is based on a road network model that supplies both road layout information and travel times along roads (arcs) at various times of the day, including the morning and evening rushes. Travel times are allowed to vary across the day and across the week. As part of configuring the system for a new city, the travel times are calibrated for both lights-and-sirens and standard travel using historical trip data, including GPS data when this is available (Mason 2006; Mason and Henderson 2010). The accuracy of Predict's travel models have been steadily improving due to the availability of ever-better road models, thanks to the growing demand for in-car and online navigation and ever-faster computers to process these more detailed networks. To illustrate this, the original travel model used for Melbourne, which was developed by the Victorian government, contained less than 18,000 arcs, while the current commercially developed model has several million arcs.

Predict allows modelling of ambulances with different capabilities both in terms of the number and skill level of the associated ambulance officers and the possible capability to transport a patient to a hospital. The rules controlling the dispatch and at-scene behaviour of these vehicles vary from city to city, and so, these are not embedded directly in the C++ code but instead are controlled using the scripting language Python (Python Software Foundation 2011). The simulation executes Python procedures to determine what vehicle(s) to dispatch and the behaviour of these vehicles at the accident scene. This approach gives a very flexible but still fast-running software system.

Ambulance availability is specified in terms of when and at what home base an ambulance is brought into operation and when it ceases to be available. This allows shifts to be effectively modelled and simulated.

The most important source of variability in an ambulance system is the observed demand, being the number, times, locations and priorities of the incoming calls. Because the simulation software reads this data as input, this source of variability can be controlled. Thus, each simulation is essentially deterministic in nature. This allows Predict to accurately replay historical data, which is important in validating the system against historical records. This also assists with variance reduction, an important technique in the analysis of simulated output data (Law and Kelton 1999). However, some stochastic components are required to handle dispatches that differ from those seen in the historic data. In particular, the mobilisation time is typically generated from a user-defined base-specific distribution unless the vehicle is already on the road, in which case a zero mobilisation delay is used. (In some rare cases, such as in the base optimisation algorithm we present below, the historic mobilisation time for a call is used for dispatches to that call from any base.) Further, the user can specify the parameters for a distribution that is used to scale the transport duration predicted by the travel model. However, to minimise variance, such variability is often turned off (and simple means used instead) in runs comparing different system configurations.

A typical Predict run will process a set of historic calls using a given set of base locations, vehicle shifts and rules for dispatch and at-scene behaviours. Predict allows the operator to view the ambulance operation unfolding on the computer screen. Extensive statistics, such as vehicle utilisation and response-time distributions, are collected during the simulation for later analysis and to assist in comparing simulation runs.

An important feature of the Predict software is the detailed graphical analysis capabilities provided. When this software was first shown to St John in 1998, it provided their first visual representation of call locations and response performance. More recent versions of Predict allow this data to be further analysed in multiple ways such as by time of day or day of week, vehicle type, call priority, calls impacted by diversion and other user-defined attributes.

Figure 11.2 shows an example of the graphical analysis that can be performed in Predict. Detailed call-filtering rules have been used to produce this figure from a completed simulation run for a problem based on Auckland, New Zealand. The percentage value in each square region shows the fraction of calls in that area that

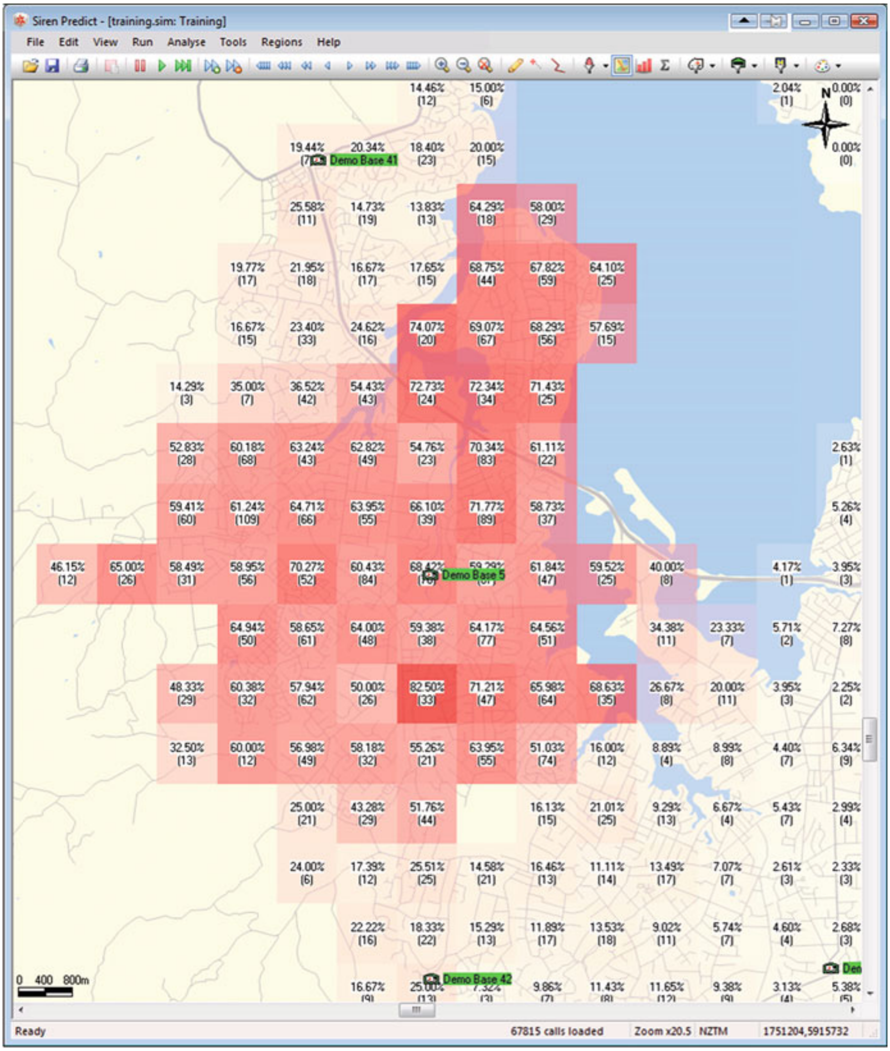


**Fig. 11.2** A screen shot from Predict showing the percentage of calls in each area that are responded to by a vehicle from base "Demo Base 5" located near the centre of this figure

have been served by a vehicle from the centremost base in the figure (base "Demo Base 5"). The value in brackets shows the number of calls that this percentage represents. Only regions with 20 or more calls in total have been included in this display.

Predict is most typically used by clients to test operational changes, such as base location changes or shift changes. Many such tests can use historic call data. However, simulation runs can also be used to predict future performance given forecast changes in call volumes. This requires creating a new set of calls (or modifying an existing call set) to match some forecast call volumes.

Predict contains a flexible call generator module that allows the user to specify a step-by-step process for creating new calls for use in a subsequent simulation run. Each call has a set of fields that must be assigned values. These include the call time, call location, call priority and so on. The call generator allows the user to assign values to these fields either by drawing samples from analytical distributions or by so-called bootstrapping in which samples are drawn from marginal distributions created from a historic call set. For example, to determine a new call's location, the user may specify call arrival rates for each time period for each suburb which are used to generate call arrival times assuming exponential inter-arrival times. Call locations could then be generated assuming a uniform distribution within that call suburb's boundaries or could be sampled from historic call locations for that suburb. Each call's priority could then be sampled from the marginal distribution of priorities formed by considering the set of historic calls in that suburb. The call priority could then be used to define a marginal distribution for the at-scene duration by considering all at-scene durations for historic calls of the given priority. Alternatively, if the user believes the at-scene durations vary by suburb, then they could sample the at-scene duration from historic calls which are in the given suburb and match the call's assigned priority. By proceeding in this manner, the user is able to specify an essentially arbitrary sequence of rules for determining values for the call's fields.

This call generator module has been used successfully by Ambulance Victoria and other clients to generate data sets to assist in planning under alternative growth scenarios and call patterns.

## 5 Practical Applications

Since being developed, the Predict software has been used to model many ambulance operations around the world. As an example of its use, a European operator wanted to investigate how response-time performance would change if single-crewed vehicles were added and the dispatch policy was changed to dispatch these single-crewed vehicles for some classes of call. These vehicles would then request that a transport-capable vehicle be dispatched only if their at-scene assessment of the patient deemed this necessary. This change was tested by modifying embedded rules in the simulation model so that once the single-crewed vehicle arrived at the scene, the at-scene patient criticality (as recorded in historic data) was examined and a transport-capable vehicle dispatched if required. Simulation results suggested this would increase a key performance metric from 62.9% to 71.5%, being an 8.5% increase. When the changes were implemented by the operator, they reported that an 8% performance increase was observed in the field.

Another example comes from Denmark, where the government used Predict simulations to redesign the ambulance operations in the Copenhagen area. The details of this, which we summarise here, are documented by Optima (The Optima Corporation 2011b) and on the Danish Capital Region's website (Region Hovedstaden 2009a).

Historically, 95% of the ambulance services in Denmark have been provided by a single private operator in association with the municipal fire services (Region Hovedstaden 2009b). However, in 2003, the Danish tender complaints board ruled that a new formal tendering process had to be adopted to meet EU competition requirements. The private operator had traditionally been given a lot of freedom in their operating procedures. Performance was traditionally evaluated in terms of a yearly average response time across a large area (Region Hovedstaden 2008), allowing potentially wide variations in response times to occur unnoticed.

In 2008, a decision was taken by the regional health officials in Denmark's Capital Region to take more direct control of the area's ambulance operation by setting up their own dispatch centres. These new centres would be responsible for controlling the approximately 160,000 dispatches made per year over a 2,561 km$^2$ area to a population of 1.6 million (The Capital Region of Denmark 2009). It was also decided at this time to improve the care being delivered by adding paramedics to some of the vehicles.

Taking control of the dispatch operation forced the region's officials to become directly involved in planning their ambulance operations, and so before the tendering process could begin, officials had to decide the numbers and types of vehicles needed across the day and the bases at which these were to be located. Service-level measures were also re-evaluated, with the historical emphasis on average response time being replaced by the more common percentage of calls reached within specified response time targets. Extrapolating arrival rates from 2005 and 2006 data, simulation experiments were conducted to determine the 2009 vehicle requirements across the day for each region. These simulation runs allowed average response times to be calculated and compared with historical data. This analysis was important in assuring the Danish public that the major changes being proposed to their ambulance operations were still going to deliver a high quality of service.

Once the vehicle requirements had been determined, a tendering process was conducted, seeking providers to deliver the required numbers of ambulances and crews across the day. Thus, the tender process was for the provision of ambulances and crews at specific times of the day instead of the traditional hands-off response-time-based approach. The separation between the dispatcher and the vehicle provider allowed for interesting tender conditions, such as the specification of a maximum mobilisation time of 90 s for high-priority calls, with large fines being imposed where these were exceeded.

The tendering process proved to be a complicated one, with new firms submitting tenders but then withdrawing them because of difficulties in concluding collective agreements in the labour market. Media interest was significant, with public criticism of the process being made by some of the parties who lost contracts. However, the tendering process was completed, and the new process commenced operation on 1st September 2009. To meet public demand for information, a website (Region Hovedstaden 2011) was created to document the performance of this new system, and provide information on how this performance was being measured. This shows that the new system is performing well. For example, in March 2011, 90% of ambulances arrived within 9.49 min, well within the 90% 13 min target.

It is interesting to consider how the actual and simulated performances compared. Actual call volumes in 2009 were 8% higher than forecast, and the proportion of high-priority calls was also higher at 54% compared with a forecast 44%. The region responded to these changes by increasing the number of vehicle hours by 6%. Re-running the simulations developed in 2007 with this corrected data gives a mean response time of 6 min 35 s, compared with the measured value in the field of 6 min 4 s, an error of just 8% (The Optima Corporation 2011b).

An important part of any project such as this is communicating the benefits (or otherwise) of possible changes to decision makers. For example, Ambulance Victoria has used Predict's graphical displays to support requests to their funding body for additional resources. More recently, Predict's visualisation capabilities were used to convince elected US officials that they should reject demands to move a resource to an affluent (and politically important) area (The Optima Corporation 2010). Simulation results showed that such a move would improve average response times for the small number of calls originating in the area but worsen performance for the much greater numbers of calls occurring in other parts of the county.

An important issue in simulation is the interpretation of results. The Predict software uses the method of batch means (Law and Kelton 1999) to estimate confidence intervals around performance measures. However, we have found that there is little awareness of the importance of statistical significance amongst practitioners when interpreting either simulated or actual performance results. For example, the performance reports of one operator are based on the daily percentage of on-time calls within zones that can have, on average, as few as 5 calls per day. The percentage of calls reached on time in each zone is reported back to management twice a day, giving an often false sense of success or failure arising purely from statistical variability. Such reporting requirements are often forced upon the ambulance organisation by their funding bodies, suggesting that there is a need to better educate ambulance operators so that they can better interpret performance figures and can also engage more actively in the contractual process to ensure statistically meaningful targets are developed.

## 6 Simulation–Optimisation for Base Locations

An important problem facing ambulance operators is to determine where their bases should be located. We were interested in addressing this problem for an ambulance operator that leased many of their base locations and thus was able to move bases relatively easily. This operator was interested in using the simulation system—with its validated model of their operations—to help determine new possible base locations. Unfortunately, the simulation approach is not itself a prescriptive tool in that it does not by itself directly answer such questions. However, we now describe a simulation–optimisation approach we have developed to find good-quality solutions to this problem while preserving the high-fidelity modelling offered by simulation.

The problem of where to locate ambulance bases, or, somewhat equivalently, which of a set of bases to assign vehicles to, has been addressed many times in the literature. Brotcorne et al. (2003) have completed an extensive review of the ambulance location and relocation models that existed up to 2002. See also Wright et al. (2006) for a more recent survey of operations research models in the wider context of homeland security and Berman and Krass (2002) for a review of general location models including ambulances. Goldberg (2004) provides an excellent layperson's overview of the problems faced by ambulance operators, including base location, and the solution approaches developed for these problems. See also the recent review by Henderson (2010).

To solve the vehicle location problem, we need to address (1) the prediction of the system performance for a given vehicles-to-locations assignment and (2) the assignment of vehicles to locations to maximise this performance. Unfortunately, these requirements conflict in the sense that if we use simplified models to predict system performance, then we can use integer-programming optimisation procedures that guarantee an optimal solution, albeit to the simplified problem. However, as we improve our modelling of the system performance, the problem becomes increasingly complicated and correspondingly more difficult to optimise.

In an ideal world, all vehicles would be idle, waiting at their bases, when the next call arrives, and so, an emergency call would be served by the closest occupied base. If this were the case, then for any given set of vehicle waiting locations, it would be easy to determine those areas which are *covered* in that the area's closest vehicle is within some specified driving time. This coverage concept underpins the early set covering approaches presented by Toregas et al. (1971) and Church and ReVelle (1974). Various modelling extensions have been proposed, including "double coverage" or "backup coverage" models (Hogan and ReVelle 1986) where the formulation counts the areas covered by two or more vehicles, models for handling multiple response-time targets (Gendreau et al. 1997), models for variable travel times (Daskin 1987) and probabilistic coverage models (Daskin 1983) where ambulances are modelled as being busy or free with some probability, and the model calculates the probability of an area being covered by at least one vehicle.

Optimisation models that explicitly consider ambulances being unavailable require estimates of the busy probabilities, where the busy probability for a base is the probability that the base will have no vehicles available to serve a call. An established approach for predicting these is the hypercube model (Larson 1974) and its approximations and extensions (e.g. Larson 1975; Budge et al. 2009). The hypercube model constructs a Markov chain that models transitions between states formed by all possible combinations of busy and idle ambulances. For small numbers of vehicles, it is possible to calculate the probability of the system being in any state and use this to predict system performance for any given assignment of vehicles to bases. When the number of vehicles becomes too large to consider all possible states, an approximate model is typically used which uses so-called correction factors to predict the response times in each suburb.

Budge et al. (Ingolfsson et al. 2008) have developed an iterative process that solves a convex optimisation problem (assuming known busy probabilities) to allocate vehicles to bases, updates these busy probabilities for the new vehicle allocation using hypercube and then repeats this process iteratively. A recent alternative to the hypercube model, which performs better in simulation experiments for problems with light to moderate vehicle utilisation, is presented by Restrepo et al. (2009).

The modelling approaches detailed above are limited in their accuracy because of their dependence upon analytical (or simpler covering) models for predicting performance. For example, such models assume exponential service times (including the driving time component; see Budge et al. (2010) for an analysis of travel times) and that vehicles are always dispatched from their bases. Our experience has been that ambulance operators are wary of such simplifications and place greater trust in a simulation model that they have been able to verify with historical data. Indeed, authors proposing analytical methods still recommend that simulation be used in any final decision making (Restrepo et al. 2009). Hence, we have focused on a simulation–optimisation approach.

The use of a closely coupled simulation/optimisation algorithm for solving ambulance logistics problems is not new. For example, Maxwell et al. (2010) and Alanis et al. (2010) have found good vehicle-to-base assignments by simulating a large number of possible solutions. Silva and Pinto (2010) developed a simulation model for Belo Horizonte, Brazil, using the Arena system (Rockwell Automation 2011). They performed a number of what-if analyses, including using the OptQuest optimiser (OptTek Systems Inc 2011) to determine how many vehicles to place at each base to meet prescribed performance targets.

A common approach in the simulation–optimisation literature, which we will adopt, is to use a local search or other heuristic to find a good local optimum. In its simplest form, this involves repeatedly evaluating a number of neighbours of some current solution and then choosing the best of these. An alternative approach is to fit a response surface that predicts how the system performance varies with the parameters being optimised. This and other simulation–optimisation options are discussed in more detail in Andradóttir (2006b) and Fu et al. (2005). In the next section, we formally present our problem and a novel solution approach which implements a local search that uses an approximate local response surface to estimate the performance of neighbouring solutions.

## *6.1 Base Location: Problem Definition and Algorithm*

Consider some simulation model of a system that includes all the real-world complexity of multiple vehicle types, vehicle shifts, multiple vehicle dispatches, diversions to higher priority calls, dispatching of on-road vehicles and cancelled calls. We fix all the operating parameters for our ambulance system except for the locations of the bases. We then wish to alter the locations of these bases to maximise the performance of the system. This problem differs subtly from those discussed above in that we are interested in moving existing base locations, not assigning vehicles to existing bases, and so, we have a continuous problem in which there are infinitely many possible base locations. We note that this objective has discontinuities in that even small changes in a base location can alter the vehicle used to serve a later call, resulting in an arbitrary change in performance. (This suggests that gradient-based simulation–optimisation methods, such as those discussed in Fu (2006), may not be the most useful for this problem.) In practice, we consider a discretised version of the problem in which we reduce the number of candidate base locations to some manageable but still large number by requiring bases to be located at nodes on the road network (where these nodes include both intersections and bends in the roads). We only move the base locations without changing other parameters such as the allocations of vehicles to bases or the shift times for the vehicles, and so, the vehicles will continue to operate as currently but with their bases moved to new locations.

We adopt a sample path optimisation approach (Fu et al. 2005) in which a single representative set of input data is used to form a deterministic equivalent problem. We follow the approach typically adopted by Predict users and use a single year's worth of historic calls as input data, with just a single replication being performed. However, to validate our approach, we will solve the model using one set of training calls and then test its solution using another evaluation set.

Let $B = \{1, 2, \ldots, |B|\}$ denote the set of bases and $C$ denote the set of historic calls that we will use in our optimiser's simulation runs. Each base $b \in B$ can be located at any node $v_b \in V$, where $V$ is the set of nodes in the road network. We assume that the simulation run uses the input data $(C, v_1, v_2, \ldots, v_{|B|}, \ldots)$, being a set of historic calls $C$, base locations $\{v_b, b \in B\}$ and other data such as shifts and road speeds.

Given this input data, the simulation then produces a range of data outputs. These include system performance measures such as the percentage of calls responded to within their target times. We generalise this by computing an average call score, where the score for a call depends on its target and actual response times. Rather than recording a call as being either on time (1) or late (0), this score function gives a numerical value that decreases smoothly with increasing response time. This has a number of advantages. It creates a smoother objective surface that is more suited to the neighbourhood search approach we use. Furthermore, Erkut et al. (2008) present a number of computational and medical arguments for using such non-0/1 reward values.

Let us denote the full set of simulation outputs by $S$. This output $S$ includes the call scores $f_1, f_2, \ldots, f_{|C|}$, where $f_c = f_c(t_c(v_1, v_2, \ldots, v_{|B|}))$ is the score that results for call $c \in C$ given the associated response time $t_c(v_1, v_2, \ldots, v_{|B|})$. Each call score depends on the call's target time and response time, where the latter depends on all the simulation inputs, including the base locations we are optimising.

We can now write our problem as follows. Given a set of simulation inputs, including the set $C$ of historic calls, we seek a set of base locations $v_1, v_2, \ldots, v_{|B|}$ that maximises the sum of call scores $F(v_1, v_2, \ldots, v_{|B|}) = \sum_{c \in C} f_c(t_c(v_1, v_2, \ldots, v_{|B|}))$. This gives rise to the following base location (BL) model:

$$\mathrm{BL}(C) : \max_{v_1,v_2,\ldots,v_{|B|}} F(v_1,v_2,\ldots,v_{|B|}) = \sum_{c\in C} f_c(t_c(v_1,v_2,\ldots,v_{|B|})) \tag{11.1}$$

We adopt a neighbourhood search strategy for solving BL($C$). We will start the search using the existing base locations. We expect to find a good solution but offer no guarantees of optimality. Instead, we are willing to accept some solution that is within the general neighbourhood of this starting solution and thus is not a major change over the existing base locations.

One approach to perform this neighbourhood search would be to perturb the location of one base, run the simulation to see if this improves the performance and then keep the change if it is better. This local search process would need to be repeated many times to find a good set of base locations. An obvious disadvantage of this method is the large time required to obtain performance values; a simulation of a year's worth of call data for a large city can take around 10 or 20 min to complete, and so, we could only test three changes per hour. To counter this, we could construct a multidimensional response surface (Andradóttir 2006a) which uses completed simulation runs to predict the system performance for new base configurations. This is the approach taken in the general purpose simulation optimiser OptQuest (OptTek Systems Inc 2011) sold with many commercial simulation packages. However, for our problem, many runs would still be required before the response surface was well developed. Instead, we developed a customised local response surface to predict performance for neighbouring solutions. This is possible because we can exploit detailed simulation trace information. This helps negate the slow runtimes often associated with simulation–optimisation techniques.

To formulate our approximate response surface, we assume the simulation output $S$ includes detailed call response information of the form $C_{\mathrm{road}}, C_1, C_2, \ldots, C_{|B|}$, where set $C_{\mathrm{road}}$ contains all calls for which the first arriving vehicle was on the road (i.e. not at its base) when it was dispatched and $C_b$, $b \in B$ contains those calls for which the first arriving vehicle was dispatched from base $b \in B$.

Using this partitioning of the calls, we can write our objective function as:

$$F(v_1,v_2,\ldots,v_{|B|}) = \sum_{c\in C_{\mathrm{road}}} f_c(t_c(v_1,v_2,\ldots,v_{|B|})) + \sum_{b\in B}\sum_{c\in C_b} f_c(t_c(v_1,v_2,\ldots,v_{|B|})) \tag{11.2}$$

However, if we assume call $c \in C_b$ will always be served by a vehicle at base $b$, we can approximate the response time $t_c(v_1,v_2,\ldots,v_{|B|})$ by the simpler response-time estimate $\hat{t}_c(v_b)$ that depends on the location $v_b$ of just the single base $b$, giving

$$F(v_1,v_2,\ldots,v_{|B|}) \approx \sum_{c\in C_{\mathrm{road}}} f_c(t_c(v_1,v_2,\ldots,v_{|B|})) + \sum_{b\in B}\sum_{c\in C_b} f_c(\hat{t}_c(v_b)) \tag{11.3}$$

Equation (11.3) assumes call $c$'s score $f_c(\hat{t}_c(v_b))$, $c \in C_b$ depends on the response time $\hat{t}_c(v_b)$ when travelling from node $v_b$. As we discuss shortly, changing the base locations may invalidate this assumption. However, as we will see, (11.3) still

**Algorithm 1:** A local search algorithm for base locations. The constant $0 < \varepsilon < 1$ defines a finish condition, while function $V_b()$ uses the simulation outputs $S^i$ to determine a new (hopefully) better location for each base $b$.

```
Basic Local Search
Initialization
Let $i = 0$
Let $(v_1^0, v_2^0, \ldots, v_{|B|}^0)$ be the set of initial base locations
Simulate: $(C, v_1^i, v_2^i, \ldots, v_{|B|}^i, \ldots) \rightarrow S^i = \{f_1^i, f_2^i, \ldots, f_{|C|}^i, C_{\text{road}}^i, C_1^i, C_2^i, \ldots, C_{|B|}^i, \ldots\}$
repeat
    Use the simulation output to improve each base location
    for each base $b \in B$
        Determine a new location $v_b^{i+1} = V_b(C_b^i)$
    Simulate using the new base locations
    Let $i = i + 1$
    Simulate: $(C, v_1^i, v_2^i, \ldots, v_{|B|}^i, \ldots) \rightarrow S^i = \{f_1^i, f_2^i, \ldots, f_{|C|}^i, C_{\text{road}}^i, C_1^i, C_2^i, \ldots, C_{|B|}^i, \ldots\}$
    $F^i = \sum_{c \in C} f_c^i$
until $(F^i - F^{i-1})/F^i < \varepsilon$
Return the best set of base locations generated
Let $j = \arg\max\{F^1, F^2, \ldots, F^i\}$
return $(v_1^j, v_2^j, \ldots, v_{|B|}^j)$
```

provides a useful estimate. Note that $\hat{t}_c(v_b)$ depends on the road speeds when call $c$ arrives, the call's priority, its call-taking time and the mobilisation delay at base $b$. These are all assumed to be invariant across simulation runs. Thus, in the simulation, the mobilisation delay for call $c$ does not change even if the call is served by a different vehicle coming from a different base.

Our local search proceeds in an iterative fashion where at each iteration $i$, a different choice of locations for the set of bases $B$ will be evaluated. Our search algorithm is shown in Algorithm 1, where superscript $i$ is used to indicate data values at iteration $i$.

An important component of this algorithm is the function $V_b()$ that uses the output $S^i$ from the $i$th simulation run to determine a new improved location for each base. We estimate the impact of moving each base $b$ by considering just the term associated with base $b$ in (11.3). Of course, in general, moving a base may change the calls that are closest to that base, hence changing $C_b^i$. Furthermore small changes in base locations may change the availability or location of a vehicle at the time of a call's dispatch, causing further changes in $C_b^i$ or in $C_{\text{road}}$. However, we might expect this approximation to be a useful predictor if the base location changes are small. Furthermore, and most importantly, this approximation allows the impact of moving a base to be estimated without running the simulation. Our approach for $V_b()$ is to simply enumerate all nearby nodes $v : |v - v_b^i| \leq r$, where $|v - v_b^i|$ denotes the Euclidean distance between nodes $v$ and $v_b^i$ and $r$ is some user-defined radius. For each of these nodes $v$, we evaluate the travel time and hence expected response time

$\hat{t}_c(v)$ and call score, for each call $c \in C_b^i$, that would result if the base were located at node $v$. The best of these nodes gives our new base location. Thus, we have

$$V_b(C_b^i) = \underset{v \in V : |v - v_b^i| \leq r}{\arg\max} \sum_{c \in C_b^i} f_c(\hat{t}_c(v)) \tag{11.4}$$

Equation (11.4) ignores factors such as changes in the resulting call-to-vehicle assignments. Therefore, after finding new locations for all the bases, we perform a full simulation run to accurately evaluate the combined impact of these new locations. This optimise-then-simulate process is iterated until no further improvement in performance is obtained.

Our local search approach has been tested using a model that is similar to that of a real system. We solved a problem with approximately 250,000 calls over a one year duration, of which about 65% of calls were high priority, about 25% medium priority and the remaining 10% low priority. About 80% of calls were responded by vehicles from their base (well below the 100% that would be assumed by a hypercube model). New base locations were found using one set of calls, and then, the effectiveness of these new base locations was tested using a different but similar test call set. Simulation using this test set showed an increase within just one iteration of approximately 8% in the number of high-priority calls responded to within 8 min and an approximately 2% increase in the number responded to within the 13 min target. Further iterations made improvements for the call set being simulated, but not for the test set. The improvements generated are significant for an ambulance operation and indicate the potential value of this approach.

## 7 Ambulance Repositioning

In a traditional ambulance operation, the vehicles follow a static return-to-home-base policy under which each vehicle has a designated home base to which it returns after completing each call. There are a number of factors that might motivate a more dynamic approach in which changes are made to these vehicle waiting locations during the day. These include:

1. Changes in the call arrival rate during the day
2. Changes in the spatial distribution of incoming calls during the day
3. Changes in the ambulance travel times during the day caused by changing road congestion patterns
4. Planned changes in the number of available vehicles
5. Random changes in the number of available vehicles caused by the dispatching to and completion of calls

The first four factors above can be addressed at the planning stage when the number and locations of the vehicles are being determined. For example, Repede and Bernardo (1994) break the day into separate time periods and solve an optimisation problem to determine the number of ambulances required in each period and their locations. A similar approach is taken by Schmid and Doerner (2010) to planning the vehicle locations in Vienna except that they solve a single full-day model that takes into account the distances vehicles must travel when moving from one configuration into the next. Rajagopalan et al. (2008) consider a similar problem but use a hypercube-based model to evaluate system performance. They do not have any vehicle repositioning costs and so can solve each time period independently using a Tabu search algorithm that allocates vehicles to bases while minimising the number of vehicles required.

We consider here the more complex problem of repositioning vehicles in real time in response to changing vehicle availability caused by the arrival and completion of calls. This problem arises from the recognition that a sequence of dispatch operations may deplete the number of available ambulances in an area, resulting in a poor response time for future calls in that area. Careful selection of the destination bases for newly available vehicles can help repair these holes in ambulance coverage. These repositioning strategies should, ideally, take into account all the factors listed above and thus are an extension of the planned repositioning models discussed above.

Repositioning approaches can deliver significant benefits. For example, Alanis et al. (2010) cite a simulation study showing that repositioning in Edmonton, Canada, generates performance improvements equivalent to having 8 additional ambulances operating 24 h a day. Such results suggest that repositioning will play an increasingly important role in future ambulance operations. We recognise, however, that repositioning can be disruptive for ambulance staff; see Henderson (2010) for comments on this as part of a wider discussion of operations research contributions to ambulance logistics. We next review the literature and then present the repositioning solution developed by Optima.

Kolesar and Walker (1974) proposed one of the first repositioning algorithms, which they designed for fire companies. One of the first proponents of vehicle repositioning for ambulances was Stout (1983), who used the term "system status management" to refer to "the formal or informal systems, protocols, and procedures which determine where the remaining ambulances will be when the next call comes in." Stout details a process where the preferred base locations for vehicles are determined by local experts for each hour of the day and for each number of idle ambulances, producing a *system status plan*. While the terminology is somewhat interchangeable, we reserve this term for such a repositioning scheme that specifies the number of ambulances $x_{t,f,b}$ to be allocated to each of the $|B|$ bases $b \in B = \{1,2,\ldots,|B|\}$ if $f = 1,2,\ldots,n$ of the $n$ ambulances are free during one of the $T$ distinct time periods $t = 1,2,\ldots,T$. Any such plan must necessarily satisfy $\sum_{b\in B} x_{t,f,b} = f,\ \forall\ f = 1,2,\ldots,n;\ t = 1,2,\ldots,T$. We consider such a repositioning plan to be an *offline* plan in the sense that any optimisation of the plan occurs well before it is used.

Unless carefully designed, system status plans can unnecessarily increase the miles driven by the ambulance vehicles. The driving distance can be reduced by using a nested plan for which $x_{t,f+1,b} \geq x_{t,f,b} \forall\ b \in B, f = 1,2,\ldots,n-1$, meaning that when a new vehicle becomes free, the desired configuration can be achieved by sending this vehicle to the unique base $b \in B : x_{t,f+1,b} = x_{t,f,b} + 1$ without having to move any of the other vehicles.

A relaxed version of the nested system status plan is proposed by Gendreau et al. (2006) who define a maximal expected coverage relocation problem (MECRP) that is solved offline to generate a plan. This model allows up to $\alpha_f$ of the $f$ vehicles available in state $f$ to be moved when going to state $f+1$, giving the constraint $\sum_{b \in B} |x_{f+1,b} - x_{f,b}| \leq 1 + \alpha_f$. (Note that Gendreau et al.'s model does not have any time dependence, and so, we have dropped $t$ in this case. As before, $f$ denotes the number of free vehicles.) They formulate an integer programme which seeks a set of vehicle configurations, one for each state $f$, $f = 1, 2, \ldots, n$, which together maximise the weighted coverage $\sum_{f=1}^{n} C(x_{f,1}, x_{f,2}, \ldots, x_{f,|B|}) q_f$, where $C(x_{f,1}, x_{f,2}, \ldots, x_{f,|B|})$ is the number of calls that are within the target driving time of an occupied base under the vehicle-to-base assignments $x_{f,b}$, $b \in B$ and $q_f$ is the proportion of time there are $f$ vehicles free, as estimated using a binomial distribution. Their model does not incorporate travel costs and, because it optimises coverage, will place only one vehicle at any base. After solving their integer-programming model for different variants of a small problem in Montreal, Canada, with between $n = 3$ and $n = 6$ vehicles, Gendreau et al. use simulation experiments to test the policy this generates. Their simulation solves a simple transportation problem to determine how to move the vehicles from the current to their new configuration for each change in the number of free vehicles $f$. They show that using a strictly nested plan ($\alpha_f = 0 \; \forall \; f$), they can increase the number of calls responded to within 8 min by between 5% and 8% over a base-line solution with no vehicle repositioning, depending on the problem variant they are considering. These values change to between 14% and 23% when non-nested system status plans are permitted. We note, however, that no details are provided on how vehicles were allocated to bases in the base-line solution, and thus, the quality of their starting point and hence of their reported improvements is difficult to assess.

As well as their work on offline system status plans, Gendreau et al. (2001) also present an online repositioning scheme in which an integer-programming model is used in real time to reposition the idle vehicles after each new call arrives. Unlike system status plans, online optimisation systems are able to explicitly calculate the cost of moving vehicles from their current positions and can balance this against the coverage benefits provided by multiple alternative candidate vehicle-to-base configurations. The Gendreau et al. online model requires all demand points to be covered within a drive time $r_2$ and some proportion $\beta < 1$ of the demand to be covered within a drive time $r_1 < r_2$. They seek to maximise the proportion of demand that is double covered within the drive time $r_1$, less some penalty term associated with changes in vehicle locations. Ideally, the model would be solved in real time every time the number of free vehicles changes. However, because the model is slow to solve, they pre-solve in parallel (using Tabu search) a large number of problems corresponding to each possible currently idle vehicle becoming busy. The vehicle movement penalty terms, which get updated before each new solve, are chosen to discourage undesirable features such as moving the same vehicle multiple times, making round trips and long repositioning movements. The authors conducted a detailed simulation study of the effectiveness of their approach in which vehicle repositioning was performed after each call arrival (but, apparently, not on other events such as call completions). They found that repositioning occurred after 38% of the call arrival events, with an average of about two vehicles (of the 40 or so available on average) being repositioned. Unfortunately, they do not report the gain achieved by their repositioning approach when compared to a static return-to-base policy.

Nair and Miller-Hooks (2009) consider an offline relocation model to generate repositioning policies for Montreal. They solve a location–relocation integer-programming model that attempts to allocate vehicles to meet single- and double-coverage requirements for demand locations under a number of network states, where a network state represents a unique combination of the call arrival distribution, the number of available vehicles and the road travel speeds. The model explicitly considers the movement of vehicles to new waiting locations whenever the system changes from one state to another, which is assumed to happen with known probabilities. They solve their problem for a system with three distinct states (the nature of which is not well described); the size of the model prevents more states being used. After making a number of approximations, they then give an analytical model for predicting the performance of the solved system. These show only minimal benefits from repositioning. The lack of any simulation results makes it difficult to quantify the value such an approach might offer to an ambulance operator.

Andersson and Värband (2007) have adopted an alternative approach in which they develop their own “preparedness” measure to predict system performance. They present work showing how this can be used to assist in both dispatch and repositioning decisions.

An alternative approach to the vehicle repositioning problem is dynamic programming; see Berman (1981a,b) for the first such models for ambulances. Compared with integer-programming approaches, these models have the advantage of more accurately modelling the stochastic elements of the problem. Zhang et al. (2009, 2012) have extended these models and used them to gain interesting insights. Optimal policies developed for small problem instances with just a single vehicle possess a number of interesting features, many of which are useful for designing repositioning algorithms for larger problems. Firstly, when deciding whether to move a vehicle, it is important to consider not just the change in coverage achieved by the move but also the call density along the vehicle movement path. For example, moving from a quiet area to a very busy area may not be optimal if it requires driving through an area with a very low call density. Where moves are made, the optimal vehicle paths are often chosen to pass through areas of high call density, even if this extends the time required to reach the destination. We are not aware of any existing repositioning models that take this effect into account. Secondly, the importance of this effect varies with the call arrival rate. Low call arrival rates make it likely that the vehicle will have completed its repositioning before the next call arrives, and thus, the optimal policy for a low call arrival rate is more likely to include repositioning operations even if these move vehicles through areas of low call density. Thus, a high call arrival rate appears to reduce the opportunities for repositioning. Third, Zhang et al. present an example showing that while achieving a high call coverage is obviously important, the system performance is also improved by reducing the service time. Thus, an optimal vehicle waiting location must consider both coverage and the expected response time to all calls, not only those reached within the target time. Zhang et al. have also considered problems with two vehicles and show that, for example, when one vehicle is dispatched to a call, any decision to reposition the other vehicle should be delayed until it is known if the first vehicle will become free at the scene or must first complete a patient transport operation. The detailed real-time tracking of vehicle status found in most modern ambulance operations means that complex dispatch policies such as this would be possible to implement if the associated optimisation models could be formulated and solved. How to best use such operational details when repositioning larger numbers of vehicles remains an important research question.

While dynamic programming can give valuable insights, the solution approach becomes impractical for realistic-sized problems. Instead, approximate dynamic programming approaches can be used in which the expected performance values associated with the system being in some state are computed approximately, for example, by using a weighted set of basis functions to approximate the optimal value function. Such systems can then perform online optimisation to choose the best of multiple reconfigurations given the costs of moving vehicles from their current positions. Maxwell et al. (2010) (see also Henderson (2010)) have made significant contributions in the application of approximate dynamic programming to ambulance repositioning. They present an approximate dynamic programme for ambulance repositioning for Edmonton and for a larger unnamed city. Using simulation, they showed that by repositioning ambulances when they became free, they could achieve an approximately 4% performance improvement for Edmonton over a good return-to-base solution in which the vehicle-to-base assignment had been determined heuristically using enumeration. This figure dropped to 2% for the larger city. Although these figures appear small, they showed that the benefits from repositioning for the Edmonton problem were similar to those gained by deploying another two additional vehicles and thus represent significant potential savings. They also showed that allowing repositioning moves to be made for all idle vehicles (not just the newly free vehicle) allowed a further 3% improvement. In subsequent work (Maxwell et al. 2011a,b), they show that they can obtain better results using a simpler approximate value function, the parameters of which are tuned using a Nelder–Mead local search algorithm instead of the more standard method of fitting to simulated results. This simpler model produces a surprisingly simple policy in which some estimated marginal benefit of adding one more vehicle to a base is used to determine the base to which a newly free vehicle should be assigned. Because they ignore vehicle movement costs, this policy would be equivalent to a system status plan if all vehicles were being repositioned. In simulations of ambulance operations in Auckland, New Zealand, (Zhang L, Personal communication, 2011) has shown that he can obtain better results by discarding the basis functions and then using local search to directly manipulate the base ordering. Research in this area is continuing.

An alternative approach has been proposed by Alanis et al. (2010) who have developed a Markov chain model for predicting ambulance performance under a system status plan. Their model assumes that for each number of free vehicles $f$, there is some given assignment of these vehicles to waiting locations. When the vehicles are in this configuration, the system is said to be *in compliance*. They model the system as a Markov chain with states corresponding to $f = 0, 1, 2, \ldots, n$ free vehicles which are either in or out of compliance. Whenever a call arrives or a service completes, the system is assumed to go out of compliance, and so, vehicle repositioning is required to regain compliance. While the Markov chain allows a call to arrive when the system is out of compliance, the underlying calculations used for determining service and repositioning times assume that the system is in compliance immediately before a call arrival or a call completion occurs. Thus, rates at which calls are completed and the system returns to compliance can be estimated using a careful analysis of expected travel times and times to reposition the vehicles assuming the system is in one of the restricted set of states formed by starting in an in-compliance state and having either one vehicle become busy (with a probability for each vehicle being given by the proportion of calls closest to that vehicle) or a vehicle become free (either at a hospital or at a demand location). A procedure based on convolution is given to predict the response-time distribution after the Markov chain probabilities have been calculated. Using comparisons with simulation runs, they have shown that this analytical model gives very good predictions of the response-time distributions. This approach is an important step forward in developing good offline algorithms for system status plans.

## 8 Real-Time Repositioning

Optima has developed an online real-time optimisation system, known as Optima Live, that is being implemented in Canada, Australia and Lee County (USA) to make real-time repositioning decisions. This optimisation system integrates with the ambulance operator's dispatching software, ensuring the latest vehicle positions and statuses are always available. An online optimisation model (described below) uses this data to generate repositioning recommendations which are then presented to the operators for their approval and subsequent automatic transmission to the vehicles. The repositioning policy generated using this model differs from a system status plan in that multiple vehicle-to-base configurations can be generated for the same number of free vehicles, with the optimised plan then being chosen based on movement costs given the current vehicle configuration. While some other software exists for moveup, Optima Live is, to our knowledge, the first such system being actually used by multiple ambulance operators that optimally solves an all-vehicle optimisation problem to generate its recommendations.

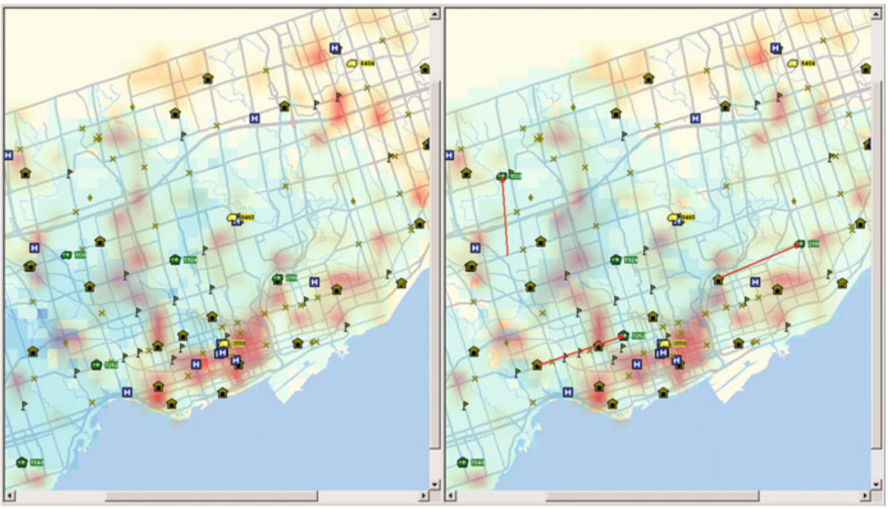


**Fig. 11.3** The Optima Live current view (*left*) and future view (*right*) showing call arrival rates (*red*) and vehicle coverage (*blue*). The three recommended vehicle repositioning moves are shown using *arrows*. These moves improve the vehicle coverage in the top right which is an area predicted to have a higher arrival rate than average

An important feature of the optimisation system is its strong visualisation capabilities that can portray both the current call coverage and the coverage expected if the recommended vehicle moves are completed. Because repositioning takes time, it is important to consider changes in the availability of vehicles that are likely to occur in the near future. Because the status and position of each vehicle are known, it is possible to predict the likely vehicle positions and availabilities once their current activities are completed. This calculation takes into account shift starts and finishes, meal breaks and vehicles completing current calls. Ideally, a scenario tree would be developed that modelled multiple alternative futures, but this would further complicate an already difficult optimisation problem. Instead, just a single future scenario is determined, and this future view is displayed to the user in the form of a predicted future call coverage map. This visual display also shows predicted call arrival rates generated using a forecasting model based on historic data. This combined future call/coverage display helps users identify areas that have poor coverage and a high expected future call demand and thus could benefit from repositioning. An example of such a display is shown in Fig. 11.3.

The data used to construct the call/coverage display is also fed into the repositioning optimisation model. This model is similar to that of Gendreau et al. (2006) but extends their work in a number of directions. While Gendreau et al. seek to maximise double coverage, the Live model allows the user to specify more general concave piecewise linear functions specifying the reward for each demand location in terms of the number of vehicles covering that location. If there are different classes of vehicles, such as when only some vehicles are transport capable, then separate piecewise linear coverage functions can be specified for (combinations of) these different vehicle classes, with a weighted combination of these piecewise functions then being used in the objective. (This approach is similar to that found in Schilling et al. 1979.) Thus, for example, the user can maximise some combination of the 8 min coverage provided by first-responder vehicles and the 20 min coverage provided by transport capable vehicles. The objective includes terms for the total reward resulting from the new vehicle positions and a penalty term that costs the vehicle movements. As in Gendreau et al. (2001), this second term penalises undesirable vehicle movements such as those involving a long travel distance, those resulting in repeated moves for the same vehicle and those that change an en-route vehicle's destination. Repositioning moves that take a vehicle too far from its designated home base can also be penalised. More details of this model are given in Appendix.

While the focus of Live has been on improving response times, the software also includes tools for monitoring upcoming meal breaks and ends of shifts. The repositioning optimisation can use this information to direct vehicles back to their home bases in anticipation of these events.

As mentioned above, Optima Live is being used in Lee County, USA. Since introducing Live, Lee County reports an 11 s decrease in average response time and a 3% improvement in their on-time performance during a period in which call volumes increased by 5.8% (The Optima Corporation 2010).

## 9 Conclusions

This chapter has reviewed a number of contributions made in the application of operations research techniques to problems faced by ambulance operators. The important problem of allocating vehicles to home bases, which many operators solve using what-if simulations, is still an area of ongoing research. Despite the availability of many integer-programming-based models, experienced researchers in the area still rely on relatively unsophisticated enumerative approaches based on simulation, hypercube or related models to solve this problem. Simulation–optimisation techniques such as that presented here can help by exploiting the fidelity provided by a validated simulation but can offer no optimality guarantees.

The problem of how to optimally reposition ambulances during the day to maximise on-time performance is a complex and challenging stochastic optimisation problem which can currently only be solved to optimality for small theoretical instances. The development of approximate dynamic programming and Markov chain models for this problem are two new significant research contributions in this area but have currently only been used for developing restricted classes of policies. Integer-programming models are more flexible but more limited in their ability to accurately model the stochastic elements of the problem. There is now substantial evidence that repositioning can improve on-time performance. However, ongoing research is required to fully exploit the possibilities this offers.

The impact of the operations research contributions discussed in this chapter is perhaps best illustrated by a quote from Ambulance Victoria. Ambulance Victoria's 2010/2012 strategic plan includes a stated goal to "Improve the efficiency of the response process by … more dynamic deployment of resource units, building on current use of the Siren [i.e. Optima Predict] decision support system" (Ambulance Victoria 2009). It is very pleasing to see the benefits of operations research being recognised at such a strategic level.

**Acknowledgements** The author would like to acknowledge his colleagues at The Optima Corporation (2011a), including Paul Day, Jim Waite, and Oliver Weide for their ongoing development efforts that have been responsible for turning the original software developed by Shane Henderson and the author into the highly polished Predict software system now available. The information gained by Optima from their Predict implementations and relayed to the author has been invaluable in preparing this chapter. Furthermore, the author acknowledges the information made available by Optima from their experiences in developing and implementing the Optima Live repositioning software. Optima's ongoing support for the author's research and that of his students is also gratefully acknowledged.

Much of the work required to develop a simulation model of the Copenhagen ambulance operations was undertaken by Dr René Munk Jørgensen of Logis A/S, Denmark (Jørgensen 2011). The author is grateful for the information provided by Dr Jørgensen and Optima.

None of this work would have been possible without Shane Henderson's contribution of the original simulation engine. The author would like to acknowledge the continuing collaboration with Shane and his ongoing research contributions and invaluable discussions. Comments on this chapter from Shane, Armann Ingolfsson and this book's editor have significantly improved its content. The author also wishes to acknowledge the contributions of his colleagues and students, including Professor Andy Philpott, David Richards, Sarah Clark and Lei (Oddo) Zhang.

## Appendix: A Real-Time Multi-view Generalised-Cover Repositioning Model

Let $Z$ be the set of demand locations (termed *zones*), $V$ the set of all *views* (defined below), $A$ the set of all available (i.e. free) ambulances and $W$ be the set of all permitted waiting locations for the ambulances. The following formulation describes the real-time multi-view generalised-cover repositioning model (RtMvGcRM) used in Optima Live. Although not shown explicitly, these sets and all constants in this model are assumed to depend on the vehicle positions, call arrival rates and road speeds applying at the time $t$ the model is solved.

$$\text{RtMvGcRM}(t): \ \max \sum_{z\in Z}\sum_{v\in V} g_{zv}(y_{zv}) - \sum_{a\in A}\sum_{w\in W} c_{aw}x_{aw} \tag{11.5}$$

$$\text{s.t.} \quad \sum_{w\in W} x_{aw} = 1 \quad \forall\, a \in A \tag{11.6}$$

$$y_{zv} = \sum_{a\in A}\sum_{w\in W} d_{awzv}x_{aw} \quad \forall\, z \in Z, v \in V \tag{11.7}$$

$$x_{aw} \in \{0,1\} \; \forall\, a \in A, w \in W \tag{11.8}$$

In this model, decision variable $x_{aw} = 1$ if ambulance $a$ is to be assigned to waiting location $w$, and $x_{aw} = 0$ otherwise. The cost $c_{aw}$ takes into account factors such as a fixed and variable cost of moving an idle ambulance, the cost of redirecting an en-route vehicle, any maximum driving times and the distance of $w$ from the vehicle's original home base. This model is defined in terms of views, where a view $v$ is a combination of a user-specified subset $A_v \subset A$ of ambulances, a set of zone-specific target response times $r_{zv}\ \forall\, z \in Z$ and a dispatch priority (either lights-and-sirens or otherwise). For example, the user may be interested in the coverage provided by transport-capable vehicles dispatched without lights and sirens to meet a 20 min target in metropolitan areas and a 30 min target in rural areas. We define $d_{awzv}$ to be the coverage an ambulance $a$ provides to zone $z$ under view $v$ when assigned to waiting location $w$. Typically, we will have

$$d_{awzv} = \begin{cases} 1 & \text{if } a \in A_v \text{ and } t_{wzv} \leq r_{zv} \\ 0 & \text{otherwise} \end{cases}$$

where $t_{wzv}$ is the driving time under the view $v$'s dispatch priority when going from waiting location $w$ to zone $z$. However, more general options are possible which recognise the benefit of arriving earlier or later than the target time or having variability in the driving time.

Variable $y_{zv}$ defined by (11.7) measures the total coverage of zone $z$ under view $v$, and $g_{zv}(y_{zv})$ is the resulting contribution zone $z$ makes to the user's performance measure. Function $g_{zv}(\cdot)$ is a piecewise linear convex function which is typically proportional to the current call arrival rate in zone $z$ and the importance associated with view $v$. For example, the standard single-coverage measure can be implemented using

$$g_{zv}(y) = \begin{cases} w_z & \text{if } y \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

where $w_z$ is the call arrival rate in zone $z$. Other choices of $g_{zv}(\cdot)$ are possible, such as that used in Daskin's MEXCLP (Daskin 1983). Function $g_{zv}(\cdot)$ can be modelled within the integer-programming formulation using additional upper-bounded variables not shown above. The objective (11.5) seeks to maximise the weighted reward generated from these $g_{zv}(\cdot)$ values less the cost of the vehicle repositioning moves, where (11.6) ensures each free ambulance is assigned one waiting location.

## References

Alanis R, Ingolfsson A, Kolfal B (2010) A Markov chain model for an EMS system with repositioning. http://apps.business.ualberta.ca/aingolfsson/documents/PDF/Repositioning.pdf. Accessed 11 Aug 2011

Ambulance Victoria (2009) Strategic plan 2010–2012, draft for consultation. http://www.ambulance.vic.gov.au/Media/docs/28336_Dec09_v2_COLOUR-51a3ac1e-b8e5-447c-bbb5-cf42ee589ad9-0.pdf

Ambulance Victoria (2010a) 2009–2010 Annual Report. http://www.ambulance.vic.gov.au/Media/docs/Report%20o%20Operations-ebb69003-7c64-4090-a48e-07ec9057f368-0.pdf. Accessed 11 Aug 2011

Ambulance Victoria (2010b) 2010 Annual Report map. http://www.ambulance.vic.gov.au/Media/docs/2010%20Annual%20Report%20Map%20v4-0139d2b6-e637-4cec-b704-118e25c3f88d-0.pdf. Accessed 11 Aug 2011

Ambulance Victoria (2011) Types of paramedics. http://www.ambulance.vic.gov.au/Paramedics/Types-of-Paramedics.html. Accessed 11 Aug 2011

Andersson T, Värband P (2007) Decision support tools for ambulance dispatch and relocation. J Oper Res Soc 58:195–201

Andradóttir S (2006a) Metamodel-based simulation optimization. In: Henderson SG, Nelson BL (eds) Handbook in operations research and management science, vol 13, Chap 20. Elsevier, Amsterdam, pp 535–574

Andradóttir S (2006b) An overview of simulation optimization via random search. In: Henderson SG, Nelson BL (eds) Handbook in operations research and management science, vol 13, Chap 20. Elsevier, Amsterdam, pp 617–631

Berman O (1981a) Dynamic repositioning of indistinguishable service units on transportation networks. Transport Sci 15(2):115–136

Berman O (1981b) Repositioning of distinguishable urban service units on networks. Comput Oper Res 8(2):105–118

Berman O, Krass D (2002) Facility location problems with stochastic demands and congestion. In: Drezner Z, Hamacher HW (eds) Location analysis: applications and theory. Springer, New York

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. Eur J Oper Res 147:451–463

Budge S, Ingolfsson A, Erkut AE (2009) Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. Oper Res 57:251–255

Budge S, Ingolfsson A, Zerom D (2010) Empirical analysis of ambulance travel times: the case of Calgary emergency medical services. Manag Sci 56(4):716–723

Church RL, ReVelle CS (1974) The maximal covering location problem. Paper Reg Sci Assoc 32:101–118

Daskin MS (1983) A maximum expected coverage location model: Formulation, properties and heuristic solution. In Transportation Science. 48–70, 17.

Daskin MS (1987) Location, dispatching and routing models for emergency services with stochastic travel times. In: Ghosh A, Rushton G (eds) Spatial analysis and location-allocation models. Van Nostrand Reinhold Co., New York. Northwestern University, pp 224–265

Erkut E, Ingolfsson A, Erdogan G (2008) Ambulance deployment for maximum survival. Naval Res Logist 55:42–58

Fu MC (2006) Gradient estimation. In: Henderson SG, Nelson BL (eds) Handbook in operations research and management science, vol 13, Chap 19. Elsevier, Amsterdam, pp 535–574

Fu MC, Glover FW, April J (2005) Simulation optimization: a review, new developments, and applications. In: Kuhl ME, Steiger NM, Armstrong FB, Joines JA (eds) Proceedings of the 37th Winter Simulation Conference, Orlando, FL, USA, 4–7, 2005. ACM 2005, ISBN 0-7803-9519-0

Gendreau M, Laporte G, Semet F (1997) Solving an ambulance location model by tabu search. Location Sci 5(2):75–88

Gendreau M, Laporte G, Semet F (2001) A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. Parallel Comput 27:1641–1653

Gendreau M, Laporte G, Semet F (2006) The maximal expected coverage relocation problem for emergency vehicles. J Oper Res Soc 57:22–28

Goldberg JB (2004) Operations research models for the deployment of emergency services vehicles. EMS Manag J 1:20–39

Henderson SG (2010) Operations research tools for addressing current challenges in emergency medical services. In: Cochran JJ (ed) Wiley encyclopedia of operations research and management science. Wiley, New York

Henderson SG, Mason AJ (1999) Estimating ambulance requirements in Auckland. In: Proceedings of the winter simulation conference, Phoenix, Arizona, vol 2, pp 1670–1674

Henderson SG, Mason AJ (2004) Ambulance service planning: simulation and data visualization. In: Brandeau M, Sainfort F, Pierskalla W (eds) Operations research and health care: a handbook of methods and applications. International series in operations research & management science, vol 70, Chap 4. Kluwer, Dordecht, pp 77–102

Hogan K, ReVelle C (1986) Concepts and applications of backup coverage. Manag Sci 32: 1434–1444

Ingolfsson A, Budge S, Erkut E (2008) Optimal ambulance location with random delays and travel times. Health Care Manag Sci 11:262–274

Jørgensen RM (2011) Logis A/S. http://www.logis.dk/. Accessed 11 Aug 2011

Kolesar P, Walker WE (1974) An algorithm for the dynamic relocation of fire companies. Oper Res 22(2):249–274

Larson RC (1974) A hypercube queuing model for facility location and re-sub-areaing in urban emergency services. Comput Oper Res 1:67–95

Larson RC (1975) Approximating the performance of urban emergency service systems. Oper Res 23

Law AM, Kelton WD (1999) Simulation modeling and analysis, 3rd edn. McGraw-Hill, New York

Mason AJ (2011) Engineering Science ambulance research website. http://www.esc.auckland.ac.nz/mason/Ambulances/. Accessed 11 Aug 2011

Mason AJ (2006) Faster map matching for emergency vehicle trip analysis. In: Proceedings of the 41st annual conference of the Operational Research Society of New Zealand, vol 41. Operations Research Society of New Zealand, New Zealand, pp 19–28

Mason AJ, Henderson SG (2010) An optimisation approach for map matching using sparse ambulance GPS data. Proceedings of the 5th INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2010) D. Sundaramoorthi, M. Lavieri, H. Zhao, eds. published by INFORMS.

Maxwell MS, Henderson SG, Topaloglu H (2011a) Equivalence results for approximate dynamic programming and compliance table policies for ambulance redeployment. http://legacy.orie.cornell.edu/~shane/. Accessed 11 Aug 2011

Maxwell MS, Henderson SG, Topaloglu H (2011b) Tuning approximate dynamic programming policies for ambulance redeployment via direct search. http://legacy.orie.cornell.edu/~shane/. Accessed 11 Aug 2011

Maxwell MS, Restrepo M, Henderson SG, Topaloglu H (2010) Approximate dynamic programming for ambulance redeployment. INFORMS J Comput 22(2):266–281

Nair R, Miller-Hooks E (2009) Evaluation of relocation strategies for emergency medical service vehicles. Transport Res Record 2137:63–73

OptTek Systems Inc (2011) Optquest. http://www.opttek.com/. Accessed 11 Aug 2011.

Priority Dispatch Corporation (2005). ProQA User Guide http://www.prioritydispatch.net/. Accessed 11 Aug 2011

Python Software Foundation (2011) Python programming language - official website. http://www.python.org/. Accessed 11 Aug 2011

Rockwell A (2011) Arena simulation software. http://www.arenasimulation.com/. Accessed 11 Aug 2011

Rajagopalan HK, Saydam C, Xiao J (2008) A multiperiod expected covering location model for dynamic redeployment of ambulances. Comput Oper Res 35:814–826

Region Hovedstaden (2008) Fremtidens ambulancekørsel og sygetransport i Region Hovedstaden (Future ambulances and patient transport in the Capital Region). http://www.regionh.dk/NR/rdonlyres/623A992C-F102-4E80-971A-A7A0E2D41C0F/0/Fremtidensambulancek%C3%B8rselogsygetransport.pdf. Accessed 11 Aug 2011

Region Hovedstaden (2009a) Ambulanceudbud 2009 (Ambulance Contract 2009). http://www.regionh.dk/menu/sundhedOghospitaler/Til+fagfolk/Akut+Medicin+og+Sundhedsberedskab/Praehospital/Praehospital+arkiv/Ambulanceudbud+2009.htm. Accessed 11 Aug 2011

Region Hovedstaden (2009b) En samlet orientering om udbudsforløbet vedrørende ambulancekørsel og sygetransport i Region Hovedstaden (A comprehensive briefing on the tender process for ambulances and patient transport, Capital Region). http://www.regionh.dk/NR/rdonlyres/E3776015-C235-465B-ADA5-C7A18778912B/0/Notatsamletorienteringomudbudsforl. Accessed 11 Aug 2011

Region Hovedstaden (2011) Responstider for ambulancer (Response Times for Ambulances). http://www.regionh.dk/menu/sundhedOghospitaler/Akut+hjaelp/Ambulancer+og+Akutl. Accessed 11 Aug 2011

Repede J, Bernardo J (1994) Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. Eur J Oper Res 75:567–581

Restrepo M, Henderson SG, Topaloglu H (2009) Erlang loss models for the static deployment of ambulances. Health Care Manag Sci 12:67–79

Schilling DA, Elzinga DJ, Cohon J, Church RL, ReVelle CS (1979) The TEAM/FLEET models for simultaneous facility and equipment siting. Transport Sci 13:163–175

Schmid V, Doerner KF (2010) Ambulance location and relocation problems with time-dependent travel times. Eur J Oper Res 207(3):1293–1303

Silva P, Pinto LR (2010) Emergency medical systems analysis by simulation and optimization. In: Johansson B, Jain S, Montoya-Torres J, Hugan J, Yücesan E (eds), IEEE available online at http://www.informs-sim.org/wsc10papers/222.pdf Proceedings of the 2010 Winter Simulation Conference

Stout JL (1983) System status management: the strategy of ambulance placement. System status management: the strategy of ambulance placement. J Emerg Med Serv 8:22–32

The Capital Region of Denmark (2009) Facts about the capital region of Denmark. http://www.regionh.dk/NR/rdonlyres/3184FA95-9C9D-4EBC-803B-8A068676EAC5/0/Faktapjece_eng_web.pdf. Accessed 11 Aug 2011

The Optima Corporation (2010) Optima case studies: Lee County, Florida. http://www.theoptimacorporation.com/case-studies/emergency-medical-services-provider-lee-county-avoids-more-than-usd-750000-in-unnecessary-costs-sees-faster-response-with-optima-livetm-and-optima-predicttm_10. Accessed 11 Aug 2011

The Optima Corporation (2011a) http://www.TheOptimaCorporation.com. Accessed 11 Aug 2011

The Optima Corporation (2011b) Optima case studies: capital region of Denmark. http://www.theoptimacorporation.com/case-studies/capital-region-of-denmark-relies-on-optima-predict-to-develop-accurate-realworld-simulation-of-ambulance-services-for-precision-planning_12

Toregas CR, Swain R, ReVelle CS, Bergman L (1971) The location of emergency service facilities. Oper Res 19:1363–1373

Wikipedia (2011) Ambulance Victoria. http://en.wikipedia.org/wiki/Ambulance_Victoria. Accessed 11 Aug 2011

Wright PD, Liberatore MJ, Nydick RL (2006) A survey of operations research models and applications in homeland security. Interfaces 36:514–529

Zhang L, Mason AJ, Philpott AB (2009) Optimization of a single ambulance move up. In: Proceedings of the 44th annual conference of the Operational Research Society of New Zealand. Operations Research Society of New Zealand, New Zealand, pp 225–226

Zhang L, Mason AJ, Philpott AB (2012) Optimising single-ambulance move-up. Forthcoming

# Chapter 12
# Planning and Managing Mass Prophylaxis Clinic Operations

**Rachel L. Abbey, Katherine A. Aaby, and Jeffrey W. Herrmann**

## 1 Introduction

Public health can be defined as "what we as a society do collectively to assure the conditions in which people can be healthy" [Institute of Medicine (IOM) 2003, p. xi]. Public health focuses not on the individual but on the population as a whole, setting it apart from the traditional healthcare system. The definition of a public health system is even more complex. It is a system that includes a network of individuals and organizations working together to create the conditions for health. It is the working together that enables them to act as a system (IOM 2003). This network can include government agencies, community organizations, healthcare providers, schools, businesses, and the media. Each network member, including local health departments, has a role to promote public health that may vary by community [National Association of County and City Health Officials (NACCHO) 2008].

An effective public health system includes the following core functions: preventing diseases, protecting against environmental hazards, preventing injuries, promoting and encouraging healthy behaviors, responding to disasters and assisting communities in recovery, and ensuring the quality and accessibility of health services (NACCHO 2003). Local health departments, the preferred term used by the NACCHO, are often the first line of defense within the public health system to support these core functions for many communities. Local health departments include all jurisdictional types and sizes. Throughout this chapter, the term "state

R.L. Abbey • K.A. Aaby
Montgomery County, Maryland Department of Health and Human Services, 2000 Dennis Avenue, Silver Spring, MD 20902, USA
e-mail: Rachel.Abbey@montgomerycountymd.gov; Kay.Aaby@montgomerycountymd.gov

J.W. Herrmann (✉)
A. James Clark School of Engineering, University of Maryland, College Park, MD 20742, USA
e-mail: jwh2@umd.edu

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_12, © Springer Science+Business Media New York 2013

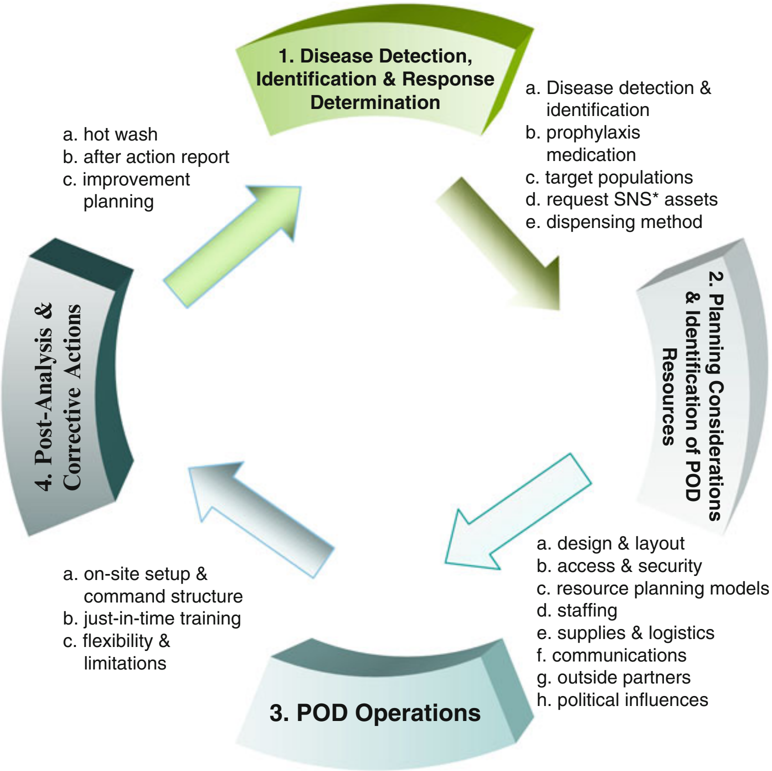

1. Disease Detection, Identification & Response Determination
a. Disease detection & identification
b. prophylaxis medication
c. target populations
d. request SNS* assets
e. dispensing method
2. Planning Considerations & Identification of POD Resources
a. design & layout
b. access & security
c. resource planning models
d. staffing
e. supplies & logistics
f. communications
g. outside partners
h. political influences
3. POD Operations
a. on-site setup & command structure
b. just-in-time training
c. flexibility & limitations
4. Post-Analysis & Corrective Actions
a. hot wash
b. after action report
c. improvement planning

**Fig. 12.1** Point of dispensing (POD) Operations Cycle. This figure outlines the four steps, along with specific processes, that public health planners work through when creating an effective efficient POD operation for a public health emergency or other public health activity. Step 1 begins at the *top* of the figure and continues clockwise through the specific processes (a–e) moving similarly through steps 2–4. (*Asterisk*) Strategic national stockpile

and local health departments or governments" is used to be inclusive of tribal health departments and governments.

This chapter will discuss the steps local health departments take to dispense medication to the public using PODs (points of dispensing) during a public health emergency or every day event, such as annual flu vaccination. The authors provide the POD Operations Cycle in Fig. 12.1 as a way to visually demonstrate the steps necessary to ensure an effective POD response. This chapter is intended to provide public health professionals an overview of the processes involved in planning and managing a POD operation from the perspective of a local public health department and local public health emergency planners (hereinafter known as "planners").

The chapter discussion applies to preparing for and responding to several public health emergency scenarios, including pandemic influenza, an aerosolized anthrax attack, a smallpox outbreak, as well as other smaller scale public health operations including annual influenza PODs. Although this chapter is written from a United

States (US) perspective, many of the concepts may apply to other countries. The authors recognize that PODs for public health emergencies and PODs for daily activities are different in scalability and urgency; however, some of the same steps can be applied to either situation. Throughout the chapter, the authors provide examples of the 2009–2010 H1N1 influenza pandemic as well as annual flu vaccine PODs to attempt to compare and contrast the two situations and the use of PODs. For more information on emergency preparedness and public health, visit the following web sites: Centers for Disease Control and Prevention (CDC) www.bt.cdc.gov and the Federal Emergency Management Agency (FEMA) www.fema.gov/areyouready.

The remainder of this chapter is organized as follows. Section 2 provides an overview of emergency preparedness and public health. Section 3 specifically addresses PODs and their role in emergency preparedness. Section 4 outlines and describes the specific steps and activities, shown in the POD Operations Cycle (Fig. 12.1), that local planners perform in POD planning and response. Section 5 discussed future research and practice challenges, including new and innovative solutions to increase the efficiency of PODs.

## 2 Background on Emergency Preparedness

As part of the public health system, local health departments, in collaboration with our state, tribal, and federal partners, have been charged with preparing for, responding to, and recovering from threats to public health. These threats can include acts of biological and chemical terrorism such as the dissemination of aerosolized anthrax spores or food product contamination and naturally occurring infectious disease threats such as pandemic influenza. Although predicting when and how such an event might occur is difficult, public health departments cannot ignore the possibility of events like the terrorist attacks on September 11, 2001, the anthrax attacks in 2001, the 2003 outbreak of severe acute respiratory syndrome (SARS), the 2004 earthquake and tsunami in Indonesia, Hurricane Katrina in 2005, the 2009–2010 novel H1N1 influenza pandemic, the 2010 Haitian earthquake, and the 2011 cascading disaster of an earthquake, tsunami, and nuclear event in Japan.

Preparing a nation to address these types of public health threats is a formidable challenge, but the consequences of being unprepared can be devastating. Today, with the increase in globalization, addressing health issues worldwide is critical. The ease and speed of transporting goods, services, and people across borders allows a small town or village to spread disease to the largest city in record time. The public health infrastructure must be prepared to prevent illness and injury that would result from a chemical, biological, radiological, and nuclear incident. As with emerging infectious diseases, the early detection and control of biological and chemical attacks depends upon a strong and flexible public health system at the local, state, federal, and international levels. In addition, primary healthcare providers must be vigilant to report and observe unusual illnesses or injuries.

In the USA, specific local state and federal laws and regulations provide guidance and authority to local health departments during an emergency. State laws grant powers to local governments in the form of police powers to quarantine, investigate disease outbreaks, and regulate facilities (Vinter et al. 2010). The federal Homeland Security Presidential Directive HSPD-21 names disease surveillance, caring for the sick and deceased, medical and nonmedical prevention strategies, and community resiliency critical to biodefense. The 2006 Pandemic All-Hazards Preparedness Act contains many functions of local health departments and emphasizes that local health departments and other medical first responders are critical to the response.

Local health departments undertake many roles and responsibilities during a public health emergency and are guided, in collaboration with many community and government partners, by specific federal frameworks to assure an organized and effective response. The National Response Framework (NRF) provides guiding principles for governments and their partners to provide a unified response to disasters and emergencies. Part of the NRF addresses the 15 emergency support functions (ESF) under which many local, state, and federal governments organize their resources and capabilities. ESF #8 is Public Health and Medical Services, which covers most of the functions of federal, state, and local health departments as well as other healthcare partners [U.S. Department of Homeland Security (DHS), FEMA 2011b]. Some of the core actions include assessment of public health/medical needs, health surveillance, managing mass fatalities, providing public health and medical information, managing behavioral healthcare, and protecting against environmental hazards (U.S. Department of Homeland Security 2008). In an effort to create a united response among agencies and organizations, local health departments follow the National Incident Management System (NIMS). NIMS can be applied across a full spectrum of potential all hazards regardless of size, location, and capacity of the incident (U.S. DHS, FEMA 2011a). For more information on NIMS, please visit www.fema.gov/emergency/nims/.

Federal program guidances such as the 2009 National Health Security Strategy, the 2010 Biennial Implementation Plan, and the 2011 Public Health Emergency Preparedness (PHEP) capabilities provide measures, benchmarks, and funding for local health departments to improve their PHEP capacity. In addition, CDC and state representatives conduct annual technical assistance reviews (TAR) to assess plans and to ensure readiness; these reviews use a 0–100 score system (U.S. CDC, October 14, 2011). However, since PHEP is relatively new to the field of public health, there is limited research and data regarding performance and program measures. Over the next few years, especially with the release of the new PHEP capabilities, additional data will be collected and analyzed to help inform future PHEP policies.

## 3 Purpose of PODs in Emergency Preparedness

If required, PODs, or points of dispensing, are a key function of local health departments in their mass prophylaxis and vaccination response to a public health and medical disaster or emergency. The goal of PODs is to medicate the population as quickly and accurately as possible to prevent morbidity and mortality. In emergency preparedness, PODs, or mass prophylaxis clinics, are the primary strategy used by local health departments to distribute medical countermeasures, such as medications (vaccine or antibiotics), to the public. PODs are also used in the everyday function of local health departments such as annual flu or routine immunization clinics. They vary in size, number, and location depending upon the jurisdiction, disease, and other factors. The annual TAR guidance requests localities to collect some baseline data which includes hourly estimated throughput, numbers of PODs, types of PODs, and levels of staffing. PODs are aimed at prophylaxis (prevention of illness), rather than at treatment (medical efforts to treat symptomatic individuals) (IOM 2008).

There are two primary POD models used by local health departments. The first is a centralized POD design in which the local health department and their partners set up several locations across a jurisdiction and ask the public to come to these PODs to pick up their medications; this is referred to as the "pull method" (Ablah et al. 2010). These could include traditional "walk-up" clinics where persons arrive one by one at multiple venues or "drive-through" clinics where cars drive up to receive medication. These clinics were popular during the H1N1 pandemic and often are used for annual flu or traditional vaccination clinics. The second design is a decentralized POD model, where the local health department and its partners would deliver directly to the public; this concept is referred to as the "push model" (Ablah et al. 2010). For example, if a target population were school children, as during annual flu season, then PODs in the schools might be the fastest method to dispense the vaccine. Local health departments may choose to use either one of these models or a combination.

In addition to using different POD models, local health departments may staff PODs using medical or nonmedical staff, or a combination. Depending on the type and scale of the incident or event, it may be necessary to use primarily nonmedical staff. For example, if the event or incident is small and manageable, primarily medical staff can provide accurate screening, triage, and exams, and could answer medical questions. If an incident is so large that it exceeds the local health department's medical staff capacity, nonmedical personnel may be used to supplement medical staff. The nonmedical workers would be able to dispense medications and triage as appropriate but would not be able to provide medical consultation or assessments (IOM 2008).

## 4 The POD Operations Cycle

The POD Operations Cycle, illustrated in Fig. 12.1, demonstrates the processes necessary for an effective and efficient POD response. The first step of the cycle, which begins at the top of Fig. 12.1, is disease detection, situational awareness, and response determination. This includes an assessment of surveillance data and additional information, which will help to determine the affected population and a decision about the scope of the POD operation. The second step involves specific planning considerations that are necessary to clinic operations as well as the identification of resources necessary for the clinic. The third step is the setup and operation of the PODs. The fourth (and last) step is post-operation analysis and corrective actions, which involve capturing and analyzing the lessons learned during the clinics' operations in order to improve future operations.

## *4.1 Step 1: Disease Detection, Identification, and Response Determination*

The first step of the POD Operations Cycle allows clinic planners to create a picture at a specific time and place to determine the type of disease (or biological agent) and to make critical decisions. The initial detection will likely occur at the local level. The detection and identification of the disease will determine the type of response which includes the type of prophylaxis medication necessary, identification of the target populations, the request for medical supplies, and the type of dispensing method(s). The order of when these processes occur may vary, and some may occur simultaneously, but the first step in the POD Operations Cycle always includes an assessment of the situation at hand and what immediate decisions need to be made in order to provide prophylaxis to the affected populations.

### 4.1.1 Disease Detection and Identification

Epidemiological data and information about the disease are critical when making an assessment of any outbreak, pandemic, or other biological event. Epidemiology is the study of the determinants and distribution of disease in the human population (Morton et al. 2001). Planners use epidemiological data to determine the type and size of the affected population to whom they will provide prophylaxis. During an emergency, understanding the epidemiological determinants and distribution of disease is important, and time sensitive, in order to save as many lives as possible. The populations that are of concern may be different and may change during the course of the event as more information develops about the disease or agent. In order to make the best-informed decisions about the disease and the affected populations, planners use many tools to gather and analyze information.

Disease surveillance systems at state and local health agencies must be in place and capable of detecting unusual patterns of disease or injury, including those caused by unusual or unknown biological agents. In the USA there are a number of disease surveillance systems used by states and localities to collect, report, and track reportable disease data. It goes beyond the scope of this chapter to discuss the different disease surveillance systems; however, they can differ from locality to locality, are often disease specific, and are fragmented (IOM 2003). This often makes it difficult to received timely and accurate data, particularly as it relates to a public health emergency.

Syndromic surveillance systems are of particular interest in the field of public health emergency response. They use health-related data such as symptoms which are often precursors to an actual diagnosis and may serve as a warning system. One particular system, Electronic Surveillance System for Early Notification of Community-based Epidemics (ESSENCE IV), which was designed by the Johns Hopkins University Applied Physics Laboratory and the US Department of Defense, is used by many local and state health departments. In practice, ESSENCE is an ongoing systematic collection of indicators of health status, grouped into health syndromes among a patient population. The application collects information from hospitals, healthcare providers, over-the-counter pharmaceutical sales, and school-based absenteeism reports. It applies statistical algorithms to detect unexpected changes in the data and provides the information to health officials in a web-based application [U.S. Department of Health and Human Services (DHHS), CDC 2011]. In particular, both exponentially weighted moving average (EWMA) models and autoregressive moving average (ARMA) models are used in ESSENCE.

The goals of ESSENCE are early detection of large-scale outbreaks, enhancement of traditional notifiable disease surveillance systems, monitoring the progress of recognized outbreaks, and ruling out existence of an emergency. The benefit of ESSENCE is that early detection accelerates response time; early response time reduces transmission, and reduced transmission limits incidence and mortality. During the H1N1 pandemic, ESSENCE provided planners a system to monitor the outbreak and information about the incident, including demographic characteristics of those who became ill and died from the disease.

In addition to syndromic surveillance, an additional system that planners use to identify an agent or disease is public health laboratories. The laboratory response network has been established to assist in a response to infectious diseases and bioterrorism. According to the association of public health laboratories, public health laboratories provide clinical diagnostic testing, disease surveillance, environmental and radiological testing, emergency response support, applied research, laboratory training, and other essential services to the community. There are central public health laboratories in every state, and the District of Columbia and many states have local public health laboratories as well that range in size and capacity.

Another system planners use in the USA is the Health Alert Network (HAN). HAN is a nationwide, integrated information and communication system whose goal is to strengthen state and local preparedness by serving as a platform for the distribution of health alerts, dissemination of prevention guidelines and other information, distance learning, national disease surveillance, and electronic laboratory reporting (U.S. DHHS, CDC 2001). In addition to HAN, many local health departments depend upon additional local and state systems to send and receive critical health information.

The US Department of Homeland Security's (DHS) BioWatch Program is also used by planners to help identify what is going on in the community. In 2003, the newly created DHS introduced the BioWatch Program. The program's objective is to swiftly detect specific biological agents that could be released in aerosolized form during a biological attack. BioWatch and infectious disease surveillance through the public health and healthcare systems are complementary. However, BioWatch has the potential to provide a timelier alert than the public health systems due to the quick turnaround time in reporting results. Also, the testing is focused on only select biological agents, unlike the wide range of infectious agents tested in the public health systems (Shea and Lister 2003).

Lastly the media (e.g., television, Internet, social media, and radio) serves as a tool to distribute timely information and should be monitored closely for analysis of any event. In addition, federal partners are also part of the information collection process. DHS and the U.S. DHHS are in constant communication during an incident of national importance. As one of the offices within the U.S. DHHS, the CDC plays an important role in communicating to state and local health departments the most current epidemiological data on the disease as well as providing guidance for prophylaxis medications. The CDC was critical in providing information on vaccine priority groups during the H1N1 pandemic.

In conclusion it is necessary for planners to collect and analyze data and information from many different sources in order to determine the type of event, the disease, and the affected population. This information will then be used by medical experts to determine the type of prophylaxis medication dispensed.

### 4.1.2 Prophylaxis Medication

As part of the assessment process, after determining the agent or disease, planners must determine the type of prophylaxis to dispense. Prophylaxis can be defined as "the prevention of or protection from disease." In public health, types of devices (e.g., condom), treatments, and medications can be referred to being prophylactic (Agnes 2005, p. 1150). Different types of diseases will require different types of prophylaxis, including vaccination and/or distribution of antibiotics. In addition to the type of prophylaxis, the route (e.g., oral, injection, and patch) and the number of doses are also important, particularly when planning for a mass prophylaxis operation.

The type, route, and dose of prophylaxis for even a common disease, like influenza, can vary. The primary prophylactic treatment for influenza is a vaccine (injection); however, since 2003 an influenza vaccine nasal spray has been available for those three years to 49 years of age (Fiore et al. 2010). In some cases, as during the early stages of the H1N1 pandemic, two doses of the influenza vaccine were initially recommended for individuals for full immunity (U.S. DHHS, CDC 2009); however, this was later reversed to only apply to children under age 10 (U.S. DHHS, CDC 2009). Another example is aerosolized anthrax, oral antibiotics, and/or a vaccine (injection) may be dispensed. Both methods include multiple doses to treat the affected population. It is important for planners to recognize that the type, route, and dose of prophylaxis, which is determined by scientific evidence from CDC, will affect the overall design, staffing, processing time, and throughput of the POD.

### 4.1.3 Target Populations

Planners must determine the size of the affected target population and identify subpopulations or priority groups that may be especially vulnerable or need special assistance. Local health departments often follow disease guidances developed by state and federal officials. The decisions made regarding the target populations, as well as any subpopulations or priority groups, are made based upon epidemiologic data of the disease as well as information collected from the various sources mentioned in Sect. 4.11.

An example scenario is influenza. The CDC's Advisory Committee on Immunization Practice (ACIP) meets regularly to update their annual flu guidance, which is based upon epidemiologic and clinical data and input from the general public (Fiore et al. 2009). This guidance is used to inform local and state public health officials as to the target population and priority groups for annual flu. It is important that there be a consistent message to the public about who is included in the priority groups and the reasons behind this decision. For routine annual influenza, the target populations have remained fairly constant. However, for the 2010–2011 influenza season, ACIP updated its guidance to include all persons aged 6 months and older. The previous guidance recommended annual vaccinations of adults aged 19–49 years, but the new recommendations were supported by evidence that annual flu vaccination is a safe and effective preventive strategy that could benefit all age groups (Fiore et al. 2010). The target population for annual flu is extremely large, and with enough vaccine to meet the demand, subpopulations or priority groups are not necessary.

During the H1N1 pandemic, the primary priority groups determined by ACIP were different than annual influenza. They included pregnant women, people who live with or provide care for infants, healthcare and emergency medical services personnel, people aged six months to 24 years, and those aged 25–64 who have medical conditions (Fiore et al. 2009). The epidemiological data showed that those groups were most at risk for the H1N1 virus. With limited vaccine at the time, these were the priority populations that local health departments were to target to limit morbidity and mortality.

There are challenges associated with outreach to target populations. Many populations may be resistant to receiving a vaccine. For example, despite the fact that Hispanics and African-Americans are more at risk for chronic illnesses (e.g., asthma, diabetes), influenza vaccination rates are still lower in these groups [Trust for America's Health (TFAH) 2010; U.S. DHHS, CDC, National Center for Chronic Disease and Health Promotion 2010]. Data from focus groups on the H1N1 pandemic found that some target populations felt that immunization or vaccine were not viewed as an important health issue, that the media exaggerated H1N1 as a pandemic, and that there was inadequate information on participants of clinical trials (Gist 2011). It is important for planners to be aware of cultural barriers, including concerns about receiving medication, and additional strategies may be required, i.e., public messages may need to target specific populations to inform them about the safety and importance of vaccination.

### 4.1.4 Request Strategic National Stockpile Assets

Once it is recognized that a significant public health incident has occurred, or is occurring, the need for supplies and resources, specifically medical, is urgent. The Strategic National Stockpile (SNS) is a national repository of medicine and other medical supplies used to supplement and resupply state and local resources. The decision to request and deploy SNS assets will be a collaborative decision made by local, state, and federal officials. The decision will begin most likely at the local level when officials first identify a potential or actual situation that they believe has the potential to threaten the health or safety of their community. The SNS supplies will be used as a supplement to state, local, and personal stockpiles (IOM 2011). The task associated with the delivery of federal SNS assets from their original warehouse location to the receiving, staging, and storing (RSS) warehouse sites and then to the dispensing sites is referred to as distribution (IOM 2008).

SNS stockpiles are strategically located in secure warehouses throughout the USA to ensure that once federal and local authorities agree that SNS deployment is needed, "12-hour push packs" of medications and/or supplies can be delivered to any designated RSS site within 12 h, while other managed inventory can be in place within 24 h of the decision to deploy. Once the SNS supplies arrive at the designated site, state and local authorities assume responsibility for the supplies and equipment and oversee storage, distribution, and dispensing (U.S. DHHS, CDC 2011).

### 4.1.5 Dispensing Method

After the target populations are identified and the request for supplies (either through the SNS or local or state stockpiles) and assistance has been sent, planners must select one or more methods for dispensing prophylactic medication. Dispensing involves providing prophylactic medication to the affected population in response to an incident or threat (IOM 2008). Ideally, those in the target population could obtain the prophylaxis from a web of entry points such as primary care providers, PODs, pharmacies, and other private healthcare providers. In some localities, prepositioning medications, or storing medications close to or in possession of those who need rapid access, is also being explored as another strategy in addition to PODs to rapidly distribute and dispense medication (IOM 2011). This variety of sources makes the prophylactic medication more accessible, thereby increasing the likelihood that people will decide to receive it to protect themselves.

Most local health departments have a mass prophylaxis plan that includes PODs. In an event requiring mass prophylaxis, the target population will be instructed to receive prophylactic medication using either one model or a combination of the POD models previously discussed in Sect. 3 of this chapter. The appropriate mass prophylaxis response for some disease agents, such as smallpox, may be accomplished over several days, because of the long incubation period. However, others, especially aerosolized anthrax, will require the target population to receive the first prophylactic medication within 48 h or less to minimize morbidity and mortality (Heymann 2004). The type of POD(s) and length of operation will vary depending upon the location site, the disease agent, and the size of the target population. Therefore, it is all the more important for planners to develop a scalable and flexible mass dispensing plan.

## 4.2 Step 2: Planning Considerations and Identification of POD Resources

After determining the type of disease and medications, requesting the SNS, targeting populations, and deciding on the dispensing method, planners move to the second step of the POD Operations Cycle. Planning considerations are made based upon information collected in the first step and include design and layout of PODs and access and security. Resource planning models can then assist to identify some of the resources necessary such as staffing and associated costs. Additional resources necessary for PODs include supplies and logistics, communications, outside partners, and political influences. Many of these planning considerations and identification of resources can be done pre-event, while others may occur during real time after the incident begins.

Throughout step 2, it is necessary to assess funding at each stage of the process. For example, during an annual influenza vaccination campaign, funding may be limited. To have the largest community impact, planners may decide to use limited resources for PODs at elementary schools to target young children, an at-risk population that can spread disease quickly within the community, along with a few selected small public PODs targeting other at-risk populations. In the case of a declared public health emergency, the Robert T. Stafford Disaster and Emergency Assistance Act states that incurred expenses may be reimbursed by the federal government; this may allow for more flexibility in providing increased outreach to the target population (U.S. DHS, FEMA 2007).

### 4.2.1 Design and Layout

The design and layout of the POD operation is critical. Design refers to conceptualizing the plan and purpose of the POD. Layout refers to the actual laying out (i.e., on paper or in a computer program) the design or plan to move people through the POD. See Figs. 12.2 and 12.3 as sample POD layouts. Ideally, much of the design and layout of the clinics is performed pre-event. However, in some cases, they may need to be created at the time of an event, if facilities are being occupied, renovated, or otherwise in use. The processes would be the same, but time will be a factor during real time.

There are some design elements that need to be considered. The size and dimensions of the facility must be well suited to handle large crowds and long lines

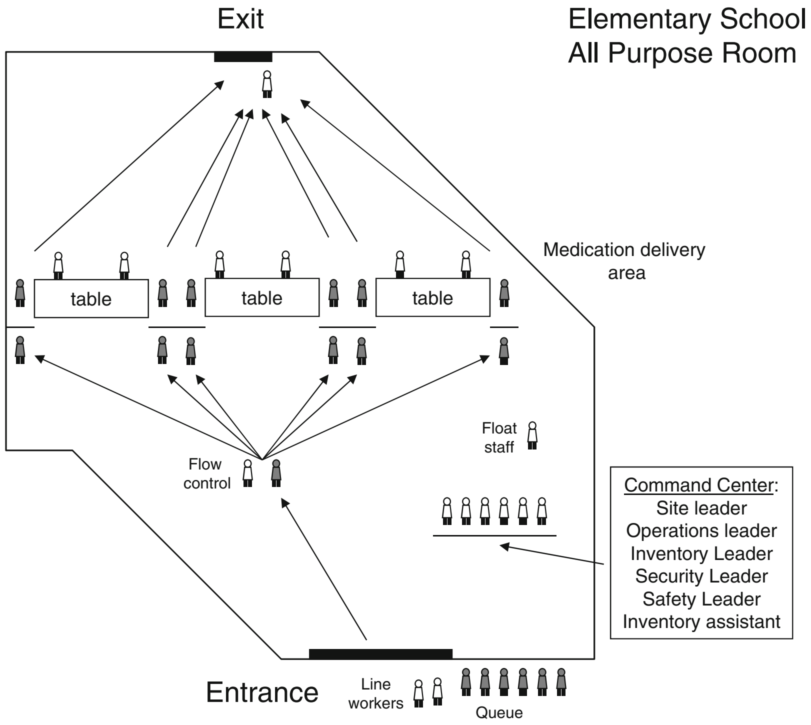


**Fig. 12.2** Point of dispensing (POD) layout A. This is an example layout of a POD using an elementary school all-purpose room, with associated staff positions. The flow of the POD begins and the entrance and follows the *arrows* to the exit. This figure was created using Microsoft® PowerPoint®

of waiting people (including those with special needs, e.g., wheelchair bound and blind), preferably under cover and sheltered from the weather. Large areas inside the facility where people can receive handouts (e.g., disease and medication fact sheets and other information) will be needed. People receiving medications should enter from one area and exit from another without having to backtrack past others who are still waiting in line. Specific stations that clients visit during the process of receiving the medication will need to be designed. These stations can include greeting, triage, registration and forms completion, medical screening, education, and dispensing of medication. Some PODs have many stations, while others will have as few as two. The number of stations and staff needed may depend upon the amount of data collection or additional patient education required by local, state, and federal governments at the POD. This may not be determined until the event.

The time to serve the target populations must also be considered. When time is of the essence (i.e., it is critical to provide a target population with medication as quickly as possible due to the epidemiological characteristics of the disease), it may be better to have as few stations as possible in order to reduce the time in the

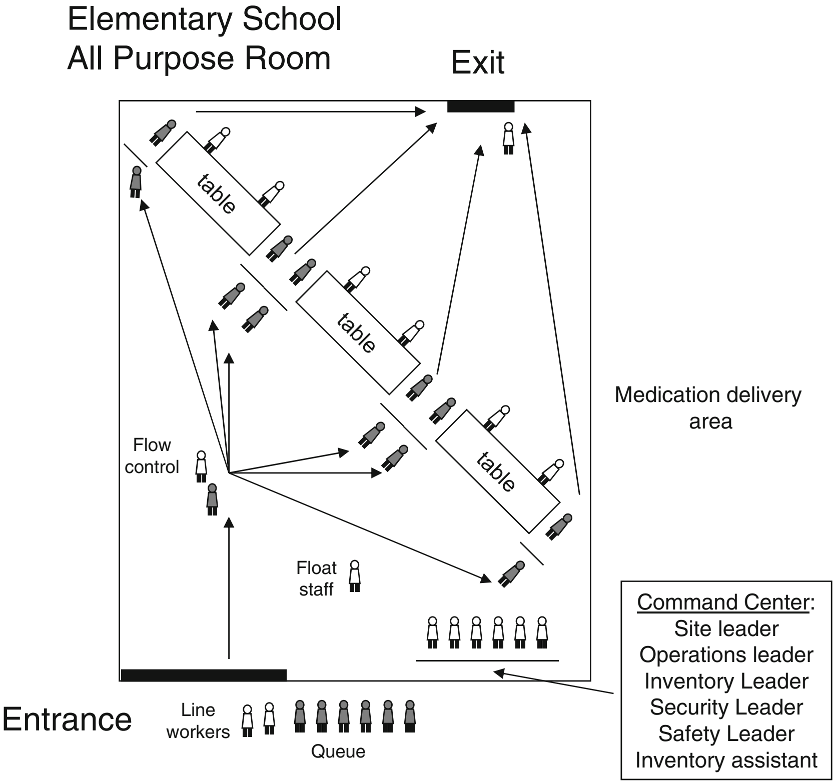


**Fig. 12.3** Point of dispensing (POD) layout B. This is an example layout of a POD using an elementary school all-purpose room, with associated staff positions. The flow of the POD begins and the entrance and follows the *arrows* to the exit. This figure was created using Microsoft® PowerPoint®

POD. For an annual flu vaccination POD, it is necessary to collect paperwork and documentation on the client or resident receiving the medication; additional stations are necessary to process this paperwork, but time is not as critical as during an emergency.

Most local mass prophylaxis plans include a general POD layout which may need to be adjusted depending upon the event. This layout is intended to be flexible and may change, expand, or contract, depending upon the characteristics of the mass prophylaxis campaign (considered in the first step). This is particularly evident in the difference between annual flu vaccination POD operations and the H1N1 pandemic flu vaccination PODs, which was treated as a public health emergency. Since the volume of patients exceeded annual flu vaccination PODs, in some cases, layouts had to be reevaluated in order to accommodate additional staff and space for screening patients.

Planners should perform a site assessment, pre-event if possible, to determine each site's capacity to meet the design needs of the POD. Once the POD sites have been selected, a site-specific plan must be developed for each site. The following additional information should be collected:

- Equipment and furniture
- Staff accommodations
- Accommodations for people with disabilities or special needs
- Location of the EMS/ambulance/first aid staging area
- Number of restrooms
- Location of the medication receiving area (if necessary)
- Environmental health concerns
- Security arrangements
- The number and location of entrances and exits
- Space available for an incident command post
- Number of parking spots
- Potential traffic problems
- 24/7 contact information for the facility manager and procedures for accessing the site
- Other basic needs of the public (Phillips and Williamson 2005)

During the H1N1 pandemic, the demand for the vaccine was initially unknown. In the beginning the demand in some areas was so high that portable toilets had to be ordered and delivered in order to accommodate the people waiting in line outside the facility. An actual layout of the floor plan will need to be designed and can be created easily using paper, pencil, and measuring tape once a location for the POD operations is established.

### 4.2.2 Access and Security

Accessibility, time, and security are essential factors that affect the design of POD operations. Many of these decisions can be made pre-event, but again some may need to be changed or modified real time.

Flu vaccination strategies that create better access for the target population can increase vaccination rates (TFAH 2010). Issues such as providing adequate traffic flow and accommodating large numbers of people often lead local health departments to partner with local school systems to use their facilities. If the schools are in session, planners may need to consider nontraditional community venues such as community centers, convention centers, sports arenas, recreation centers, libraries, armories, churches, and private businesses. During the H1N1 pandemic, PODs were opened in schools, hospitals and large provider groups, public health departments, pharmacies, and occupational and institutional PODs (Rambhia et al. 2010). It is important to use existing places of congregation in the community that people already know and can easily access.

In order to maximize accessibility, it is important to have PODs at venues and times that best reach the target population. For example, if the target population is seniors, it may be advantageous to open PODs at senior centers or other gathering venues for seniors in the community. Working adults may prefer evening hours and drive-through PODs that they can visit on the way home from work.

Security for POD sites is a local law enforcement responsibility. It is critical for mass dispensing during a public health emergency, especially in controlling large crowds and/or in dealing with a shortage of the medications. Local planners must coordinate with local law enforcement to ensure that thorough security assessment is conducted by law enforcement on each designated POD site and that a security plan is written by law enforcement.

### 4.2.3 Resource Planning Models

After determining the scope and strategy of the response, planners must estimate the resources required for executing the response. Because of the scale of mass prophylaxis operations, tools such as planning models can help planners create estimates of resources, especially staffing requirements, and generate and evaluate plans for the logistics of distributing supplies. Many of the models discussed here can be used for pre-event as well as real-time planning and as an evaluation tool for POD plans. It is always important to keep in mind that models provide predictions (not guarantees), and their accuracy is limited by the quality and uncertainty of the information used to build the model. Also, in general, a key tradeoff in modeling is that models that are more accurate (i.e., they capture more details of the situation) usually require more information and more time to build and run.

Planning models are operations research models that are implemented as web sites, spreadsheets, and computer software. These models require inputs about the scope of the mass prophylaxis operation and the steps that will be performed. Many of these inputs are gathered from mass prophylaxis and POD plans or previous POD data collected. From this information, planning models estimate the staff required to operate the PODs. Some models can also estimate the congestion, which affects how long people will wait in line and how much space should be allocated for lines.

The Weill/Cornell Bioterrorism and Epidemic Outbreak Response Model (BERM) was the first widely available mass prophylaxis planning tool that performs capacity analysis to estimate staffing requirements (Hupert and Cuomo 2003). The model has been implemented as a spreadsheet and as an interactive web site. Version 2.0 is available online (http://www.ahrq.gov/research/biomodel.htm). This model simplifies the modeling process by limiting the analysis to a small number of options, which reduces the time required to build a model.

RealOPT (Lee et al. 2009) is another popular POD planning model. The RealOPT suite of models uses a combination of simulation and optimization to solve a variety of planning problems, including selecting locations for PODs, estimating staffing needs, and allocating staff to stations. RealOPT is a software application available upon request from the developers, who are based at Georgia

Tech (http://www2.isye.gatech.edu/~evakylee/medicalor/). The use of simulation and optimization requires additional computational resources but allows the tool to model many details and find automatically good staffing plans.

The Clinic Planning Model Generator (CPMG) (http://www.isr.umd.edu/Labs/CIM/projects/clinic/) is a spreadsheet-based tool that generates, for a specific POD configuration, a customized capacity-planning and queuing model spreadsheet (Aaby et al. 2006a, b; Herrmann 2008). The model allows planners to enter known population information and set time constraints specific to their applications. The immediate results include the minimum staff levels required, along with detailed POD information regarding waiting times, queue lengths, and cycle time. This approach can model a wide range of POD designs and avoids lengthy simulation runs by using queuing network approximations to estimate how long persons will wait.

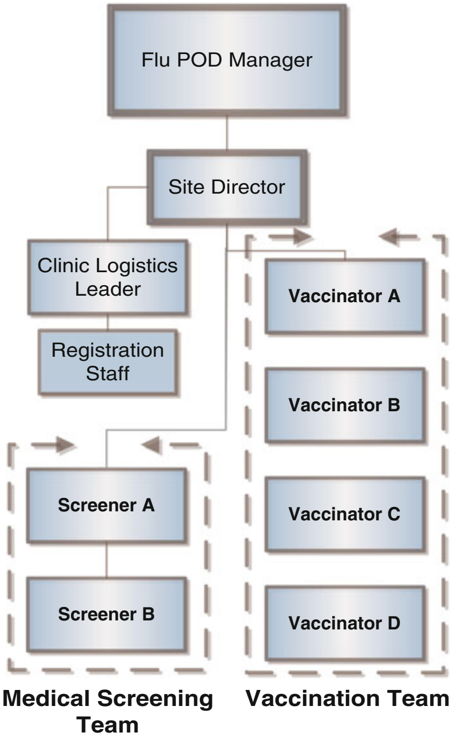
Using the spreadsheet that CPMG creates, planners can easily adjust staffing levels and various inputs until they are satisfied with the efficiency of the PODs. Users can accept default values if they have little information about their PODs, or input more detailed information, such as routing probabilities and process times. Planners can use the model spreadsheets, which they have created using the CPMG, to determine the number of staff members they need to achieve the capacity they need and to design PODs that avoid unnecessary congestion. The CPMG uses data collected from POD exercises, and they have been validated by those exercises and by public health professionals.

Planning mass prophylaxis PODs also requires the close tracking of funds. This will assist local and state health departments when requesting for reimbursement from the federal government or if limited resources are available. A mass prophylaxis budgeting tool (Cho et al. 2011) allows planners to estimate the cost of operating a POD. The model includes setup costs, labor costs, supplies and materials, and facility costs. The model requires the user to enter data about the resources required and their value. Based on this data, the model aggregates everything to determine the overall cost of the POD, with costs by activity.

### 4.2.4 Staffing

Staffing is one of the largest challenges for local health departments when planning PODs. Considerable pre-event planning must be devoted to recruiting, training, and maintaining sufficient numbers of POD staff and volunteers to open and operate the PODs. Many local health departments will depend upon government employees, Medical Reserve Corps members, and other medical and nonmedical volunteers to serve as staff in PODs. As described above, planning models can be used to determine the number of direct service staff (i.e., those who provide the medication to the resident) to assure maximum capacity and to prevent unnecessary bottlenecks. The non-service-related staff (i.e., those who oversee the POD operations, assist with line direction, and serve as replacements) also need to be determined by planners.

**Fig. 12.4** Point of dispensing (POD) organizational chart small venue. It includes suggested staff for a flu clinic serving 300–500 clients. It might take place at a medical clinic or other small site



The NIMS or Incident Command System (ICS) provides a structure to assure a clear chain of command, communication, and supervision and includes the functional areas of finance, logistics, operations, and planning. ICS is the on-site structure used at the scene of an event (U.S. DHS, FEMA 2011c). Because the public health functions during an emergency are much more complex than they are during typical public health activities and require collaborating with many different agencies and disciplines, having a common structure like the ICS is key to the function of the response, which include PODs (Landesman 2001). POD plans should include detailed ICS organizational staffing charts for use in emergencies, and it may be necessary to include other agencies, such as law enforcement, to demonstrate a unified response. For annual flu vaccination PODs and other nonemergency PODs, a modified version of these organizational charts can be used. Nonemergency POD operations provide an ample opportunity to practice using the ICS structure and NIMS. A sample organization chart for a small venue (300–600 clients) flu vaccination POD is seen in Fig. 12.4; a large venue (5,000–7,000 clients) flu vaccination POD is seen in Fig. 12.5.

During a large public health emergency, once their staffing resources have been exhausted, local health departments may need additional staff from other organizations such as colleges, universities, and community-based organizations (e.g., Rotary Clubs, Lions Clubs, sororities and fraternities, professional associations,

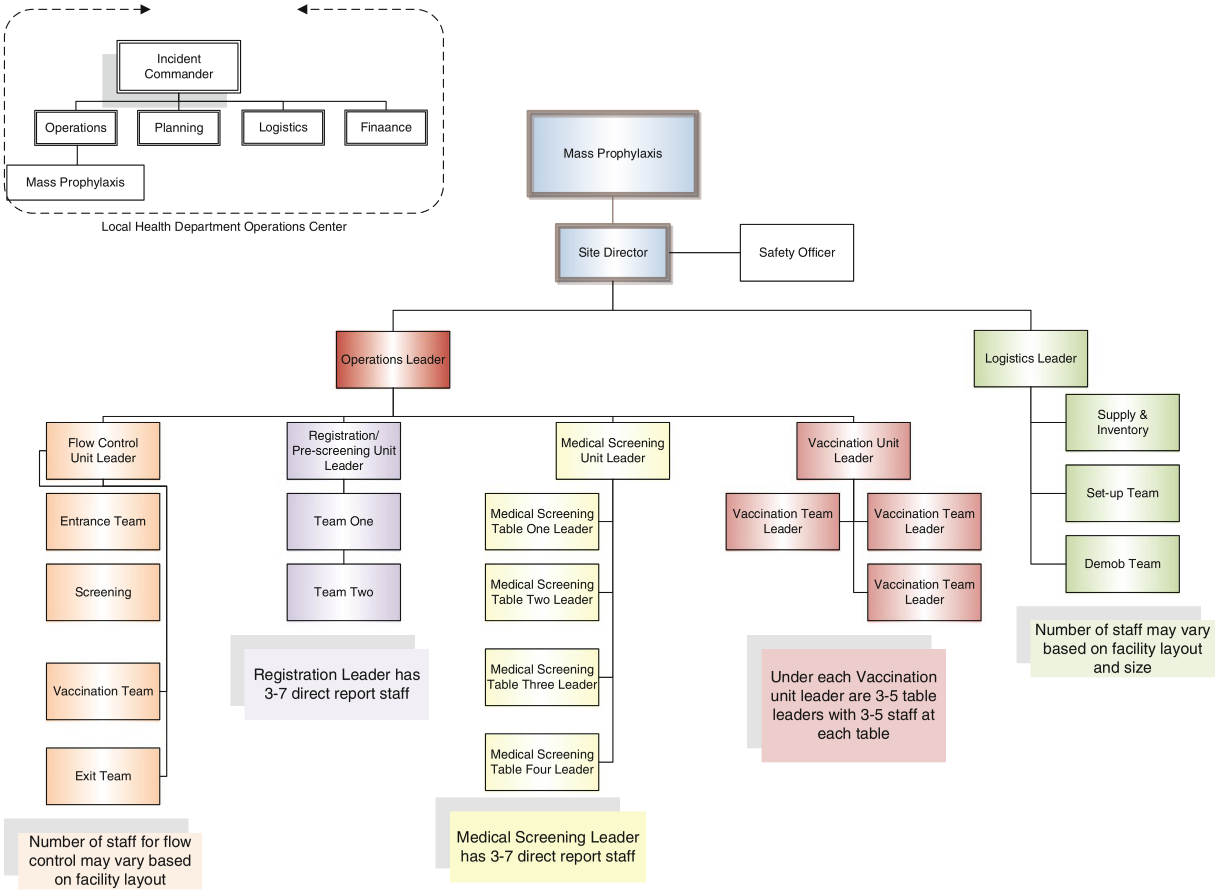

Incident Commander
Operations
Planning
Logistics
Finaance
Mass Prophylaxis
Local Health Department Operations Center
Mass Prophylaxis
Site Director
Safety Officer
Operations Leader
Logistics Leader
Flow Control Unit Leader
Entrance Team
Screening
Vaccination Team
Exit Team
Number of staff for flow control may vary based on facility layout
Registration/ Pre-screening Unit Leader
Team One
Team Two
Registration Leader has 3-7 direct report staff
Medical Screening Unit Leader
Medical Screening Table One Leader
Medical Screening Table Two Leader
Medical Screening Table Three Leader
Medical Screening Table Four Leader
Medical Screening Leader has 3-7 direct report staff
Vaccination Unit Leader
Vaccination Team Leader
Vaccination Team Leader
Vaccination Team Leader
Under each Vaccination unit leader are 3-5 table leaders with 3-5 staff at each table
Supply & Inventory
Set-up Team
Demob Team
Number of staff may vary based on facility layout and size

**Fig. 12.5** Point of dispensing (POD) organizational chart large venue. It includes suggested staff for a flu clinic serving 5,000–7,000 clients. This might take place at a large school, university, or other large facility

faith-based groups, and private industry). Formal agreements (e.g., memoranda of understanding or agreement) should be in place in advance so that this type of mutual aid can be requested and provided promptly.

Once the positions and duties are determined, staff must be told where and when to report to duty. This can be accomplished through notification methods (e.g., phone trees and electronic alert systems) that contact staff at any time, including outside normal hours. Call down drills that test core leaders and workers should be conducted at least quarterly, and improvements should be made prior to the next test.

Finally, planners should create a detailed staffing schedule for multiple shifts if necessary. As part of their POD plan, the planners should be encouraged to develop procedures for the care and feeding of the POD staff and volunteers.

### 4.2.5 Supplies and Logistics

The demands placed upon a local health department to manage supplies and logistics can be overwhelming. Some of the supplies will arrive real time (e.g., vaccine); however, supplies may also be stockpiled (e.g., gloves, alcohol wipes, antibiotics) as part of the pre-event planning. Planners should determine how much medication to order or to request. During annual flu season, supplies and medication may be ordered using existing contractual services within the health departments. The actual amount of annual flu vaccine may be determined by the amount of funding available (i.e., federal, state, and local funds) to purchase vaccine and by the demand from the previous flu season.

The trucks that will move supplies from the RSS, which may be overseen by the state or local health department, to the PODs will follow routes determined by the planners or the trucking company (if one is used). Planners at the state or local level will want to create routes that are short and require little time to distribute the supplies while taking into account the capacity of the trucks (how much they can carry) and the requirements for supplies at each POD. CDC has made available the TourSolver software to assist with this planning task (C2Logix 2011).

During a pandemic influenza situation, where there is a declared emergency, the vaccine supply chain will require the federal government to deliver medications to state health departments which in turn distribute them to the local health departments. During H1N1, localities were solely dependent upon this vaccine supply chain. Vaccine deliveries were sporadic and limited, and challenges arose such as the wrong size needles and syringes were provided and the vaccine required refrigeration. These were real logistical challenges for local health departments.

In addition to medication, other medical supplies (e.g., needles, alcohol wipes, and gloves) need to be considered as well as the removal of hazardous waste. Supplies for staff such as vests for identification, directional signs, communication tools (e.g., walkie-talkies), printed materials for the public (e.g., screening forms, drug/vaccine information), and other event-specific materials should also be included. Local health departments should have an inventory system that is used to ensure that other supplies are easily accessible and on hand for the POD. These supplies may be stockpiled at a specified location and rotated on a regular basis after use.

### 4.2.6 Communications

Public messaging and risk communications, communicating to the public the true health benefits and health risks, is important in planning for a POD operation. In assessing resources, some messages may have been created prior to the event and are part of an overall risk communications plan. The primary communicators with the public are the public information officers at the state and/or local levels. They are charged with ensuring that messages address the target audience and utilize communication methods commonly used among that population. For example, during the H1N1 pandemic, some planners used bus advertisements, designed and field tested by the CDC, in specific county locations to encourage the African-American community to get vaccinated. It is important to communicate clear and concise messages to the public about who is at risk (i.e., target populations) and what preventive measures can be taken (i.e., flu vaccinations). Public information messages should be developed to inform the public that symptomatic persons should not go to a POD to receive prophylactic medication but to seek medical treatment at hospitals or other facilities or from their private physicians. During annual flu vaccine campaigns, the federal government provides public service announcements to encourage the public to get vaccinated; state and local efforts focus more on providing information on when and where to receive the vaccination (Rambhia et al. 2010).

In addition to risk communication, communication between staff and agencies is a critical function. In order to ensure that appropriate and accurate information is being shared with the public and/or internally between agencies and staff, the State and Local Emergency Operations Center or a Local Health Department Operations Center serves as the primary communicator of messages to operations staff during an emergency. During an annual flu event, using the ICS structure, the incident commander or operations section chief may serve as the primary communicator of messages. To ensure successful communication during normal and emergency operations, information technology support is important. To support an effective emergency response, the following items must be in place in all Operations Centers: primary power and back-up power, computers with controlled access and security policies, telephone (cell and landline) systems, radio systems, and information technology staff to support these devices.

### 4.2.7 Outside Partners

Planners must assess the resources and capabilities of outside partners. During pre-event planning, when no PODs are operational, local health departments should collaborate on some level with outside partners in the community on emergency preparedness activities. During a mass prophylaxis operation, these partners (e.g., the Medical Reserve Corps, Emergency System for Advance Registration of Volunteer Health Professionals, Community Emergency Response Teams, volunteers, community-based organizations, pharmacies, faith communities, healthcare organizations, the American Red Cross, and others) may be able to assist by providing resources for the response.

In some cases local health departments may partner with businesses (e.g., retail chains, large employers, banks, drive-through businesses) or healthcare organizations [e.g., Health Maintenance Organizations (HMO) and private PODs] to provide a POD operation. Most hospitals will not serve as public PODs because of the increase in the number of emergency room visits; PODs will hopefully divert those who are not ill. During the H1N1 pandemic, an increased number of state and local health departments used pharmacies to administer vaccine [Association of State and Territorial Health Officials (ASTHO) 2009]. That partnering with pharmacies during H1N1 opened the door for local and state health departments to continue the relationship for future emergencies. Outside businesses and organizations may assist with staffing, outreach to at-risk populations, physical facilities, security, storage of supplies, and other logistical support (IOM 2008).

#### 4.2.8 Political Influences

Political influences are a factor during a POD operation. Sometimes these can be predicted, but sometimes they cannot, and political considerations may trump public health concerns. Politicians must make critical policy decisions with limited information and in a short amount of time in a public health emergency. During the H1N1 pandemic, many jurisdictions closed schools for days and weeks in attempting to contain the disease; however, there continues to be ongoing political and scientific debate over whether the effectiveness of this strategy outweighs the greater economic impact (Cauchemez et al. 2011; Lempel et al. 2009; Jackson et al. 2011; Gift et al. 2010). It is important to be aware that some decisions made prior to, during, and after the POD operations may be solely or partially based on political influences.

### *4.3 Step 3: POD Operations*

POD operations begin in step 3 of the POD Operations Cycle. The POD operations will last for one or more days depending upon the scope of the mass prophylaxis campaign. Annual flu vaccination POD operations are approximately 1–3 months, depending upon public demand and the availability of vaccine and resources. During the H1N1 pandemic, POD operations began in September 2009 and continued beyond May 2010. The schedule of the PODs changed as public demand and the availability of vaccine changed. POD operations will generally flow more smoothly if sufficient time (pre-event) is allocated to the planning considerations and identification of resources discussed in the above section.

### 4.3.1 On-Site Setup and Command Structure

POD staff should arrive at the POD location several hours prior to the start of POD operations to set up the POD. If the POD is large and will be open for many hours, it may be useful to staff a setup team and a demobilization team to perform these tasks in order to prevent fatigue among the staff working the first and last shifts. The staff should utilize the detailed POD layout and design that was created from the site assessment and floor plan of the facility during step 2 of the POD Operations Cycle. Tables, chairs, directional signs, and supplies should be positioned correctly according to the POD layout.

Staff not participating in setup activities should arrive before the start of POD operations to sign in and receive just-in-time training for their POD duties and assignments. There may be last minute reassignments due to absences or other changes.

For most POD operations, the NIMS model is utilized in order to ensure clear leadership roles, delegation of duties, chain of command, personnel reporting system, identification of personnel, and record keeping (Phillips and Williamson 2005). On site the staff follow the ICS communications structure in that there is only one person who oversees the operation and no one supervises more than five direct staff. It is important to note that the ICS structure is flexible and is used for a variety of incidents across all levels of nongovernmental and governmental organizations and many different disciplines. Everyone who is part of the response effort should be trained in the basics of NIMS and the ICS structure (U.S. DHS, FEMA 2011a, c). Please refer back to Sect. 4.2.4 for sample POD organizational charts.

Off-site command personnel will depend upon the size of the entire POD operation. A command post will be set up on site at a predesignated location. The POD leaders will oversee operations from the command post.

### 4.3.2 Just-in-Time Training

Before the POD begins serving the public, it is important to provide just-in-time training to staff, usually led by the team leaders to the service staff. Ideally, all staff should have received training, participated in an exercise or previous PODs, or received information on POD operations prior to the event, but this may not happen. Therefore, it is important to plan for and implement just-in-time training immediately before a POD operation to ensure that all personnel, including those from outside organizations, are able to participate fully. Effective just-in-time training can help staff to:

- Increase their knowledge of the duties they are being asked to perform, which may be different from their normal tasks
- Feel more confident to perform these duties in an unfamiliar environment and under high-stress circumstances
- Work better with unfamiliar people and clients

Just-in-time training has been shown to be effective if it supports the responder, provides opportunity to practice, and takes into account the cultural environment (Cress et al. 2010).

#### 4.3.3 Flexibility and Limitations

Once the POD is open to the public, there will be challenges. Planners should recognize this and realize that no one plan can address every possible challenge. The important thing is to be flexible when trying to solve problems. For example, if a queue of people has formed outside the POD, ask those in the line to fill out the paperwork while they are waiting (rather than when they arrive at registration).

Some challenges require solutions that cannot be implemented feasibly during operations. Planners should note such problems and analyze them thoroughly afterwards.

### *4.4 Step 4: Post-analysis and Corrective Actions*

The fourth (and final) step of the POD Operations Cycle involves capturing lessons learned and identifying next steps. This includes the things that worked well during the POD operation and the challenges. This step is an important opportunity for local health departments to learn and to document how they will improve their mass prophylaxis plans in the future.

Many state and local governments are strongly encouraged to follow the Homeland Security Exercise and Evaluation Program (HSEEP) standards (https://hseep.dhs.gov). HSEEP is used to provide a standardized policy, methodology, and language for designing, developing, conducting, and evaluating all exercises. However, some of the templates can also be adapted for actual events (Montgomery County, Maryland Advanced Practice Center for Public Health Preparedness and Response 2007). After-action reports developed by state and local health departments for the H1N1 pandemic as well as annual flu vaccination PODs were created using HSEEP standards.

### 4.4.1 Hot Wash and After-Action Report

Although a hot wash is usually conducted after an operations-based exercise, many local health departments and other emergency responders also find it useful after a real event. A hot wash is a facilitated discussion that allows participants in the POD operations to engage in self-assessment of their roles and responsibilities and to help form an overall assessment of the response. Ideally, the hot wash is conducted soon after the POD operation is complete, preferably the same day while information is still fresh, and by a facilitator who was not part of the operation. The facilitator works to ensure that the discussion is constructive and brief and focuses on both the strengths of the operation and the areas for improvement. Some local health departments may develop evaluation forms that can be distributed to all participants, while others may choose to designate someone to take notes during the hot wash discussion. The HSEEP standards provide a template for the hot wash minutes. Whatever form this review may take, it is imperative to document this information in order to include it in the after-action report (AAR) (U.S. DHS 2007).

The AAR serves as the primary documentation of what happened during a POD operation. The AAR describes what happened, outlines best practices or strengths, identifies areas that need improvement, and suggests recommendations for improvement (U.S. DHS 2007). The HSEEP standards provide a template for an AAR and specific requirements including an improvement plan.

### 4.4.2 Improvement Planning

Improvement planning is the final piece of an effective POD operation. Planners should use the recommendations recorded in the improvement plan in the AAR and put them into action. A facilitated after-action conference can be used to bring together all the agencies and organizations involved in the real event to first identify the corrective actions and second who and by when they will be completed. Concrete corrective actions are then prioritized, tracked, and incorporated into a continuous quality improvement plan. Many of the actions may require changes in POD operations plans, fine-tuning policy and procedure manuals, or partnering with additional organizations to acquire additional staff. It is crucial to begin to implement these recommendations in order to improve POD operations in the future.

## 5 Future Research and Practice Challenges

Although PODs remain the primary means for local health departments to administer prophylactic medication to the population, there are additional modalities being studied. Other modalities, such as MedKits, the US Postal model, and stockpiling of pharmaceuticals, are critically important to explore, especially for other diseases or agents when time is of the essence. These modalities could be used by local jurisdictions to support POD operations by closing gaps and increasing access. In addition, there are new and innovative technologies that local health departments have developed in order to improve efficiency of PODs. In the remainder of this section, the authors discuss opportunities for future research, and practical challenges to be overcome, related to these additional modalities.

### *5.1 Additional Modalities to Support PODs*

In 2006 the CDC partnered with the Missouri Department of Health and Senior Services to conduct an 8-month evaluation on the effectiveness of using MedKits as another means of providing timely medication to the public during a public health emergency. MedKits is a concept that prepositions medications, in this case antibiotics for anthrax, in individual households. The MedKit concept was developed in collaboration with the Food and Drug Administration (FDA) and met all federal and state regulatory requirements. For this research study, approximately 4,076 households participated, mainly from St. Louis City, St. Louis County, and St. Charles County. Each household was asked to maintain a MedKit in the home as directed and to reserve it for emergency use. Results found that 97% of the study respondents returned their MedKits upon completion of the study; 75% of the respondents reported that having the MedKit in their home increased their awareness to prepare for a public health emergency and the majority, 94% or more of each cohort, acknowledged that they would like to have a MedKit in their homes (U.S. DHHS, CDC 2007). The FDA and CDC continue to explore MedKits as one of the different modalities for increasing the nation's capacity to respond to a public health emergency requiring medical countermeasures.

In 2009 President Obama issued an Executive Order which directs DHHS, DHS, and the US Postal Service (USPS) to establish a model that allows postal workers to deliver medicine directly to residences. The Postal Model, as it is known, serves as another modality for states, cities, and counties to use to enhance their existing mass prophylaxis plans [U.S. DHHS, Office of the Assistant Secretary for Preparedness and Response (ASPR) 2011]. Funding was provided to localities in 2011 to test specific components of the model.

Currently, additional medical countermeasure modalities are being explored by federal government officials. The priority is to study strategies that can increase local access to lifesaving medications in a timelier manner. One strategy involves prepositioning medications. This could be caches of pharmaceuticals stored at or near where they will be dispensed such as workplaces, pharmacies, and other healthcare facilities (IOM 2011). Another strategy is prepositioning or predispensing of medications with first responders as a strategy to increase the response time to biological or chemical emergencies (U.S. DHHS, CDC 2007; IOM 2011). The September 2011 Institute of Medicine Report, *Prepositioning Antibiotics for Anthrax*, describes in detail the pros and cons of these strategies and encourages more research if the potential benefits outweigh the potentials risks and increased costs (IOM 2011).

It is important that local as well as federal and state health officials continue to research, evaluate, and share additional strategies that can be used to support POD efforts. PODs should be viewed as only one strategy state, local, and tribal jurisdictions can use to quickly and effectively dispense medical countermeasures to their populations.

## *5.2 Using Technology to Improve POD Functionality*

Federal, state, and local health departments continue to create best practices and develop innovative ideas and solutions to improve the overall function of PODs. The following are examples of innovative ways local and state health departments are exploring technology to address POD functionality, specifically bottlenecks or congestion of people in PODs.

One innovative practice is the use of handheld devices to perform the screening. The Montgomery County, Maryland Advanced Practice Center, and the University of Maryland have studied this possibility using personal digital assistants (PDAs), Blackberry®, and iPhone® devices. They have developed a basic patient screening for both anthrax and hepatitis A scenarios. The PDA version was field tested in 2009 during an actual drive-through clinic in Tarrant County, TX, for a hepatitis A vaccine. The results found that screeners using the PDA screening were two times faster than those using the traditional paper screening (Tarrant County, Texas Advanced Practice Center 2009). More information about this project can be found at http://www.isr.umd.edu/Labs/CIM/projects/clinic/.

The Bay Area Mass Prophylaxis Working Group (BAMPWG) in California has created an online screening form for anthrax prophylaxis based on the concept that residents can easily prescreen themselves for medications prior to arrival at a POD (see http://www.bayareadisastermeds.org/). The residents answer a limited number of questions for up to 20 people then receive a printout of which resident should receive which drug, along with dosing instructions, if applicable. The site, pretested with 8,000 people per minute, was created in web-based software that can be modified easily by the BAMPWG. Some of the benefits in using an online screening form are that residents can go directly through an "express" route at the POD and that throughput increases at the POD by reducing the number of people who need to go to screening (Relucio and Pine 2010).

Another online screening form used during the H1N1 pandemic was developed by Yolo County Health Department, California. Created in order to ease the screening process for staff and to increase throughput at the POD, it was field tested in 2008 at three exercises and used real time during the H1N1 pandemic. The health department had anticipated approximately 10% of the population would arrive at PODs with printed screening forms including answers to questions as well as identifying the medication. In testing, it was found that 11.4% utilized this process. In the future Yolo County is considering having the screening tool available in other languages and compatible with smartphones (Carey 2010). Strategies that use technology and innovation are the future. As resources are reduced, it only makes sense to utilize these strategies to improve POD functionality.

## 6 Conclusion

Planning and managing a POD operation is an important function for local health departments in conjunction with local county government and community partners. The POD Operations Cycle represents the steps and activities associated with organizing a successful operation. However, it should be remembered that the Cycle is meant to act as a guide only and that it is important for plans and operations to be scalable and flexible. The importance of including lessons learned and follow-up on improvement plans for real events and exercises is vital to improving response efforts.

As the threats become more complex, the response effort by local health departments will continue to remain a challenge. The use of technology and innovative solutions to tackle these challenges will improve the planning and response to these threats.

## References

Aaby K, Abbey R, Herrmann JW, Treadwell M, Jordan C, Wood K (2006a) Embracing computer modeling to address pandemic influenza in the 21st century. J Public Health Manag Pract 12(4):365–372

Aaby K, Herrmann JW, Jordan C, Treadwell M, Wood K (2006b) Montgomery County's Public Health Service uses operations research to plan emergency mass-dispensing and vaccination clinics. Interfaces 36(6):569–579

Ablah E, Scanlon E, Konda K, Tinius A, Gebbie K (2010) A large-scale points-of-dispensing exercise for first responders and first receivers in Nassau County, New York. Biosecur Bioterror 8(1):25–35

Agnes M (2005) Webster's new college dictionary. Wiley, Cleveland

Association of State and Territorial Health Association (ASTHO) (2009) Operational framework for partnering with pharmacies for administration of 2009 H1N1 vaccine. Retrieved from the Association of State and Territorial Health Association. Retrieved April 2, 2012 website: http://www.astho.org/Display/AssetDisplay.aspx?id=2613

C2Logix (2011) SNS TourSolver. Retrieved April 2, 2012 from http://snstoursolver.c2logix.com/

Carey D (2010) On-line medical screening: the advent of technology for mass prophylaxis or mass vaccination activations. SNS Summit, 26–29 July 2010

Cauchemez S,Bhattarai A, Marchbanks TL, Fagan RP, Ostroff S, Ferguson NM the Pennsylvania H1N1 working group (2011) Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. Proc Natl Acad Sci USA 108(7):2825–2830. doi:10.1073/pnas.1008895108 (Early Edition 31 January 2011, 1–6)

Cho BH, Hicks KA, Honeycutt AA, Hupert N, Khavjou O, Messonnier M, Washington ML (2011) A tool for the economic analysis of mass prophylaxis operations with an application to H1N1 Influenza vaccination clinics. J Public Health Manag Pract 17(1):E22–E28

Cress C, Spitzer J, Stephens A, Oxman G (2010) Enhancing training during public health emergencies: an inclusive just-in-time training (JITT) approach. Retrieved 4 March 2011: http://web.multco.us/sites/default/files/health/documents/whitepaper_jitt.pdf

Fiore AE, Uyeki TM, Broder K, Finelli K, Euler GL, Singleton JA et al (2009) Use of influenza A (H1N1) 2009 monovalent vaccine: recommendations of the advisory committee on immunization practices (ACIP), 2009. MMWR Recomm Rep 58(RR-10):1–8

Fiore AE, Uyeki TM, Broder K, Finelli K, Euler GL, Singleton JA et al (2010) Prevention & control of Influenza with vaccines—recommendations of the advisory committee on immunization practices (ACIP) 2010. Morb Mortal Wkly Rep 59(RR08):1–62

Gift TL, Palekar RS, Sodha SV, Kent CK, Fagan RP, Archer WR, Edelson PJ et al (2010) Household effects of school closure during pandemic (H1N1) 2009, Pennsylvania, USA. Emerg Infect Dis 16(8):1315–1317. doi:10.3201/eid1608.091827

Gist AW (2011) Slides for presentation on H1N1 statewide focus group experience. Maryland Department of Health and Mental Hygiene, Office of Minority Health and Health Disparities, Baltimore

Herrmann JW (2008) Disseminating emergency preparedness planning models as automatically generated custom spreadsheets. Interfaces 38(4):263–270

Heymann DL (2004) Control of communicable diseases manual, 18th edn. American Public Health Association, Washington, DC

Hupert N, Cuomo J (2003) Bioterrorism and epidemic outbreak response model (BERM). Retrieved 2 April 2004: http://www.ahrq.gov/news/press/pr2003/btmodpr.htm

Institute of Medicine (IOM) (2003) The future of the public's health in the 21st century. The National Academies Press, Washington, DC

Institute of Medicine (IOM) (2008) Dispensing medical countermeasures for public health emergencies: workshop summary. The National Academies Press, Washington, DC

Institute of Medicine (IOM) (2011) Prepositioning antibiotics for Anthrax. The National Academies Press, Washington, DC

Jackson C, Mangtani P, Vynnycky E, Fielding K, Kitching A, Mohamed H et al (2011) School closures and student contact patterns. Emerg Infect Dis 17(2). doi:10.3201/eid1702.100458

Landesman LY (2001) Public health management of disasters: the practice guide. American Public Health Association, Washington, DC

Lee EK, Chen C, Pietz F, Benecke B (2009) Modeling and optimizing the public-health infrastructure for emergency response. Interfaces 39(5):476–490

Lempel H, Hammond RA, Epstein JM (2009) Economic cost and health care workforce effects of school closures in the U.S. Retrieved on April 26, 2011 from Brookings Institution website: http://www.brookings.edu/papers/2009/0930_school_closure_lempel_hammond_epstein.aspx

Montgomery County, Maryland Advanced Practice Center for Public Health Preparedness and Response (2007): April 2, 2012 Notes from the field: a collection of emergency preparedness exercise and evaluation reviews. Retrieved from http://www.montgomerycountymd.gov/hhstmpl.asp?url=/content/hhs/phs/apc/notesfromthefield.asp

Morton RF, Hebel JR, McCarter RJ (2001) A study guide to epidemiology and biostatistics, 5th edn. Aspen Publishers, Gaithersburg, p 1

National Association of County and City Health Officials (NACCHO) (2003): April 2, 2012 Promoting and protecting healthy communities: a city officials guide to public health. Retrieved from NACCHO's website at: http://www.naccho.org/advocacy/resources/upload/City-Official-Guide.pdf

National Association of County and City Health Officials (NACCHO) (2008): April 2, 2012 The 2008 national profile of local health departments fast facts. Retrieved from NACCHO's website at: http://www.naccho.org/topics/infrastructure/profile/resources/2008report/upload/profilebrochure2009-10-17_COMBINED_post-to-web.pdf

Phillips FB, Williamson JP (2005) Local health department applies Incident Management System for successful mass influenza clinic. J Public Health Manag Pract 11(4):269–273

Rambhia KJ, Watson M, Sell TK, Waldhorn R, Toner R (2010) Mass vaccination for the 2009 H1N1 pandemic: approaches, challenges, and recommendations. Biosecur Bioterror 8(4):321–330. doi:10.1089/bsp. 2010.0043

Relucio K, Pine A (2010) Selecting an antibiotic for individuals requiring anthrax prophylaxis: standardizing practices across 11 California counties. SNS Summit, 26–29 July 2010

Shea DA, Lister SA (2003) The biowatch program: detection of bioterrorism. Congressional research service report no. RL 32152, 19 November 2003. Retrieved on 14 April 2011: http://www.fas.org/sgp/crs/terror/RL32152.html

Tarrant County, Texas Advanced Practice Center (2009) Summary of evaluation of eMedCheck for hepatitis A drive-through POD. Pennington Field, Bedford

Trust for America's Health (TFAH) (2010) Issue brief: fighting flu fatigue. Retrieved on 7 March 2011: http://healthyamericans.org/assets/files/TFAH2010FluBriefFINAL.pdf

U.S. Department of Health and Human Services (DHHS), Centers for Disease Control and Prevention (October 9, 2009) Update on influenza A (H1N1) 2009 monovalent vaccines. Morb Mortal Wkly Rep 58(39):1100–1101

U.S. Department of Health and Human Services (DHHS), Centers for Disease Control and Prevention (CDC) (2001) The health alert network (HAN). Retrieved on 18 April 2011: http://www.bt.cdc.gov/DocumentsApp/HAN/han.asp

U.S. Department of Health and Human Services (DHHS), Centers for Disease Control and Prevention (CDC) (2011) Assessment of ESSENCE performance for influenza-like illness surveillance after an influenza outbreak—U.S. Air Force Academy, Colorado, 2009. Morb Mortal Wkly Rep 60(13):406–409

U.S. Department of Health and Human Services (DHHS), Centers for Disease Control and Prevention (CDC) (2011) Strategic national stockpile (SNS). Retrieved on 4 April 2011: http://www.cdc.gov/phpr/stockpile.htm

U.S. Department of Health and Human Services (DHHS), Centers for Disease Control and Prevention (CDC) (2011) Public health preparedness: 2011 state-by-state update on laboratory capabilities and response readiness planning. Retrieved on 14 October 2011: http://www.cdc.gov/phpr/pubs-links/2011/documents/SEPT_UPDATE_REPORT_9-13-2011-Final.pdf

U.S. Department of Health and Human Services (DHHS), Office of the Assistant Secretary for Preparedness and Response (ASPR) (2011) Postal model for medical countermeasures deliver and distribution. Retrieved on 5 April 2011: http://www.phe.gov/Preparedness/planning/postal/Pages/default.aspx

U.S. Department of Health and Human Services (DHHS), Centers for Disease Control and Prevention (2007) CDC's division of strategic national stockpile emergency MedKit evaluation study summary. Retrieved 26 April 2011: http://www.bt.cdc.gov/agent/anthrax/prep/pdf/medkit-evaluation-summary-2007.pdf

U.S. Department of Health and Human Services (DHHS), Centers for Disease Control and Prevention (2009) Advisory committee on immunization practices (ACIP). Summary report, 29 July 2009. Retrieved on 20 March 2012: http://www.cdc.gov/vaccines/recs/acip/downloads/min-archive/min-jul09.txt

U.S. Department of Health and Human Services (DHHS), Centers for Disease Control and Prevention (CDC), National Center for Chronic Disease Prevention and Health Promotion (2010) Reach U.S. finding solutions to health disparities: at a Glance 2010. Retrieved on 7 March 2011: http://www.cdc.gov/chronicdisease/resources/publications/aag/pdf/2010/REACH-AAG.pdf

U.S. Department of Homeland Security (DHS) (2007) Homeland security exercise and evaluation program (HSEEP): volume III: exercise and evaluation and improvement planning. Retrieved on 4 March 2011: https://hseep.dhs.gov/support/VolumeIII.pdf

U.S. Department of Homeland Security (DHS) (2008) Overview: ESF and support annexes coordinating federal assistance in support of the national response framework. Retrieved 26 April 2011: http://www.fema.gov/pdf/emergency/nrf/nrf-overview.pdf

U.S. Department of Homeland Security (DHS), Federal Emergency Management Agency (FEMA) (2007) Robert T. Stafford disaster relief and emergency assistance act, as amended, and related authorities. Retrieved on 18 October 2011: http://www.fema.gov/pdf/about/stafford_act.pdf

U.S. Department of Homeland Security (DHS), Federal Emergency Management Agency (FEMA) (2011a) NIMS frequently asked questions. Retrieved on 4 April 2011: http://www.fema.gov/emergency/nims/FAQ.shtm#item1c

U.S. Department of Homeland Security (DHS), Federal Emergency Management Agency (FEMA) (2011b) Emergency support function#8-public health and medical services annex. Retrieved on 30 September 2011: http://www.fema.gov/pdf/emergency/nrf/nrf-esf-08.pdf

U.S. Department of Homeland Security (DHS), Federal Emergency Management Agency (FEMA) (2011c) Incident command system (ICS) overview. Retrieved on 4 October 2011: http://www.fema.gov/emergency/nims/IncidentCommandSystem.shtm

Vinter S, Lieberman DA, Levi J (2010) Public health preparedness in a reforming health system. Harvard Law Policy Rev 4(2):337–360

# Chapter 13
# Emergency Departments: “Repairs While You Wait, No Appointment Necessary”

**Kenneth N. McKay, Jennifer E. Engels, Sahil Jain, Lydia Chudleigh, Don Shilton, and Ashok Sharma**

## 1 Introduction

The flows within and surrounding Emergency Departments have been the subject of systematic, mathematical analyses since the 1950s (e.g., Newell 1954). While there has been a long history, the degree of quantitative analysis using methodologies such as simulation and queuing theory has been relatively limited compared to other areas of healthcare (e.g., see reviews by Fries 1979; Boldy 1976; Pierskalla and Brailer 1994; Flagle 2002; Fomundam and Herrmann 2007; Wiler et al. 2011). In general, the quantitative analyses performed on the Emergency Department flows have been in response to classical operations management challenges about demand and supply, wait times, time in system, queuing, staffing, and supply chain interactions. With a few exceptions, the detailed research on flows has not included specific resource characteristics and behavior, implicitly or explicitly assuming that such factors do not significantly influence the research effort or results.

K.N. McKay (✉)
Department of Management Sciences, University of Waterloo, Waterloo, ON, Canada
e-mail: kmckay@uwaterloo.ca

J.E. Engels
Manulife Financial Corporation, Business System Analyst, Kitchener, ON, Canada

S. Jain, MD
Consultant, Toronto, ON, Canada

L. Chudleigh
VP Quality and Performance Management, St. Mary’s General Hospital, Kitchener, ON, Canada

D. Shilton
St. Mary’s General Hospital, President, Kitchener, ON, Canada

A. Sharma, MD
St. Mary’s General Hospital, Grand River Hospital, Joint Chief of Staff, Kitchener, ON, Canada

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_13, © Springer Science+Business Media New York 2013

In this chapter, we focus on research targeted towards the operations of the Emergency Department and investigate the underlying problem characteristics and salient features. We do not address clinical decision making within emergency medicine, nor topics such as patient safety, or nurse and physician planning models. These are all important topics which must be considered in parallel when analyzing an Emergency Department in a holistic fashion.

The intent of this chapter is multifold: (a) to summarize and review relevant research identifying strong contributions and solid foundations upon which to move forward, (b) to discuss the closest analogy to the Emergency Department and identify shared characteristics which might assist with understanding the Emergency Department conundrum, and (c) to present possible detailed modeling characteristics that might be relevant when building and validating models at the operational level.

The chapter contains a discussion of Emergency Departments and the relationship to repair shops, modeling challenges, a literature review of recent results in aggregate and detailed modeling of Emergency Departments, a field study methodology section on how an Emergency Department might be studied, followed by subsections on key constraints and concepts found in the literature and additional modeling requirements observed in a field study at St. Mary's General Hospital. The chapter concludes with a discussion on potential research topics. Throughout the chapter, we will interleave theoretical and applied insights, as appropriate, based on the authors' experience with an Emergency Department at St. Mary's General Hospital, Kitchener, Ontario, Canada.

## 2 Repair Shop Analogies and Translational Insights

A major challenge when performing research on an Emergency Department is understanding how the department functions throughout the day and year, how various resources relate to each other, what input flow implies, what types of uncertainty exist, and so forth. While an Emergency Department has some similarities to many other service situations, there are a number of traits which are unique or extreme. As noted in the chapter title, an Emergency Department can be described as a "repairs while you wait, no appointment necessary" situation.

Repair shops are a subset of the dynamic job shop problem combined with service characteristics. In a repair shop, specialized skills and knowledge are needed to assess the specific service required and perform the repair service. The possible priorities, risks, and benefits associated with the repair will depend on the context. In Emergency Departments, the context is centered on human life, medical diagnosis, treatment, and the patient's well-being. The healthcare context implies a service/repair situation with high expectations, risks, and many constraints. In this chapter, the repair shops we are referring to are not maintenance shops or the product repair areas located in a factory associated with a manufacturing line (e.g., Sleptchenko et al. 2005) but are small self-standing repair enterprises.

## 2.1 Repair Shop Characteristics and Implications

Some of the extreme or severe service/repair traits attributable to the Emergency Departments and not found in simple service cases are:

- Triage or initial inspection requires experience and specialized training to assess and to route the prioritized work to the appropriate area. The complexity or specialty associated with the case is also assessed to augment the routing decision.
- The time in queue or before service is critical because the work can deteriorate further, increasing the priority and complexity. For example, repairs on materials can be susceptible to aging, rupture, and exposure, and might also have some form of breakage or leakage causing negative impact during queue time.
- External, specialized consultations from scarce resources are relatively common for both diagnosis and treatment advice.
- There are many legal and professional policies and practices mandated throughout the process (e.g., charting).
- Once certain types of repairs are started, they cannot be set aside, stopped, or interrupted.
- Customers expect correct and timely made critical repairs; return visits and delays are not desired or acceptable.

These extreme characteristics can be seen in certain service/repair fields other than healthcare. For example, many of these aspects can be observed in a repair facility servicing nuclear plants and the mining industry (McKay 1987). In healthcare, some of the extreme constraints may drive particular behaviors. For instance, the real or perceived implications of patient transfer between two attending physicians during a shift change or anticipated patient complexity can possibly affect which patients a physician will see towards the end of a shift and what activities the physician does towards the end of a shift. Some physicians may only attend to extremely critical patients, or new patients who can be processed in their entirety without a handoff, and they will try to clear out existing patients to reduce the amount of transfer to the incoming physician.

If we make the assumption that Emergency Departments are a subset of the small, independent service/repair shops, there are a few observations that can be made which assist in understanding how to study, analyze, and improve the department:

- In repair shops, there are many resource conflicts ranging from floor space, specialized equipment, supporting services, technicians, and infrastructure. It is usually not possible to have enough tools and resources to simultaneously satisfy all possible requirements.
- If a repair shop is running 24 × 7, it is rare that all supporting services are also running 24 × 7. This creates resource and skill shortages beyond the day shift.
- Certain types of repairs or situations can require substantially greater amounts of resources and capacity than others. If the required resources are in scarce supply but are in frequent demand, other repair jobs will be often delayed. There can

also be a high variance in requirements and in processing times. Specific work mix can create floating bottlenecks or dramatically affect the performance of a bottleneck resource.

- Repair shops use loosely formed teams in a collaborative fashion. Communication flows and efficiency are critical for successful execution. If communication practices are subpar, mistakes and delays result.
- Repair shops typically work on more than one repair at a time. If there are too many repairs active, the workflow is too fragmented and all repairs incur extended delays as the repair team attempts to work on all jobs in parallel; however, if there are too few repairs active, the repair resources might be idle waiting for secondary tasks to be completed before they can complete their own work.
- Repairs may require additional specialist advice or specialized tests to be performed. If the advice or tests are not available in a timely fashion, workflows start to slow down which can cause additional work to be started, leading to congestion in the area.
- Repairs may span shift boundaries and require debriefing and transfer of knowledge about the problem, diagnosis, treatment plan, work done to date, and current status. While the transfer is being processed, actual repair work is not being performed on any job. This can delay all work in the queue. Repairs across shift boundaries can also increase the risk of miscommunication, possible errors, and increased inefficiency.

Observations such as the above can inform a field study and data analysis. At St. Mary's General Hospital, this knowledge meant that we specifically looked for floating bottleneck resources (e.g., nurses, beds, physicians, testing, inpatient admitting), patient mix and arrival patterns that implied substantial differences in resource requirements and processing times (e.g., old and young patients), issues that would starve or block bottleneck resources (e.g., delays in lab tests, lack of beds), shift-to-shift hand-off processes (e.g., how long resources do the processes take), external dependencies (e.g., specialists), and any special "out of the ordinary" events or situations which had the potential for derailing the department (e.g., isolation requirements, code announcements in other parts of the hospital to which the Emergency Department personnel respond to).

The service/repair nature of Emergency Departments implies that most of the classic challenges will exist: nonuniform demand patterns throughout the day, high variance in the demand patterns, and peak demand servicing. Additional operational challenges are:

1. High demand, process variability, and $24 \times 7$ work patterns will challenge traditional learning curves, skill acquisition, and organizational learning. While some tasks are routine and performed often enough on each shift for the learning processes to take place, other specific tasks may not take place often enough for every nurse, physician, or other staff member on every shift to have gained prior experience through practice. For the classical learning theory to apply, the same

“standard” tasks need to take place with cognitive reflection many times within a relatively short time window and also to have ongoing practice opportunities.

2. High demand, process variability, and 24 × 7 work patterns will challenge the creation, deployment, and sustainability of any improvement activity. Many of the continuous improvement and lean methods originating in industry assume the flow characteristics associated with high-volume and low-mix situations. Emergency Departments will rarely satisfy the underlying assumptions related to analytical techniques such as statistical process control.
3. Medical diagnostic and treatment situations are often a web of activities involving teams, multiple resources with non-standardized routings, and flows created dynamically for each patient. Floating bottlenecks, unintended starving, and blocking of bottlenecks can create ripple effects affecting many other direct and indirect resources and their related processes.
4. Extensive relationships with other parts of the hospital such as diagnostic testing and admittance into wards create interdependencies, flow pinch points, and possibly flow stoppage. These types of relationships and dependencies are likely rare in self-standing repair shops.
5. In many jurisdictions (e.g., US EMTALA policy, see www.cms.gov), the Emergency Department cannot refuse or turn away a patient under certain conditions, and while some repair shops have a “never say no to a possible job”, the legal requirements create a tightly constrained situation beyond the authority and control of the Emergency Department.
6. Risk management practices are also more extreme in healthcare situations because of malpractice and the direct impact on human health and well-being.

Variance is perhaps the most important, general aspect to focus upon in a production setting and has been noted as early as 1832 (Babbage 1832). Internal and external sources of variance impact fixed and variable costs, yield, storage requirements, and management methods. Variance is not always equal throughout a process, and variance towards the end of a process does not carry the same implications as variance near the beginning. For example, variance beyond a desired level towards the end of a process can result in losing all of the investment made in previous steps, as well as perhaps not leaving enough time before the delivery due date for recovery. Relatively common sources of variance can be managed and the negative impacts mitigated, but rare events are more difficult to deal with and can greatly increase variance when they occur. The size of the Emergency Department will also affect how important variance will be to daily operations. For example, in a small community hospital, there might be two physicians on duty, and if one is required to help elsewhere in the hospital, the impact is greater than at a larger teaching hospital having more resources on duty.

One of the challenges in a repair shop situation is creating a reasonably balanced workload for the majority of resources. For example, if one resource does a batching process instead of a pipeline transfer, the lumpy deliveries can create peaks and troughs of work undulating through the department. In the Emergency Department, this might take place when a physician processes a number of charts before providing test order instructions to the other Emergency Department staff. Then the staff members have multiple tests to be processed in parallel instead of having a slow trickle feed, spaced apart.

## 3 Modeling Framework

The focus of this chapter is on operational, tactical, and planning models at the hospital and not the types of models used in policy making by governments. In mature research areas such as manufacturing assembly lines, there have been many quantitative efforts on problem abstraction and model frameworks (e.g., Buzacott and Shanthikumar 1993). It is difficult to find similar quantitative research abstracting the Emergency Department situation, creating modeling frameworks. It appears that most of the Emergency Department literature pertains to a specific site with a specific research question. Exceptions do exist at higher, conceptual levels. For example, the conceptual model by Asplin et al. (2003) provides a starting point for framing and understanding the challenges faced by Emergency Departments and for guiding research efforts.

Although general, abstract models have not been developed or broadly adopted at the detailed, quantitative level, there have been a number of claims about generalizable simulation and analytical models of Emergency Departments. For example, Facchin et al. (2010) state in their paper about a generalized, flexible simulation model of an Emergency Department that “no essential characteristics of any department behavior is neglected; the model may be adapted to many different services.” Fletcher et al. (2007) describe a generic UK model for Emergency Departments that was used in policy making on the broad scale and then at specific hospitals. The terms “generalizable” and “generic” are problematic because the claims may be appropriate for the specific units involved in the research but not for all Emergency Departments worldwide (e.g., Shingo 2010).

Fletcher and Worthington (2009) provide a discussion about types of generic models, create a four-tiered taxonomy for discussing generic models, and then conduct a review on the modeling literature covering various parts of the hospital interacting with emergency flows. They note that “more evidence of specific A&E (accident and emergency) models than generic” exists. They observe possible issues for the three generic A&E models reviewed with either validation or implementation. Fletcher et al. (2007) describe their own experience using the generic UK emergency model in a local setting, and while they used the generic model in various ways with ten Emergency Departments, none of the experiences went as far as to “test if the predicted improvements were made.”

Unfortunately, supporting evidence that any of the developed models and tools are indeed general beyond a specific context has not been forthcoming, and there was no evidence found that generic models developed by one research effort have had ongoing research performed with them, nor that they have been adopted by another researcher or practitioner.

The operation of Emergency Departments is not dissimilar to that facing relatively small, dynamic repair shops in industry, which also have not been systematically or broadly studied as part of the industrial engineering or operations management research agendas. While it is sometimes possible to translate results from one domain to another to gain insights or to create a foundation, it is not possible in this case as an insufficient amount of results and insights exist for translation. There has been work on maintenance shops and on production repair areas directly associated with a production unit, but there have not been similar efforts on a small independent repair shop. We can use general behavior and flow dynamics from repair shops to help understand the Emergency Department as described in Sect. 2 of this chapter, but there are not any analytical results for comparison or to transfer.

One goal of this chapter is to discuss constraints and dynamic relationships which might be relevant when studying or modeling a specific Emergency Department or when thinking about general models for Emergency Departments. Before discussing specific modeling issues or delving into the literature review, it is useful to consider why truly general models of the Emergency Department might not have been developed yet, and might never be. There are four framework challenges which are relevant to this question: (a) vertical or decision scope integration, (b) horizontal or multipurpose inclusion, (c) specific purpose-built models, and (d) in situ relevance. To put it more concisely: depth, breadth, specialty, and context.

The vertical or decision scope challenge relates the three level taxonomy of models dating back to the seminal work of Anthony (1965): strategic, tactical, and operational. The three levels of hierarchical decision making are for different decision horizons and scope, with:

- Strategic and long-term planning level: decisions typically focused on location, capacity and size, and type of business or specialization (e.g., decisions affecting many years, major investments in time and funds).
- Tactical level: decisions affecting time frames measured in months or fiscal quarters, dealing with policies and intermittent decisions such as hiring and staffing levels, guided by the strategic decisions.
- Operational level: decisions made daily, weekly, or monthly addressing the day-to-day details of what is done, when, how, and by whom in the most efficient and effective fashion as guided by the constraints imposed by the strategic and tactical levels.

The vertical decomposition often implies different degrees of granularity with operational decisions requiring finer detail and many factors included and the strategic level using aggregated data and models. While it is sometimes possible to combine two levels (e.g., strategic and tactical, or tactical and operational), it is rare to find models that combine all three vertical levels of decision making because of the data and model component granularity.

Horizontal or multipurpose inclusion suggests that a general, horizontal model is possible at each layer, addressing similar types of decisions, using a common level of information granularity. This is sometimes possible in manufacturing and service situations where the processes are well-defined and integrated models crafted. These situations are relatively clean and pristine (e.g., a controlled supply chain with highly automated factories). Imagine the complexity and challenges that an Emergency Department tactical model might entail when combining all relevant and interdependent tactical topics such as physician/nurse staffing levels, physician skill and training, department flow policies (e.g., based on triage), ambulance locations, ambulance redirection, policies for nurse and physician decision making with respect to standing orders for diagnostics, and interdepartmental polices for inpatient admission. These are just some of the Emergency Department factors at the tactical level. None of these factors really exist in isolation from the others. The list does not address a fully integrated, tactical model, including all of the hospital elements associated with the Emergency Department (e.g., diagnostic imaging, labs, surgical suites, inpatient wards, and clinics).

Specific, purpose-built models are targeted towards a single theme within a horizontal layer. Assuming that integrated vertical and horizontal models are not reasonable, it might be reasonable to construct general models for a specific topic, at a specific level, such as creating a general model for physician staffing in the Emergency Department. In fact, a number of research results and efforts fall into this category (e.g., Carter and Lapierre 2001; Gendreau et al. 2006; Jones et al. 2008). Topics such as physician staffing are not dependent upon any operational characteristics associated with the emergency flow and are more suitable for general models.

It is possible that the real-world characteristics associated with Emergency Departments will make even a general, specific purpose model almost impossible for application. Consider the following potential reasons for lack of relevance of a general model:

1. There are significant differences between situations where there is one hospital in a smaller community versus situations where two or more hospitals are in the same community working together. Two hospitals working together can create specific ambulance flows related to specialties, as well as flows when one hospital becomes busier than the other. For example, one hospital may specialize in trauma and pediatrics versus another hospital that may specialize in cardiac care. The two hospitals might also want to model the interaction between the two hospitals when one becomes congested or has diminished capacity. Multiple hospitals also imply the ability to redirect and transfer patients from one to another, as part of the normal flows and treatment plans. While some of these characteristics might be possible to model as part of the patient arrival patterns, some of the issues impact the flow and disposition options.
2. The implications of the size and capacity of Emergency Departments can be significant. For example, there are extreme differences when resources are relatively few (e.g., only one physician on duty) and where sufficient resources exist to absorb perturbations and can maintain other patient flows. In addition, small, rural hospitals create many challenges as they are faced with very low volumes and fixed, minimum, and in some cases indivisible supply characteristics associated with the professional caregivers.

3. Teaching and research hospitals have different processes and resources compared to community, nonteaching hospitals and models for one would not be appropriate for the other. Interns and medical students constitute resources that are not available to other hospitals to the same degree. Furthermore, attending physician time is used differently as students and learners are monitored, instructed, and debriefed by the attending physician.
4. The range of services needed within the population base served by the Emergency Department can be homogeneous with low variety or heterogeneous with high variety. For example, an inner city hospital surrounded by low-income housing and close to high-volume traffic patterns will likely have a different demand pattern compared to a small, community hospital close to a skiing area with a population dominated by retirees. The situations can also be complicated when two or more hospitals work together to stream patient flows based on presenting symptoms, allowing for hospital-based specialization. In low-variety situations where generalizations are possible, it is feasible to use simpler models and more robust assumptions which address a greater portion of the patient pool. For example, the percentage of pediatric patients might not be significant and modeling this age group as a separate factor might be unnecessary. These types of variations can create difficulty when analyzing a single hospital using a general model, or when comparing two institutions.
5. There are differences between publicly and commercially funded healthcare systems. Public health systems can have a greater number of imposed policies and budget constraints. In addition, commercial enterprises can apply more traditional business models with fewer constraints when it comes to justifying resources and acquiring them.

These domain-centric observations suggest that instead of monolithic and "generic" models, a collection of models, each with limited scope and purpose, might be required for each "specific purpose" type of analysis. For example, a model designed to address location, size, and emergency service specialties offered by an Emergency Department might need model variants recognizing the possible implications of sister facilities, public versus private funding, high- versus low-mix population bases, and geospatial factors.

The contextual or scoping differences are relatively easy to identify in the literature where the researchers are stating their assumptions and explaining why their models may or may not fit the specific hospital situation motivating them. We will return to these domain-centric ideas occasionally in the chapter to explain modeling issues which may or may not be relevant in a given situation. In Sect. 5, we will present a number of possibly relevant constraints and modeling issues from both the reviewed literature and a study conducted at St. Mary's General Hospital, a medium-sized, community, nonteaching hospital which shares the population base with another hospital.

It is important to remember that the intent of this chapter is to highlight possible constraints and issues that might be required, and to make both researcher and practitioner aware of the possible requirements, not to mandate their inclusion in all models and in all studies. In some cases, a particular aspect will be important and will bias the results if not included; in other cases, the same concept would be unnecessary. For example, a researcher may decide to exclude process characteristics of the highest severity cases in an aggregate analysis because the severe cases are relatively few and this exclusion might make sense depending on the research question being addressed; however, if the research is probing why the Emergency Department stalls occasionally and there is a sudden outbreak of unduly long wait times, it is important to include any events which might occupy an unusual number of resources for an extended time.

## 4 Literature Review: Scope, Methods, Data, and Results

To understand the Emergency Department challenges in context, Hoffman (2006) is recommended as it provides an excellent historical overview of the US Emergency Department evolution. Hoffman notes: “emergency rooms have been described as ‘in crisis’ since the late 1950s. The nature and intensity of the crisis, however, has changed over time.” For example, the long-term issues with nonurgent care, overcrowding, and longer than desired wait times are discussed. Even though the specifics are based on the US experience, it appears that most of the issues are global and that many regions of the world have experienced a similar journey. It is important to realize that there are many possible differences in the way that Emergency Department services can be delivered. For example, Shingo (2010) points out the differences between the traditional emergency service model in Japan and the North American model which has been recently introduced in Japan.

Although there have been a number of heuristic approximations, optimization formulations, and simulations conducted during the last seven decades, the Emergency Department has remained a place of perplexing complexity, mystery, and challenge. It is an intricate and difficult problem for healthcare providers. Overall efficiency in the Emergency Department remains problematic (e.g., Taylor 2006; CIHI 2011; NHS 2011; MHNZ 2011; GAO 2009; Jayaprakash et al. 2009; Wilper et al. 2008). There are many expectations, goals, and policies set by governing bodies for Emergency Departments, some of which have been retracted or altered after deployment (e.g., 4 h rule in UK—Fletcher et al. 2007; Mayhew and Smith 2008; Topping and Campbell 2010; Bell 2011). There have been many discussions of possible solutions and remedies, but these have not yielded any generic solutions or strategies that address the widespread concerns about efficiency and effectiveness.

Over the last half century, there have been a number of review articles summarizing the various applications of operations research in healthcare, including the Emergency Department (e.g., recent reviews can be found in Fomundam and Herrmann 2007; van Sambeek et al. 2010; Rais and Viana 2010; Turner et al. 2010; Wiler et al. 2011). There have also been texts and chapters in which the quantitative analysis of healthcare is reviewed and summarized (e.g., Hall 2006; Gaur 2008). This is just a brief list of reviews; there are many others to be found in the literature.

We concluded our review by searching the operations research and healthcare literature including the applied methodology components of operations research, management sciences, operations management, industrial engineering, and applied mathematics, and the application research arising from the medical fields such as emergency physician education and practice, emergency medicine management, emergency services, nursing education and practice, and general healthcare management.[1]

Inclusion and exclusion are additional factors when considering a review. Reviews such as van Sambeek et al. (2010) provide a clear description about the criteria they used. This type of clarity is critical because they provide some insights about the literature with respect to Emergency, but explicitly exclude any articles with a primary focus on predicting demand or length of stay, two key topics of Emergency Department research.

Any researcher probing the topic of modeling Emergency Departments should be aware of these issues and limitations. For example, this review and chapter is limited to what information can be found in English language sources. We did attempt to cover both methodology and application literature using the variety of descriptors and terms noted above.

What is clear from reviewing the literature on modeling Emergency Departments is that while there has been a long history of quantitative analysis of the challenges facing Emergency Departments, including mathematical programming, statistical analysis, queuing theory, and simulations, the body of work is not large when compared to the similar types of research performed on manufacturing topics, other service situations, or healthcare aspects such as the surgical suite.

As with most applied topics, there are four dimensions to consider: scope or topics addressed, data used, the methodologies employed, and results obtained. While in this section we will overview the methodologies used to study the Emergency Department, more specific details about particular methodologies can be found in other chapters. We will also provide a review of the various scopes, data used, and results.

## *4.1 Scope*

The most common theme in the Emergency Department literature seems to focus on wait times and related issues such as crowding, resource utilization, capacity planning, room shortages, and triage flows. Topics such as batching of work by

[1] Terminology is also a factor in performing literature searches on this topic. For example, various terms are used to describe the functionality and include emergency room, Emergency Department, trauma unit, medical assessment unit, and accident and emergency unit. Spelling is an additional issue with terms such as queuing and queueing. Phrasing is also important as queuing theory versus queuing model will provide different results. These are just the English language semantics and do not address the challenges associated with research disseminated in different languages.

physicians has been studied, but it is relatively rare (e.g., Dobson et al. 2012). Secondary topics include such areas as physician and nurse scheduling (e.g., Beaulieu et al. 2000; Gascon et al. 2000; Carter and Lapierre 2001; Pinedo 2009; Yang et al. 2009), and arrival pattern analysis and forecasting (e.g., Hoot et al. 2008; Jones et al. 2008). Tertiary topics look at the interfaces with other parts of the hospital, for example, inpatient admissions (e.g., Bagust et al. 1999). Early research such as Newell (1954) and Handyside and Morris (1967) focused on bed usage in Emergency, and Newell (1965) and Bailey (1956) addressed Emergency arrival patterns.

Over time, the components and targets of the analyses have varied in response to the type of crisis facing the hospitals. For example, the concept of triage was introduced to hospitals from the battlefield in the early 1960s (e.g., Weinerman and Edwards 1964; Weinerman et al. 1965) and was mentioned as a topic of study and investigation in the 1970s to determine how effective it was (Hamilton 1974). Hoffman (2006) notes that triage was not accepted by all and there were debates as to its merit. It is now common in hospitals, and it is also common for almost all studies, to include various triage levels. The benefit of triage is no longer debated in many areas but is still relatively new to some jurisdictions (e.g., Andersson et al. 2006). How triage might work and who does the triage is still debated (e.g., Holroyd et al. 2007; Ting et al. 1991; Han et al. 2010), but most researchers appreciate the need to include the triage levels and their impact on resource utilization and treatment plans. Boarding (admitting ED patients when no inpatient rooms are available, and similar to banks in flow lines) and overcrowding caused by slow movement of admitted patients towards (i.e., blocking) have been the subject of research for an extended time as well (e.g., Hamilton 1974; Lynn and Kellermann 1991). The arrival demographics and the impact of nonurgent care patients have also been the subject of studies and commentary since the 1960s (e.g., Coleman and Errera 1963; Hamilton 1974).

The tertiary topics also have a long history. In some cases, it is possible to include the Emergency Department component as a black box via an approximation or input distribution. Alternatively, the Emergency Department is modeled at some level of detail, and this information is used to guide planning of the tertiary area. The interaction of emergency patients with other parts of the hospital such as elective or scheduled surgery has been a major focus since the early 1960s. An early example of the first type of tertiary inclusion was provided by Fetter and Thompson (1965). They describe a simulation model taking into account scheduled and unscheduled events on beds and operating rooms. Hannan (1975) is an example of the second type of tertiary model. Hannan simulated a holding or observation area as part of the Emergency Department to act as a buffer area to reduce premature admissions to inpatient wards. Another tertiary interaction is the general impact of emergency admissions on the hospital resources. Kolesar (1970) presented an early queuing analysis of scheduled and unscheduled admissions.

## *4.2 Methods*

A variety of methods have been used to study aspects of Emergency Departments including queuing (e.g., Green 2006; Soni and Saxena 2011; Dobson et al. 2012) and simulation analyses (e.g. Jun et al. 1999; Jacobson et al. 2006; Seila and Brailsford 2009; Paul et al. 2010). While there are examples of dynamic programming and queuing theory applied in the area, the majority of research publications focus on simulation or queuing with some empirical component. The inclusion of empirical data clearly enhances the understanding of the problem and provides a base for validation, as suggested by Fisher (2007). Challenges relating to simulation in the healthcare field were described in detail by Carter and Blake (2005). Recent simulation efforts such as Duguay and Chetouane (2007) and Facchin et al. (2010) illustrate two general categories of simulation efforts. Duguay and Chetouane’s purpose was to reduce the wait time at a specific hospital, while Facchin et al. were trying to create a generic, flexible simulation base with which to “correctly describe many existing emergency services” through simple adaptations of the base tool. Duguay and Chetouane do an excellent job in explaining how their modeling effort relates to prior work such as Saunders et al. (1989), their data collection and analysis techniques, the model, limitations, and results. Many simulation-based articles gloss over many of these aspects, focusing on the model and results. It is important for a reader to understand what the model is actually modeling, and papers such as Duguay and Chetouane achieve this clarity, as does Blake et al. (1996).

Facchin et al. also present a good discussion of the existing literature and issues found in simulation efforts. In discussing general simulation models, it is difficult to make supported claims and understand the limitations of the model. Facchin et al. were able to deploy the model at several regional hospitals and to provide value; however, it is a recently developed tool and it is not clear how useful their model will be at different levels of decision making, with various scopes or target purposes, or in different contexts beyond their current, regional experience.

The ability to claim successful adaptation of a general tool to a specific case has always been a tricky topic; you can adapt the tool with less than 10% of the original effort, or with 90% of the original effort, or with more than the original effort. Simulation methodologies have also been applied in very innovative ways. For example, Hoot et al. (2008) use a sophisticated simulation approach for the real-time forecasting of overcrowding in Emergency Departments. The specifics of the technique might need to be adjusted for different hospital situations, but the general approach is intriguing and appears to have merit.

Although the vast majority of research has been conducted using traditional discrete event simulation and queuing methods, other methods have also been explored. For example, Lane et al. (2000) use systems dynamics, Jones and Evans (2008) agent-based models, Yeh and Lin (2007) genetic algorithms, and Malakooti et al. (2004) group technology.

## 4.3 Data

Models do not exist in a vacuum. They are created to address a specific question or issue, or are created to assist a line of inquiry. They exist to help make some decision. The data used in a model will depend on the question or line of inquiry. In reviewing the literature, we found a wide range of data granularity used in the research efforts. At the highest level of abstraction, the data will not have per patient records but will be summarized by fitted distributions for arrivals, service, and disposition (e.g., discharged home or admitted to an inpatient ward). This level is adequate for getting to ballpark estimates for regional distribution of resources, specialized services, impact of nonurgent care clinics, approximate number of physicians, nurses, beds, and so forth, with the accuracy possibly within an order of magnitude.

The next level of granularity addressed topics such as wait times, lengths of stay, admissions to inpatient wards, flows to different areas such as Acute and Subacute, ambulance diversion, physician and nurse counts, impact of physician assistants, nurse practitioners, beds, bays, and stretchers, and auxiliary processes such as diagnostic imaging and blood work. Most of the reviewed literature used rather simple models and constructs. For example, the simplest "black box" data does not have any detailed patient or treatment characteristics. As the models become more detailed, the triage level and some measures of average processing or wait time based on triage level are included. Data driving the resource performance can also be simple with standards or estimates for physician, nurse, and test times, with and without shifts. The models using this type of data assume many detailed differences are not significant for the analysis being conducted. For example, this level of data assumes that the triage level is a sufficient predictor of flow, that there are no significant differences between physicians, that physician productivity is the same for each hour of the day, and so forth. While it might be possible to obtain aggregate results over an appropriate horizon, it is difficult for this level of modeling to explain what has been happening or might happen at finer resolutions.

Some of the research used more detailed data about each patient visit. For example, information about specific tests might be involved, and the patient is tracked through the process in order to start creating a more detailed model (e.g., Samaha et al. 2003). In most cases, additional information about the patient is not used (e.g., past medical history, age, gender, and drug usage), nor are specific nurses or physicians identified. The patients are still considered to be abstract entities at a very high level. This type of data modeling can be seen in many models as the researchers look at the steps associated with triage, diagnostic assessment by nurses, physicians, tests ordered, and final dispositions (e.g., Saunders et al. 1989; Sinreich and Marmor 2005).

Saunders et al. (1989) and Sinreich and Marmor (2005) are two examples that may extend the model the greatest at this level of data construct to include many small processing steps within and around the Emergency Department; however, there are still many assumptions in both studies. For example, there are no detailed personalizations at the patient or physician experience levels. The patient level

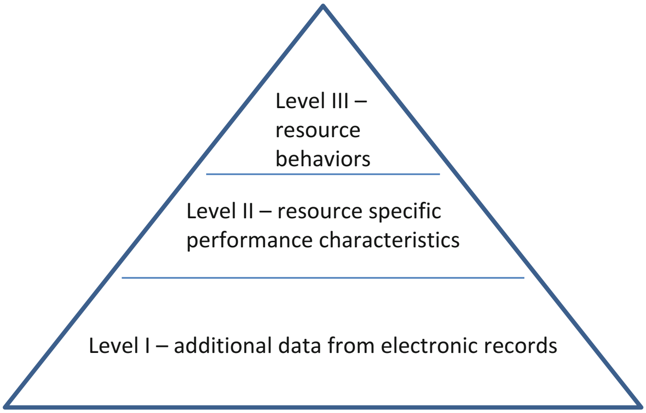

Level III –
resource
behaviors
Level II – resource specific
performance characteristics
Level I – additional data from electronic records

**Fig. 13.1** Extended data modeling levels

modeling in Saunders et al. was based on triage levels, and Sinreich and Marmor distilled eight different types of patients (derived from detailed patient information including age, gender, and presenting complaints). Patients were assumed to be generic in both studies, as were the physicians. Sinreich and Marmor had the physicians grouped by assessment and by treatment tasks, but physicians within these groupings were generic. Another model, which is reasonably detailed and also used processing information on a per patient basis, can be found in Connelly and Bair (2004).

The research efforts noted above either implicitly or explicitly assume that it will not matter what physician is on duty, what the experience level is the physician, if the patient is overweight, if the patient is geriatric and requires additional resources, what happens during a shift change, or if the patient has to be isolated because of possible infection concerns. If it is the goal of the model to provide an idealistic, general baseline, then these types of assumptions are quite acceptable. It is also possible that these differences are not important and that the limited model can provide realistic insights; however, it is incumbent upon the researcher to justify the latter case and provide evidence that the assumptions are valid.

In our own research, we encountered three additional categories of data to consider. They may or may not matter depending on the Emergency Department situation being modeled, and they are context dependent (e.g., similar to the data and heuristics discussed in McKay 1992). The three levels are illustrated in Fig. 13.1.

*Level I*—The first level of additional information includes more information from the patient and hospital records. For example, it is common to find presenting symptoms, patient age, whether or not the patient has a family physician, and diagnosis. If the department has a significant number of elderly and pediatric patients, modeling this difference may be important. An elderly patient may imply more tests, lengthier diagnosis and assessment time because of preexisting conditions, and other challenges such as dementia. Pediatric patients may imply other types of specialists, equipment, tests, and processes.

We have seen some research that uses secondary data, such as presenting symptoms, to help guide the analysis (e.g., Sinreich and Marmor 2005). With respect to other personal data about the patient, we have seen some examples that used the patient's age for routing control (e.g., to a pediatric area, Chan et al. 1997); however, we have not seen any research that uses information about the patient's age to predict the types of diagnostics required or specialists who might be required for consultation on a case-by-case basis as part of the physician assessment. At a higher level, Asaro et al. (2007) statistically analyzed a large number of input and output factors (including a number of patient characteristics) with respect to a regression model for overall throughput times, but note the limitations of the data available in the electronic records to which they had access. They analyzed over 176,000 records and were able to create a regression model that explained approximately 25% of the variance using a combination of system, time, and patient factors. Innes et al. (2005) did an in-depth analysis of possible factors affecting emergency physician workload and included a large number of patient and case factors. They looked at the procedure required, triage level, arrival by ambulance, Glasgow Coma Scale score, age, any comorbidity, and number of prior visits. Their model explained approximately 31% of the variance in the physician's time per patient visit.

*Level II*—The second level of additional data addresses unique performance characteristics about resources. For example, specific information about physicians can help decode or understand what is going on and what might happen in the future. We are not aware of any published work on Emergency Departments which include specifics on individual physician productivity or skill level based on actual physician analysis.

Gunal and Pidd (2006) modeled junior and senior physicians in a generic fashion and used this information to control the amount of multitasking a physician will do and the time a physician will take. This was the only research we have encountered in the literature that included some experimental control of physician levels and productivity. Random physician differentiation was done by Jones et al. (2008). They varied the patients per hour capacity for a physician using a Poisson distribution. While not working with individual physician data, Crane and Noon (2011) provided a discussion and modeling concept for how physicians spread their time across a patient case and suggested how this can be incorporated in an analysis and how it could improve the result validity. The modeling of emergency nurse resources at a detailed level is also limited. The Komashie and Mousavi (2005) study was the only one we have seen that included different levels of nurses with different processing times.

*Level III*—The third level of additional data requires observation periods and longer research efforts. This level of information pertains to behaviors, and what people actually do in the Emergency Department during their shifts, and what may or may not happen with regard to unusual events and their implications. With the exception of the work noted above by Gunal and Pidd (2006), Jones et al. (2008), and Crane and Noon (2011), the inclusion of any individual work patterns is rare. None of the reviewed quantitative research included any extensive discussion or modeling of individual work behavior or work patterns. The Gunal and Pidd research is noteworthy in that they used results from the work of Chisholm et al. (2000) on physician multitasking and interruptions to guide their modeling effort. Laxmisan et al. (2007) performed an ethnographic analysis on physician decision making, and they discussed multitasking, interrupts, and patient handoff. Levin et al. (2006, 2007) also studied shift changes and the timing impact of the change on physician workload.

Other research efforts that have studied the Emergency Department and that may provide insights for quantitative modeling at a deeper level include works such as:

1. Mentis et al. (2010) and Mentis (2010) explored emotions in the emergency room with respect to electronic record keeping and possible side effects to safety and efficiency.
2. Andersson et al. (2006) looked at nurse decision making at triage and discussed possible errors and inconsistencies in the triage process. This can be important if a researcher wants to consider a realistic situation where triage errors occur and need to be handled in the subsequent flows.
3. Ullman et al. (1975) did an extensive demographic analysis of emergency patient characteristics and implications on hospital resources. Their research can provide insights about nonrandom patterns in the arrival stream and how this can be related to evolving community populations.
4. Coiera and Tombs (1998), Fairbanks et al. (2007), and Woloshynowych et al. (2007) probed communication behavior, linkages, and patterns, all of which can impact detailed flows and interaction points between resources.

In summary, almost all of the current and past literature uses generic data at the aggregate or probability distribution level. Models and analyses using additional data from the individual records, data about individual resource performance, and data about the resource behavior appear to be rare. It is important to realize that the additional data may or may not be important in a specific situation, and it is important to know the limitations of any model which excludes relevant data. For example, Downey and Zun (2007) perform a detailed analysis on an Emergency Department, including many factors such as age, gender, presenting illness, as well as linkages to other, supporting areas of the hospital. They found that many factors beyond the department itself should be included in the analysis as the turnaround time from other units (e.g., psychiatry, medicine, and laboratory) was the largest contributor to extended wait times.

## *4.4 Implementation in Practice*

The literature review of Sect. 4 was based on approximately 200 sources, not all of which have been cited. We found that research focusing on the internal flows of an Emergency Department is limited. We did not find studies in the Emergency Department process context that focused on methods, nor did we find a substantial number of generic models or modeling frameworks at the process level for this application domain. The generic models in the literature do not appear to have been adopted beyond the originating research group. It appears that change and implementation in the healthcare domain are lengthy activities, and it is possible that positive results will be forthcoming; however, this delay adds to the difficulty of validating claims of generality and value.

In the applied sense, it is clear that discrete event simulation can be used to capture the core processes in a specific Emergency Department and provide some form of claimed value, either theoretical or of practical value to the studied institution. Some but not many of the simulation-based studies have seen their results used in an actual setting and have been able to claim substantial impact (e.g., Blake et al. 1996). There have been some studies using queuing (e.g., Green et al. 2006), but again with few actual implementations. If the applied research does not result with changes implemented by the hospital, there is always the question about the validity of the model and the assumptions underlying the model and validation. It is hard to know with certainty if the "necessary and sufficient" real-world constraints and factors have been included for real-world adoption.

Even in cases when changes have been implemented, the claims are not always transparent or self-evident. For example, were the research results (observations or recommendations) previously unknown to the hospital? Did the research confirm and quantify what the hospital expected? Did the research provide support for moving forward? Were the what-if scenarios suggested by the hospital or were the solutions and recommendations the result of academic and theoretical inspiration? Did the effort required to document and build the model result in the changes, or were the changes driven by the actual analysis and results? Were any recommendations implemented and sustained with quantitative evidence of impact?

Several applied publications have appeared in recent years. For example, Crane and Noon (2011) and Shiver and Eitel (2009) provided guidance for how to improve Emergency Departments using a combination of qualitative and quantitative methods. Books such as these can serve as a starting point for practitioners and academics who wish to understand some of the general aspects of the problem space and how some of the techniques can be applied. The material is generally case study style and the data modeling typically does not go beyond levels I and II, but the texts do describe the value of good analysis, mapping and understanding the problem, understanding variance, and how to use the methods in an applied way.

## 5 A Field Study

In 2011, an extensive field study was conducted of the Emergency Department at St. Mary's General Hospital, in Kitchener, Ontario, Canada. Approximately 1,000 person hours were involved in the study and analysis, and of these, 500 h were spent in the department observing day-to-day flows and processes. Roughly three-quarters of this effort was at the senior, consulting analyst level, the rest at the intermediate, postgraduate level. Detailed time studies were not done with the exception of key

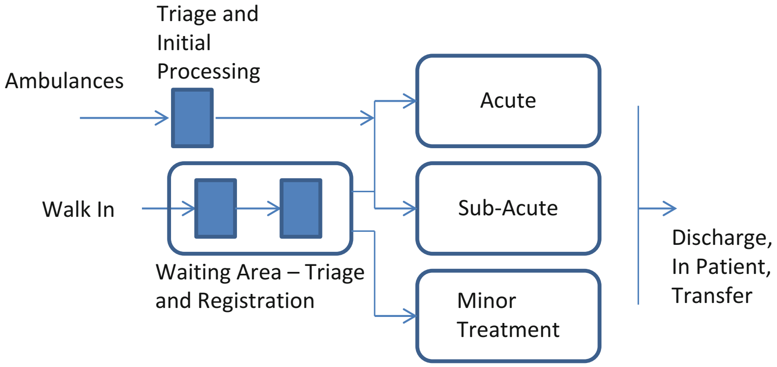

Triage and Initial Processing
Ambulances
Walk In
Waiting Area – Triage and Registration
Acute
Sub-Acute
Minor Treatment
Discharge, In Patient, Transfer

**Fig. 13.2** General flow of St. Mary's Emergency Department

narratives for unusual flows. The majority of time was spent on understanding the dynamic flows, how this related to the data captured by the hospital, and how the data could be analyzed. The hospital is a medium-sized community, nonteaching hospital with a 25-bed Emergency Department which sees over 45,000 patients each year. The hospital works with other regional healthcare facilities to provide specialized services and ambulance routing. Another community hospital provides a more general service, and St. Mary's is considered to have a less-varied patient portfolio (e.g., majority of trauma, mental, and pediatric are routed elsewhere). As such, the modeling and contextual constructs at St. Mary's are likely to be less challenging compared to other situations and are possibly the best case scenario for simplicity. A combination of field and quantitative methods was used as part of the study (the general field methods are described in McKay 2011).

Figure 13.2 shows the general flow and structure of St. Mary's Emergency Department.

The detailed flow through Acute and Subacute areas is shown in Fig. 13.3. There are multiple exit points and additional loops and streams associated with orders (lab and imaging), and specialists. There are two points in the flow where blocking or flow limits exist that constrain the number of patients in process (e.g., the physicians and nurses constrain the number of patients being seen at the same time).

The department has been recently renovated, is considered "modern," uses triage and fast tracking to stream noncritical patients to a minor care area, deploys nurse practitioners and a physician assistant, and has had a number of improvement activities in the last few years. There are also clerks used for registration, organizing movement into wards, and facilitating communication and flow throughout the area. Porters perform the room preparation, assist with patient care, and are responsible for moving patients to the imaging department. The Emergency Department also uses electronic tracking and flow monitors in each of the treatment areas. They also recently introduced a "lean" continuous improvement process throughout the hospital, including Emergency.

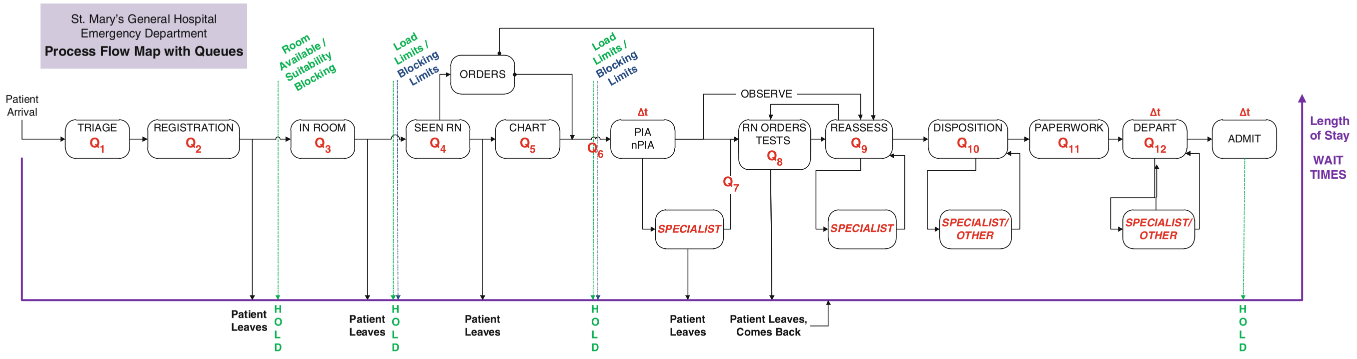


**Fig. 13.3** Detailed flow—St. Mary's Emergency Department

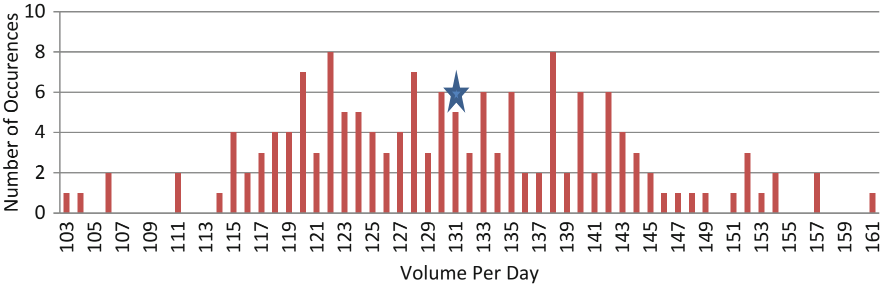


**Fig. 13.4** Patient volume distribution (*star* indicates the mean and median—131 per day)

Historically, there seemed to be contradictory variation in patient flow and wait times. There was a great deal of speculation about cause and effect relationships, what to change, and how improvement would be obtained. Furthermore, a number of early gains associated with recent changes seemed to disappear over time.

Data from December 1, 2010, to March 31, 2011, was analyzed as a Flu Season. Data from April 1 through August 31, 2011, was used as a Non-Flu Season. For the examples presented in this chapter, the Non-Flu Season data was used unless explicitly noted. The insights are grouped into the following categories:

- Arrival patterns
- Age dependencies
- Physician dependencies
- Wait time, flow analysis
- “Out-of-the-ordinary” impacts
- General modeling

## *5.1 Arrival Patterns*

Sinreich and Marmor (2005) presented one of the best discussions and analysis of arrival patterns. They developed a method for arriving at a pattern that seemed to fit the mean hourly pattern at the hospital they were working with. It is not clear how the fit is by-hour or by-day, or what the standard deviation is by-hour, by-day for each point. They used 24 months of data and found weekly, daily, and hourly effects.

At St. Mary’s, we restricted the analysis to cases since December 2010 because the data electronically captured was more extensive than prior to that time. We broke the time period into two distinct seasons, Flu and Non-Flu. Just as others have noted in studies, there were differences at the month, week, day, and hourly levels. At St. Mary’s, the mean and median arrival of patients per day for the 6 months, Non-Flu sample period was 131 (highlighted with a star). Unfortunately, this “average” day only occurs 3% of the time, and the frequency distribution is presented in Fig. 13.4.

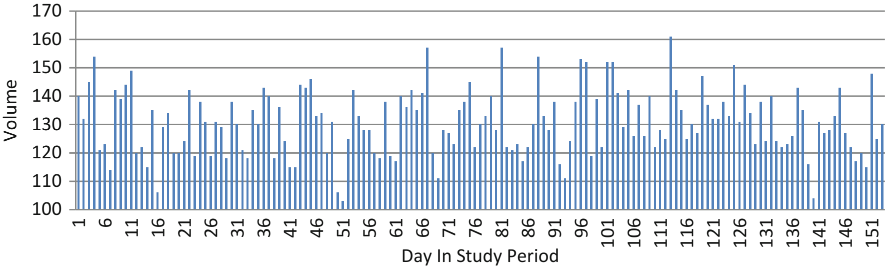


**Fig. 13.5** Patient volume by day

The volumes were distributed across the 153 days as shown in Fig. 13.5. Mondays are considered by the hospital staff as always being the busiest day of the week. We found that the total, aggregate volume for Mondays was indeed the highest when compared to the other days; however, analyzing the data week by week, Mondays were the busiest day only 33% of the time. Figure 13.5 points out another type of pattern, sequences of days below and above the mean of 131, suggesting that each day may not be independent of what happened the previous day, and that some form of underlying, short term trend might exist.

Another interesting pattern was that the patient flow times were shorter on the weekend compared to the weekdays. Volumes are relatively high on Saturdays and Sundays, so this was not the causal factor. There has been some speculation about the causes, but nothing definitive has yet been identified. For example, some tests are not done on the weekends and therefore patients are scheduled to return on Monday, speeding up the weekend flow while contributing to the higher congestion on Monday.

Many studies have pointed out the importance of modeling the hourly arrival times. We also found this a challenge at St. Mary's. The St. Mary's electronic tracking system starts with the triage time, and not when the patient actually arrives. Thus, the triage time gives a slightly distorted picture of arrival patterns, further confounded when two nurses, instead of just one, are triaging during busy periods. The electronic system does not measure the time a patient is waiting for triage. We were able to obtain the true arrival times for half of the time period via a different method, but not for the full analysis, and this then required a merge and match procedure to properly synchronize the data. If the triage data is being used to model the patient arrivals, then the analysis and discussion would need to start after this point (e.g., the pattern arriving at registration) and should not be used to model arrivals into triage or the triage process itself. The true arrival time is possibly more important to a researcher than a practitioner. If the hospital and governing body start the wait time and length of stay measurement clocks at the point of triage, they might not be as concerned about patients waiting pre-triage. A researcher studying the whole system is likely to be concerned about pre-triage as well.

A major weakness in our analysis was the inability to assess complexity of an incoming patient. We had the usual acuity, presenting complaint data, diagnosis,

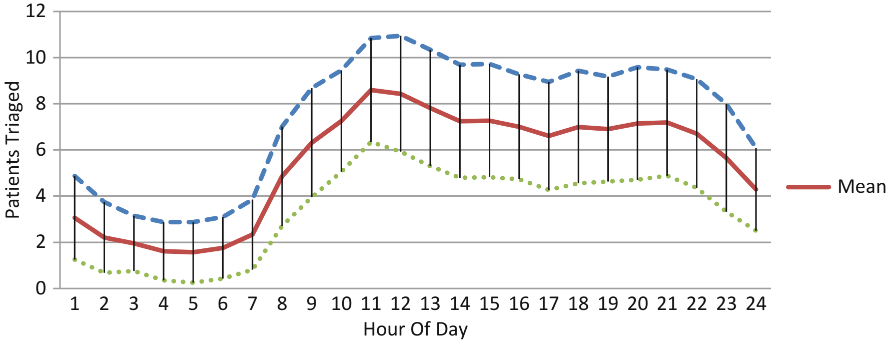


**Fig. 13.6** One standard deviation confidence interval for hourly arrivals

age, and medical disposition, but we did not have an indication of the patient’s complete condition. For example, you can have two patients with the same acuity and presenting complaint, but one might take far fewer resources to process and a shorter elapsed time compared to another patient. Specifically, we did not know from the data if the patient required special isolation handling due to possible infections (e.g., a high-risk patient), if the patient would be problematic to get information from (e.g., elderly with dementia), if the patient would need additional resources for any task (e.g., because of a high body mass index), or if the patient was part of a police investigation, and so forth.

St. Mary’s has the Emergency Department split into areas for Acute, Subacute, and Minor Treatment (e.g., fast tracking using nurse practitioners). The highest triage levels of acuity (I and II) almost always go to the Acute area, and the nonurgent (IV and V) go to Minor Treatment when it is open. Level III cases are the most numerous and might be routed to any of the three areas based on the presenting complaint and the expertise of the triage nurse. Electronic triage has just been implemented at St. Mary’s, and we will be investigating the use of the electronic data for predictive purposes on a case-by-case basis. During the study period there were 1,100 different complaint codes recorded and not all were unique. We do not know yet if this path of reasoning will be fruitful, but it looks like a potential path to pursue.

The arrival patterns are often modeled at the hourly level, and many papers present a graph showing the hourly means (although most papers do not share variance information). The hourly mean and the upper and lower bounds based on a standard deviation for the St. Mary’s Non-Flu Season are shown in Fig. 13.6.

We found that there were volume and “acuity mix” differences by hour and by day. For example, Mondays and Thursdays usually presented different volumes and acuity patterns which are possibly related to weekends and Wednesday family physician office closures which are common in the region. The acuity mix and arrival patterns were also found to be different in what is considered Flu Season versus Non-Flu Season.

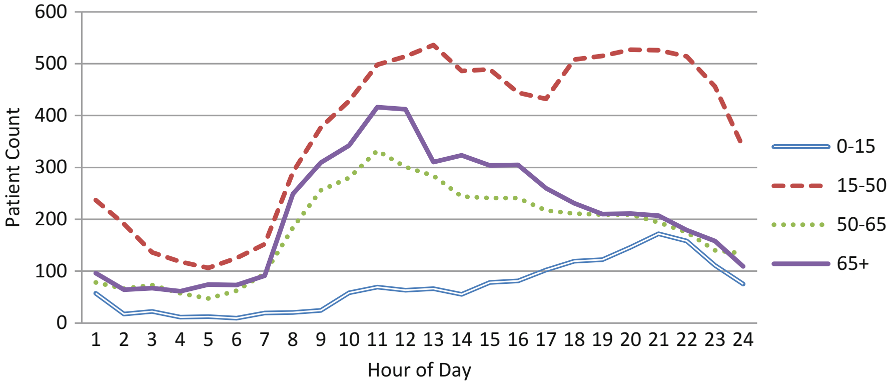


**Fig. 13.7** Total patient arrivals by age grouping per hour

Most studies we reviewed assumed that the arrivals are unique and are independent of each other. We did not find that this was the case at St. Mary's. For example, during the Non-Flu period, approximately 6–7% of the arrivals were at the department less than two days earlier. They returned for a number of reasons and the numbers varied by physician. In total, approximately 20% of the arrivals were associated with patients who visited the department two or more times during the Non-Flu period. The returning patient phenomenon is a topic being looked at by many hospitals, and it is also a topic for future research. It will require more medical analysis to understand the reasons for returning as some of the patients are returning for secondary treatments, tests, or instructions and not for new issues. For example, we know that deep vein thrombosis (DVT) follow-ups are scheduled Monday through Friday and are part of the emergency arrival pattern. At St. Mary's, the 20% is a significant portion of the total department arrivals, and this will need to be accounted for when modeling arrivals or when analyzing past history.

## *5.2 Age Dependencies*

In some studies, the age of the patient has been used when there are special pediatric areas. Figure 13.7 shows the hourly arrival pattern by age segment for the 6 months period. There were also some minor differences per day of week, but not significant. Four age segments were used and the segmentation was based on the normal hospital grouping whenever age was discussed: 0–15 capturing the young and juvenile, 15–50 for the normal young adult to middle-aged grouping, 50–65 for the middle-aged, but not yet senior citizen, and the 65+ category for the elderly.

We found that the age of the patient affected the acuity mix (using the Canadian Triage Assessment Scale—CTAS—http://caep.ca/resources/ctas), care flows, and diagnoses which followed. Figure 13.8 shows the age distribution across the acuity levels.

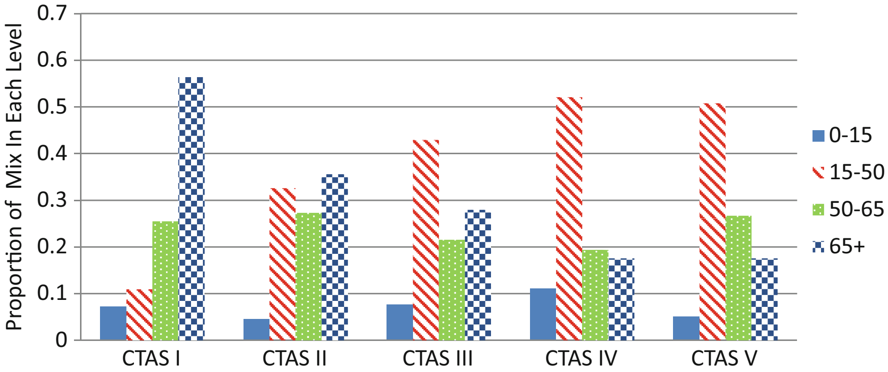


Fig. 13.8 Distribution of each age group in each acuity level

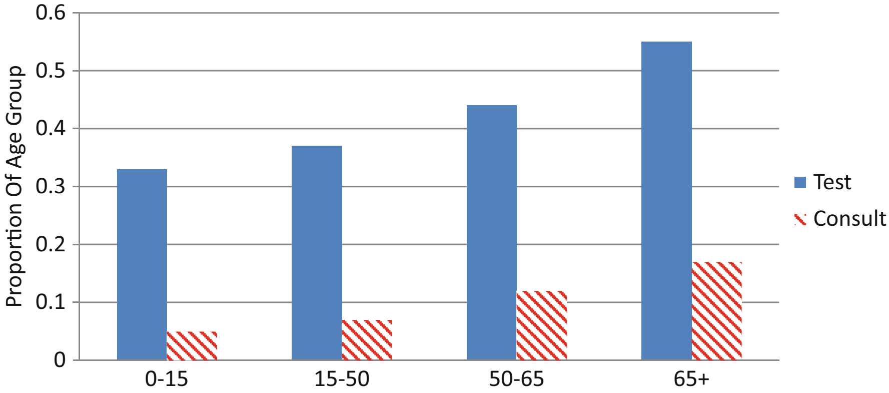


Fig. 13.9 Proportion of patients requiring a specialist or an imaging test

There are also differences by age regarding presenting complaints and final disposition. Unfortunately, everything is a mixture, and it is not possible to say that X is always related to Y when it comes to age. There are patterns, and the general patterns would need to be modeled to ensure an appropriate mixture. This is highlighted in Fig. 13.9 which shows two of the major activities that affect the length of stay in the department: specialist consultation (during diagnosis) and tests associated with diagnostic imaging.

On average, patients with an imaging test will spend an additional 2 h in the department, and patients requiring a specialist consultation will have their length of stay extended by approximately 4 h. When both are needed by a patient, the length of stay will be extended by approximately 5–6 h. It is clear that age is a significant contributor and that the acuity level is not sufficient for modeling the type of services a patient will need. The relationship between tests, consultant involvement, and length of stay speaks to the potential of flagging patients requiring both testing and consultant involvement as the highest priorities to have testing done as soon as possible.

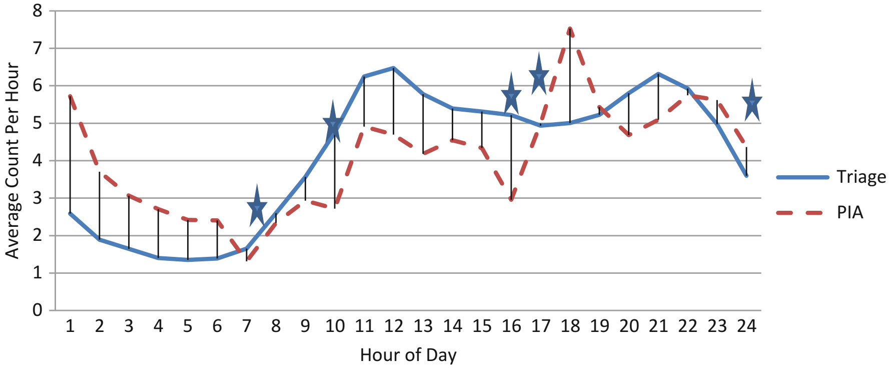


**Fig. 13.10** Triage and physician assessment (*stars* indicate start times of physician shifts)

## 5.3 *Physician Dependencies*

During the St. Mary's study, a number of insights were developed about physician behavior and productivity. First, several shift-based patterns were observed and most physicians seemed to follow it. Figure 13.10 shows the triage arrivals per hour and the physician initial assessments per hour. The graph starts at midnight. Physician start times are staggered through the day with start times at midnight, 7:30 AM, 10:00 AM, 4:00 PM, and 5:00 PM. The physician shifts are 8.5 h each. As the graph shows, there were definite shift patterns around the start and end of shifts which are highlighted with stars.

There were observed phases to a physician shift. The first phase deals with inheriting the department at the start of shift and then ramping up to a steady state. The start of shift can range from 10 to 30 min (or more) and see a decrease in patients-per-hour of 25–75%. The physician builds up an active portfolio of patients being seen, and while this number can vary, a typical number is 4–6 active charts at a time (e.g., the multitasking behavior noted by Gunal and Pidd 2006). Towards the end of a shift, the physicians will see any critical patients but will also be selective about which noncritical patients will be seen, leaving certain patients to the next shift. Extra paperwork for admissions and charting is also done during this time. The last part of the shift is used to interact with the incoming physician(s). This is illustrated in Fig. 13.11.

The mean day-shift productivity per physician was approximately 33% less than on the evening and midnight shifts. This was independent of individual physician (individual physician activity was analyzed as part of the study). There are a number of possible factors contributing to this consistent drop in productivity for this particular shift, but more analysis and field data are needed before any causal relationships can be stated with confidence. For example, we know that the elderly patient arrivals peak in the morning and it is possible that the additional time needed for elderly patients contributes to the lower efficiency. We also know that a number

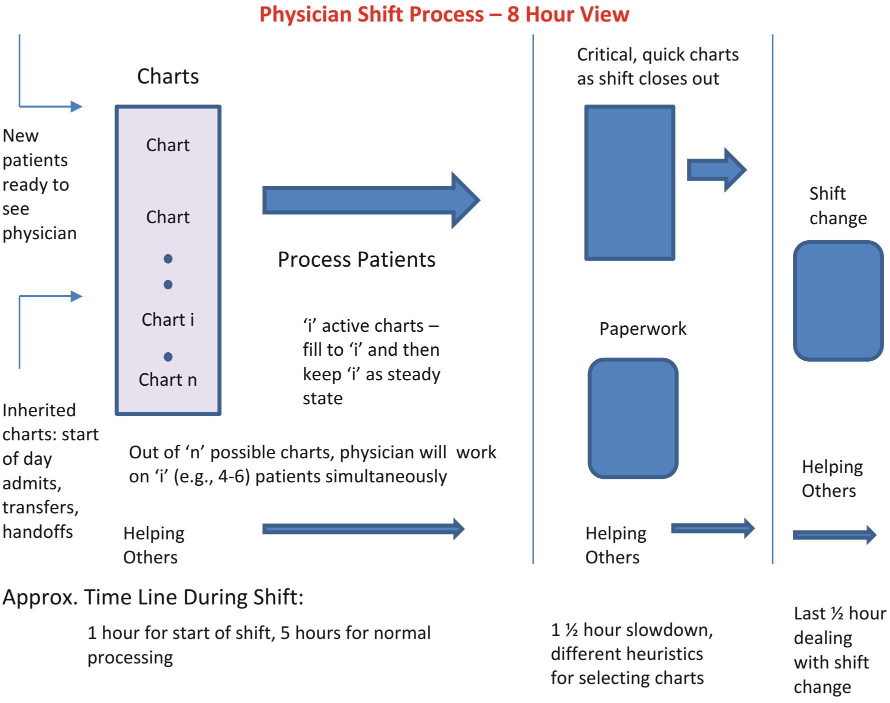

Physician Shift Process – 8 Hour View
Charts
New patients ready to see physician
Chart
Chart
Chart i
Chart n
Inherited charts: start of day admits, transfers, handoffs
Process Patients
‘i’ active charts – fill to ‘i’ and then keep ‘i’ as steady state
Out of ‘n’ possible charts, physician will work on ‘i’ (e.g., 4-6) patients simultaneously
Helping Others
Critical, quick charts as shift closes out
Paperwork
Helping Others
Shift change
Helping Others
Approx. Time Line During Shift:
1 hour for start of shift, 5 hours for normal processing
1 ½ hour slowdown, different heuristics for selecting charts
Last ½ hour dealing with shift change

**Fig. 13.11** Physician shift pattern

of return patients are scheduled for the morning and these could also contribute to the backlog and congestion. At this point, more analysis is necessary to probe possible linkages.

The field study identified different doctor flow patterns related to work flow strategies (e.g., batching of patients, areas focused upon, when different triage levels are attended to). These are important at the detailed level since a physician’s work flow can cause peaks and troughs of work to flow through various parts of the department. For example, if a physician does not cycle through the areas quickly enough, there will be lumpy demands for tests, treatment plans, etc., and this can cause momentary traffic jams of flows, resource shortages, and expediting. To confound the matter further, different work flow patterns and heuristics exist if the physician is working with another physician on the day and night shifts or is the sole physician on the night shift.

During the study period in this department, there was a mixture of medical providers: emergency physicians, medical students, a physician assistant, and nurse practitioners. The NPs can deal with certain types of acuity without requiring an attending physician. Each NP has different qualifications, and these control the types of acuity, presenting complaints, diagnoses, and treatment plans that the NP has independent control over. Depending on the mix, there is little or almost no interaction between the NP and physician. There are times when the NP is in consultation with the physician, and this takes the physician out of service for that time interval.

The medical students, interns, and physician assistant "helpers" turn the physician into multiple servers. The physician is involved with the helpers' patients for the diagnosis review and treatment plan but also sees to their own patients. If the individual physicians are analyzed, the presence of a helper can increase the physician's overall shift rate (i.e., patients assessed per hour) by 20–30%, depending on the number of shifts a helper is present (note: the attending physician is the healthcare provider of record). The time for physician initial assessment is better on these shifts since there are multiple, parallel servers, but the length of stay is longer due to the consultations on every case.

As part of the study, we examined the relationship between physician experience and skill against performance. There were three major physician groupings of experience and skill: junior, intermediate, and senior. We found definite patterns of performance based on skill level.

## *5.4 Wait Time and Flow Analysis*

A number of constraints and relationships were discovered when focusing on the wait times and process flows. The topics have been broken down into area or shift, individual staff, and external factors.

### 5.4.1 Area or Shift

This category of constraints and relationships is similar to a repair shop situation. There are some common tools and resources found in every Emergency Department room at St. Mary's, but there are some resources which are not duplicated. This issue can create a resource conflict and extend the processing times for some patients. There are also some rooms with a specialization (e.g., fractures, isolation), and there are specific heuristics used for room allocation (e.g., which ones are the last to be allocated).

The shift patterns also affect the patient flow times. For example, patients arriving "in room" just before a shift change will possibly face three factors which can affect flow. First, the assessment of noncritical patients might not occur until the new shift (to reduce the transfer of patients among providers). Second, there is a period of time at the start of shift when everyone huddles, gets debriefed, and prepares for the shift. During this time, the "in room" to "seen RN" times will be extended compared to the rest of shift, and any treatment or test processes will also be extended. Third, there might also be regular events during a shift or day such as "bed" meetings to discuss the patients being admitted. Events similar to these decrease the capacity for extended periods of time and become important for modeling. This is especially true whenever small numbers of resources are involved.

The Emergency Department is not always fully staffed due to illness or other factors. This affects the allocation of nurses and other personnel during the short-staffed periods and can affect the patient flow times. Knowing if a time period was short staffed is useful in interpreting performance metrics and building distributions. We did find that it was very rare for the physicians to be suddenly short staffed, caused by a physician calling in without having arranged coverage; however, unexpected short staffing was not rare in the other personnel areas such as nursing or clerks. When the area is short staffed, this creates an “out-of-the-ordinary” condition which often derails one or more of the departmental treatment areas. For example, the Acute area will almost always be fully staffed by nurses, and the short staffing will be felt elsewhere.

### 5.4.2 Individual Staff

Regarding individual staff constraints and relationships, there are a number of finite resources which can contribute to wait times and flow patterns. The nurses may or may not be pooled by area or be assigned rooms or specialties for the shift. They also have different efficiencies based on Emergency Department experience. These types of distinctions are important to ensure balanced workflow that matches the real situation. Porters, communication clerks, bed allocators, and registration clerks also have the potential for blocking or starving the nurses and physicians. For example, if the communications clerk is busy trying to help track down a specialist, they are not monitoring and moving patients from the triage/waiting area back to the treatment areas. If the porter is tied up transporting a patient to the diagnostic imaging area, they are not doing other duties which may impact flow for other patients. If the bed allocator is having trouble finding isolation beds for patients being admitted to the inpatient wards, the beds in Emergency remain occupied and delay new incoming patients. While the patient remains in Emergency, they will require additional nursing resources, which may impact the flow of other patients.

At St. Mary’s, there are a number of medical directives which relate to initial care and assessments that can be possibly performed by the registered nurses. The medical directives are optional, and they can affect the timing and types of tests ordered. It cannot be assumed that all orders occur after the physician assessment, and it is not possible to assume that all possible medical directives are applied.

The nursing processes might also affect patient flow. For example, when is the nurse charting done and when can the physician turn their attention to the patient? In most cases, once the nurse completes the charting, the physical patient record becomes available for the physician to process. When a level I acuity arrives, the normal process is preempted by the emergency situation. This type of distinction is important if electronic records are used to generate simulated time distributions as there are different ways to interpret the time data based on the acuity.

### 5.4.3 External Factors

Modeling the constraints and relationships external to the department at St. Mary's is important, and there are differences by time of day as well. For example, the time it takes to track down a specialist for consultation is not the same during the day as it is between midnight and 7:30 AM. The time to admit, transfer, or have diagnostic imaging performed is also variable by time of day.

If the medical disposition involves anything other than simple departure, it is likely that the departure paperwork and communications will be more extensive and will require more resources. It is not common for patients to expire at St. Mary's Emergency Department (i.e., it is not the trauma Emergency Department for the community), yet when this happens, it is a major impact on physician and nursing resources.

## *5.5 "Out-of-the-Ordinary" Impacts*

During the observation period of the study, we consistently observed periods when the work flowed smoothly and the length of stay times were relatively short. Staff and patient flows were balanced and there was little stress in the department. This flow pattern was confirmed by looking at the patterns of flow and the number of times patients received quick service (40–50% of the time). The balanced and smooth flow periods were also physician independent for the most part. There was some difference based on physician experience, but even the junior physicians had significant portions of their patients receiving quick service.

We also noted periods of time when the work flow was not balanced, patients were waiting for extended periods of time, and the personnel were dealing with a chaotic situation. The "bad days" were not always related to volumes, days of the week, or any obvious pattern trigger. The staff was often unable to explain why one day was better than another just by looking at the daily performance numbers. The overall behavior was similar to what had been observed in lean or fragile production systems where the capacity is tuned to the mean situation. Whenever something unusual presented itself, the system became overloaded as it does not have any spare capacity or resources to deal with the changing situation. Depending on the type and magnitude of the perturbation, other work flows were impacted. These observations led to a specific analysis for symptoms and causes of the department process deteriorating and leading to extended length of stays.

There were occasionally specific events which impacted multiple resources for an extended time. During these events, the resources were unable to attend to other patients or move any of the processes forward. For example, if a Code Blue (patient incurring a cardiac arrest) occurred elsewhere in the hospital, Emergency Department personnel (including a physician) were part of the response team. When such a situation occurred, the department's capability was reduced and depending on when the event took place, quite a few patients were affected. Other observed special events included fire alarms, machine breakdowns in the test area, patients accompanied by police officers, isolation cases, long-term drug users, patients with dementia, agitated patients admitted without their permission, language barriers, and certain treatment plans which required frequent nursing attention.

There are four possible impacts whenever a special event occurs. First, more resources than normal are allocated to the patient. Second, any multitasking is preempted and the resources are removed from the average work distribution pattern. Third, the dispatching or work flow after a special event is not normal as special heuristics are used to restore the department to its usual behavior. Fourth, when the resources return to normal processing, a wave of activity and requirements may flow downstream. When you often have one or more special events happening per shift, it is important to include such events and subsequent work flows in the model. At St. Mary’s, approximately 75% of the days studied had some form of special event, and any detailed model will have to include this type of situation as a noise-generating event.

It is possible that a categorization schema could be developed for the most common types of special situations and distributions developed for analytical purposes. One of the problems is gaining sufficient information about when a special event or situation has occurred and what is the impact. This type of data is not normally gathered and entered on a patient’s chart. The occurrence and impact of special situations is an area for further research at St. Mary’s. Analyzing the nonrandom clumping of slow patients has assisted in understanding the types of potential improvement associated with continuous improvement versus special operational procedures to deal with special cases.

## *5.6 Operations Research Models*

In a relatively complex situation such as the Emergency Department, it is important to validate any detailed operational model. It is not sufficient to simply look at one or two average output metrics and state that the model has internal validity. For example, at St. Mary’s, the internal flow patterns for the major population groups (e.g., triage level and age) must also be validated on a base case if different operational options are being explored. Specifically:

- Any known drivers of resource or extended wait times would need to match (e.g., arrival patterns, acuity mix, presenting complaints, seasonal indices).
- The physician productivity model would be a key component for validation. Does the patient per hour rate make sense?
- The flow times for major population segments (e.g., age) for the process steps and services (e.g., imaging tests) would be important to validate.
- Because bottlenecks may float, any starving and blocking patterns on bottleneck resources, as well as resource conflict patterns must match the real situation. For example, in the base line case, does the queueing behavior match reality?

- Low probability out-of-the-ordinary patterns must be generated, and the system components react accordingly.

As noted, this would be for the baseline before introducing any changes. These five aspects should have a close resemblance to a known case if claims are to be made for the results of an intervention.

## 6 Conclusions

It appears from the academic and practice literature that the majority of research on Emergency Department has been at the aggregate level interfacing with ambulance arrival or bed utilization, resource scheduling for nurses or physicians, or a simulation style analysis of a specific Emergency Department. There have been fewer studies of multiple Emergency Departments, research involving specific resource performance characteristics, or resource behavior; however, as these studies have shown, there are many relationships and details at the operational level that should be taken into account when preparing a detailed model. It is also important to clearly state assumptions and validation criteria to ensure that the limits of the modeling effort are well understood.

As an applied research field, Emergency Department problem(s) can be considered relatively immature and undeveloped in the context of resource characteristics and behavior. Research efforts and practitioners should be aware of this and should treat results with appropriate care. It is necessary to shift from one-off simulation studies of specific hospitals and develop systemic and holistic research agendas and efforts focused on dynamic relationships, sources of uncertainty and variance, and resource behavioral patterns.

Section 5 presented a number of observations that arose from an extensive study of the flows and behaviors at a particular Emergency Department, and we believe that many of these observations would be relevant to other Emergency Departments; however, it is clear that any single concept or recommendation has to be understood in the context of the different Emergency Department practices found within a region. There are also significant differences between teaching hospital situations and community hospitals. When healthcare resources in a community are collaborating on emergency care, this can also create unique flow patterns and relationships. We believe that it is important for researchers to understand the context they are studying and how this might limit the generality of any results.

The Emergency Department is a fascinating applied research problem. There are many inter-connected components and dynamic relationships which are not immediately obvious from the static data found in the electronic records. At an aggregate modeling level, these factors are not important but become critical when speculating about individual patient flows and daily efficiency and effectiveness. We hope that this chapter has helped shed some light on the modeling challenges and what might be considered in detailed research models. We also hope that we have been able to provide insights and guidance to practitioners who are interested in understanding the issues and relationships to be found in an Emergency Department.

**Acknowledgements** The chapter has benefited from the critical suggestions of the reviewers and by Dr. Blair Egerdie, Gary Black, Fatih Erenay, John Buzacott, Reha Uzsoy, and Brian Denton. Their suggestions have been very insightful and beneficial. The chapter would not have been possible without the tremendous support and assistance of the hospital and the people working in the Emergency Department. During the study the analysis team received a great deal of help from the nurses, clerks, porters, and physicians. Additional valuable assistance came from other units in the hospital as well. The analysis team consisted of Jennifer Engels, Sahil Jain, and Kenneth McKay. At the time of the study, Jennifer Engels was completing her MASc and volunteered to be a research assistant on the project. Sahil Jain was a senior medical student and participated in the study as part of his educational program.

## References

Andersson AK, Omberg M, Svedlund M (2006) Triage in the emergency department—a qualitative study of the factors which nurses consider when making decisions. Nurs Crit Care 11(3): 136–145

Anthony RN (1965) Planning and control systems: a framework for analysis. Harvard Business School, Boston

Asaro PV, Lewis LM, Boxerman SB (2007) The impact of input and output factors on emergency department throughput. Acad Emerg Med 14(3):235–242

Asplin BR, Magid DJ, Rhodes KV, Solberg LI, Lurie N, Camargo CA Jr (2003) A conceptual model of emergency department crowding. Ann Emerg Med 42(2):173–180

Babbage C (1832) On the economy of machinery and manufactures, 2nd edn. Charles Knight, London

Bagust A, Place M, Posnett JW (1999) Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. BMJ 319(7203):155–158

Bailey NTJ (1956) Statistics in hospital planning and design. J R Stat Soc Ser C 5(3):146–157

Beaulieu H, Ferland JA, Gendron B, Michelon P (2000) A mathematical programming approach for scheduling physicians in the emergency room. Health Care Manag Sci 3:193–200

Bell J (2011) A&E waiting times increase sharply. The Guardian, Tuesday 5 April 2011. Web link: http://www.guardian.co.uk/society/2011/apr/04/accident-emergency-waiting-times-increase. Accessed 25 Nov 2012

Blake JT, Carter MW, Richardson S (1996) An analysis of emergency room wait time issues via computer simulation. INFOR 34(4):263–273

Boldy D (1976) A review of the application of mathematical programming to tactical and strategic health and social services problems. Oper Res Q 27(2):439–448

Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice-Hall, New Jersey

CIHI—Canadian Institute for Health Information (2011) Wait times in Canada—a comparison by province. CIHI

Carter MW, Lapierre SD (2001) Scheduling emergency room physicians. Health Care Manag Sci 4:347–360

Carter MW, Blake JT (2005) Using simulation in an acute-care hospital: easier said than done. In: Brandeau M, Sainfort F, Pierskalla W (eds) Handbook of operations research in healthcare. Kluwer, Boston, pp 191–215

Chan L, Reilly KM, Salluzzo RF (1997) Variables that affect patient throughput times in an academic emergency department. Am J Med Qual 12(4):183–186

Chisholm CD, Collison EK, Nelson DR, Cordell WH (2000) Emergency department workplace interruptions: are emergency physicians “interrupt-driven” and “multitasking”. Acad Emerg Med 7(11):1239–1243

Coiera E, Tombs V (1998) Communication behaviours in a hospital setting: an observational study. BMJ 315:673–676

Coleman JV, Errera P (1963) The general hospital emergency room and its psychiatric problems. Am J Public Health 53–8:1294–1301

Connelly LG, Bair AE (2004) Discrete event simulation of emergency department activity: a platform for system-level operations research. Acad Emerg Med 11:1177–1185

Crane J, Noon C (2011) The definitive guide to emergency department operational improvement. Productivity Press, New York

Dobson G, Lee HH, Sainathan A, Tilson V (2012) A queueing model to evaluate the impact of patient “batching” on throughput and flow time in a medical teaching facility. MSOM 14(4): 584–599

Downey LA, Zun LS (2007) Determinates of throughput times in the emergency department. J Health Manag 9(1):51–58

Duguay C, Chetouane F (2007) Modeling and improving emergency department systems using discrete event simulation. Simulation 83(4):311–320

Facchin P, Rizzato E, Romanin-Jacur G (2010) Emergency department generalized flexible simulation model. In: 2010 IEEE workshop on health care management (WHCM). Venice, Italy, pp 1–6

Fairbanks RJ, Bisantz AM, Sunm M (2007) Emergency department communication links and patterns. Ann Emerg Med 50(4):396–406

Fetter RB, Thompson JD (1965) The simulation of hospital systems. Oper Res 13(5):689–711

Fisher M (2007) Strengthening the empirical base of operations management. MSOM 9(4): 368–382

Flagle CD (2002) Some origins of operations research in the health services. Oper Res 50(1):52–60

Fletcher A, Halsall D, Huxham S, Worthington D (2007) The DH accident and emergency department model: a national generic model used locally. J Oper Res Soc 58:1554–1562

Fletcher A, Worthington D (2009) What is a ‘generic’ hospital model?—a comparison of ‘generic’ and ‘specific’ hospital models of emergency patient flows. Health Care Manag Sci 12:374–391

Fomundam S, Herrmann J (2007) A survey of queuing theory applications in healthcare. ISR technical report, pp 2007–2024

Fries BE (1979) Bibliography of operations research in health-care systems: an update. Oper Res 27(2):408–419

GAO (2009) Hospital emergency departments—crowding continues to occur, and some patients wait longer than recommended time frames. Report, United States Government Accountability Office

Gascon V, Villeneuve S, Michelon P, Ferland JA (2000) Scheduling the flying squad nurses of a hospital using a multi-objective programming model. Ann Oper Res 96:149–166

Gaur KN (2008) Operations research in hospitals. In: Srinivasan AV (ed) Managing a modern hospital, 2nd edn. Sage, Thousand Oaks

Gendreau M, Ferland J, Gendron B, Hail N, Jaumard B, Lapierre S, Pesant G, Soriano P (2006) Physician scheduling in emergency rooms. In: Proceedings of practice and theory of automated timetabling conference (PATAT) 2006. Brno, Czech Republic, pp 2–14

Green L (2006) Queueing analysis in healthcare. In: Hall RW (ed) Patient flow: reducing delay in healthcare delivery, Springer International Series. Springer, New York, pp 281–307

Green LV, Soares J, Giglio JF, Green RA (2006) Using queueing theory to increase the effectiveness of emergency department provider staffing. Acad Emerg Med 13(1):61–68

Gunal MM, Pidd M (2006) Understanding accident and emergency department performance using simulation. In: Proceedings of 2006 winter simulation conference. Monterey, US, pp 446–452

Hall RW (ed) (2006) Patient flow: reducing delay in healthcare delivery, Springer international series. Springer, New York

Hamilton WF (1974) Systems analysis in emergency care planning. Med Care 12(2):152–162

Han JH, France DJ, Levin SR, Jones ID, Storrow AB, Aronsky D (2010) The effect of physician triage on emergency department length of stay. J Emerg Med 39–2:227–233

Handyside AJ, Morris D (1967) Simulation of emergency bed occupancy. Health Serv Res Fall-Winter 2(3):287–297

Hannan EL (1975) Planning an emergency department holding unit. Socio Econ Plann Sci 9(5):179–188

Holroyd BR, Bullard MJ, Latoszek K, Gordon D, Allen S, Tam S, Blitz S, Yoon P, Rowe BH (2007) Impact of triage liaison physician on emergency department overcrowding and throughput: a randomized controlled trial. Acad Emerg Med 14:702–708

Hoffman B (2006) Emergency rooms: the reluctant safety net. In: Stevens RA, Rosenberg CE, Burns LR (eds) History & health policy in the United States: putting the past back in. Rutgers University Press, New Jersey, pp 250–272

Hoot NR, LeBlanc LJ, Jones I, Levin SR, Zhou C, Gadd CS, Aronsky D (2008) Forecasting emergency department crowding: a discrete event simulation. Ann Emerg Med 52(2):116–125

Innes GD, Stenstrom R, Grafstein E, Christenson JM (2005) Prospective time study derivation of emergency physician workload predictors. Can J Emerg Med 7(5):299–308

Jacobson SH, Hall SN, Swisher JR (2006) Discrete-event simulation of health care systems. In: Hall RW (ed) Patient flow: reducing delay in healthcare delivery, Springer International Series. Springer, New York, pp 211–252

Jayaprakash N, O’Sullivan R, Bey T, Ahmed SS, Lotfipour S (2009) Crowding and delivery of healthcare in emergency departments: the European perspective, Western. J Emerg Med 10(4):233–239

Jones SS, Evans RS (2008) An agent based simulation tool for scheduling emergency department physicians. In: AMIA 2008 symposium proceedings. Washington, US, pp 338–342

Jones SS, Thomas A, Evans RS, Welch SJ, Haug PJ, Snow GL (2008) Forecasting daily patient volumes in the emergency department. Acad Emerg Med 15:159–170

Jun JB, Jacobson SH, Swisher JR (1999) Application of discrete-event simulation in health care clinics: a survey. J Oper Res Soc 50(2):109–123

Kolesar P (1970) A Markovian model for hospital admission scheduling. Manag Sci 16(6): B384–B396

Komashie A, Mousavi A (2005) Modeling emergency departments using discrete event simulation techniques. In: Proceedings of 2005 winter simulation conference. Orlando, US, pp 2681–2685

Lane DC, Monefedlt C, Rosenhead JV (2000) Looking in the wrong place for healthcare improvements: a systems dynamics study of an accident and emergency department. J Oper Res Soc 51(5):518–531

Laxmisan A, Hakimzada R, Sayan OR, Green RA, Zhang J, Patel VL (2007) The multitasking clinician: decision-making and cognitive demand during and after team handoffs in emergency care. Int J Med Inform 76:801–811

Levin S, France DJ, Hemphill R, Jones I, Chen KY, Rickard D, Makowski R, Aronsky D (2006) Tracking workload in the emergency department. Hum Factors 48(3):526–539

Levin S, Aronsky D, Hemphill R, Slagle J, France DJ (2007) Shifting toward balance: measuring the distribution of workload among emergency physician teams. Ann Emerg Med 50(4): 419–423

Lynn SG, Kellermann AL (1991) Critical decision making: managing the emergency department in an overcrowded hospital. Ann Emerg Med 20(3):287–292

Malakooti B, Malakooti NR, Yang Z (2004) Integrated group technology, cell formation, process planning, and production planning with application to the emergency room. Int J Prod Res 42(9):1769–1786

Mayhew L, Smith D (2008) Using queuing theory to analyze the Government’s 4-h completion time target in accident and emergency departments. Health Care Manag Sci 11:11–21

McKay KN (1987) Conceptual framework for job shop scheduling, M.A.Sc. Dissertation, University of Waterloo

McKay KN (1992) Production planning and scheduling: a model for manufacturing decisions requiring judgement. Ph.D. Dissertation, University of Waterloo

McKay KN (2011) Inter-domain translational research on planning and scheduling—operating rooms versus job shops. Int J Plann Scheduling 1(1/2):42–57

Mentis, H.M. (2010) Emotion awareness and invisibility in an emergency room: a socio-technical dilemma. Ph.D. Dissertation, Pennsylvania State University

Mentis H, Reddy M, Rosson MB (2010) Invisible emotion: information and interaction in an emergency room. CSCW 2010:311–320

Ministry of Health NZ (2011) Targeting emergencies—shorter stays in emergency departments. Report

NHS National Services Scotland (2011) Emergency department activity. Report

Newell DJ (1954) Provision of emergency beds in hospitals. Br J Prev Soc Med 8:77–80

Newell DJ (1965) Unusual frequency distributions. Biometrics 21(1):159–168

Paul SA, Reddy MC, Deflitch CJ (2010) A systematic review of simulation studies investigating emergency department overcrowding. Simulation 86(8–9):559–571

Pierskalla WP, Brailer DJ (1994) Applications of operations research in health care delivery. In: Pollock SM et al (eds) Handbooks in OR & MS, vol 6. Elsevier Science, Amsterdam, pp 469–505

Pinedo ML (2009) Planning and scheduling in manufacturing and services. Springer, New York

Rais A, Viana A (2010) Operations research in healthcare: a survey. Int Trans Oper Res 18:1–31

Saunders CE, Makens PK, Leblanc LJ (1989) Modeling emergency department operations using advanced computer simulation systems. Ann Emerg Med 18(2):134–140

Samaha S, Armel WS, Starks DW (2003) The use of simulation to reduce the length of stay in an emergency department. In: Proceeding of the 2003 winter simulation conference. New Orleans, US, pp 1907–1911

Seila AF, Brailsford S (2009) Opportunities and Challenges in Health Care Simulation. In: Advancing the Frontiers of Simulation, Alexopoulos C et al (eds) Intl Series in Operations Research & Management Science, Springer, 133:195–229

Shingo H (2010) Emergency medicine in Japan. Keio J Med 59(4):131–139

Shiver JM, Eitel D (2009) Optimizing emergency department throughput. Taylor Francis, New York

Sinreich D, Marmor YN (2005) Emergency department operations: the basis for developing a simulation tool. IIE Trans 37:233–245

Sleptchenko A, van der Heijden MC, van Harten A (2005) Using repair priorities to reduce stock investment in spare part networks. Eur J Oper Res 163:733–750

Soni K, Saxena K (2011) A study of applicability of waiting line model in health care: a systematic review. Int J Manag Tourism 19(1):75–91

Taylor J (2006) Don't bring me your tired, your poor: the crowded state of America's emergency departments. Natl Health Policy Forum—Issue Brief July 7, 2006 (811): 1–24

Ting HH, Lee TH, Soukup J, Cook EF, Tosteson ANA, Brand DA, Rouan GW, Goldman L (1991) Impact of physician experience on triage of emergency room patients with acute chest pain at three teaching hospitals. Am J Med 91:401–408

Topping A, Campbell D (2010) Waiting targets for accident and emergency to be scrapped. The Guardian, Thursday 10 June 2010. Web link: http://www.guardian.co.uk/politics/2010/jun/10/accident-and-emergency-waiting-time-nhs. Accessed 25 Nov 2012

Turner J, Mehrotra S, Daskin MS (2010) Perspectives on health-care resource management problems. In: Sodhi MS, Tang CS (eds) A long view of research and practice in operations research and management science, International series in operations research & Management Science. Springer, New York, p 148

Ullman R, Block JA, Stratmann WC (1975) An emergency room's patients: their characteristics and utilization of hospital services. Med Care 13(12):1011–1020

Van Sambeek JRC, Cornelissen FA, Bakker PJM, Krabbendam JJ (2010) Models as instruments for optimizing hospital processes: a systematic review. Int J Health Care Qual Assur 23(4): 356–377

Weinerman ER, Edwards HR (1964) ‘Triage’ system shows promise in management of emergency department load. J Am Hospitals Assoc 38(4):55–62

Weinerman ER, Rutzen SR, Pearson DA (1965) Effects of medical “triage” in hospital emergency service. Yale Studies Ambulatory Med Care 80(5):389–399

Wiler JL, Griffey RT, Olsen T (2011) Review of modeling approaches for emergency department patient flow and crowding research. Acad Emerg Med 18:1371–1379

Wilper AP, Woolhandler S, Lasser KE, McCormick D, Cutrona SL, Bor DH, Himmelstein DU (2008) Waits to see an emergency department physician: U.S. trends and predictors, 1997–2004. Health Aff 27(2):w84–w95

Woloshynowych M, Davis R, Brown R, Vincent C (2007) Communication patterns in a UK emergency department. Ann Emerg Med 50(5):407–413

Yang CC, Lin WT, Chen HM, Shi YH (2009) Improving scheduling of emergency physicians using data mining analysis. Expert Syst Appl 36:3378–3387

Yeh JY, Lin WS (2007) Using simulation technique and genetic algorithm to improve the quality care of a hospital emergency department. Expert Syst Appl 32:1073–1083

# Chapter 14
# Location Models in Healthcare

Bjorn P. Berg

## 1 Introduction

Decisions about the location of critical resources in a geographic network are important in many industries. For healthcare delivery these decisions have included the location of emergency care services (Toregas et al. 1971; Revelle et al. 1977), preventive care (Verter and Lapierre 2002), and public health planning (Lapierre et al. 1999). Location decisions are important because they determine the travel time between health resources (supply) and patients (demand). Location decisions are particularly important for emergency preparedness, where location decisions determine the response time and therefore access to care.

Decisions that form the basis for location models include the number of resources that should be located, where to locate the available resources, and how customers should be assigned to resources locations. The criteria used to evaluate the quality of solutions to location models include the total travel distance between for customers, the total cost of locating the decided number of resources, the maximum distance that any customer will need to travel, or the number of customers that have access to a resource within a predetermined distance tolerance. However, the decisions surrounding location models are difficult due to the large number of possible resource location. Systematically comparing every set of location decisions can be a computationally cumbersome process and may not be possible within a reasonable time frame for many decision makers.

The purpose of this chapter is to present an overview of location models and their application to healthcare settings. We begin by reviewing examples of location models applied to problems in healthcare. We do not intend to provide an exhaustive survey of the literature; rather, the goal is to identify examples of location modeling

B.P. Berg (✉)
George Mason University, Fairfax, VA 22030, USA
e-mail: bberg2@gmu.edu

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_14, © Springer Science+Business Media New York 2013

that present the nuances that healthcare problem settings present. Next, we present an overview of model formulations that are commonly used in location models for the purpose of identifying each model's benefits and limitations. We then use a case study based on the location of stockpiled nerve gas antidote in North Carolina to compare the alternative models discussed. Multiple performance criteria for the nerve gas antidote problem are identified and compared. Finally, we discuss the chapter's conclusions and identify important areas for future research.

## 2 Examples of Healthcare Location Models

In this section we present examples of application areas where location models have been used in healthcare. These examples demonstrate the wide variety of problems within healthcare that benefit from the insights that location model formulations offer. For a more inclusive review and classification of location models in healthcare, the reader is referred to Daskin and Dean (2005).

The authors of Price and Turcotte (1986) describe the challenges in deciding where to locate a blood bank. Many of the challenges identified are common to other types of location problems in healthcare and the public sector in general, including having many stakeholders, needing to serve a large and diverse population, and making decisions based on little or no data. In the process of locating a blood bank in Quebec City, Canada, high levels of public access needed to be obtained while maintaining convenience for surrounding clinics that deliver blood to the bank. The authors used donor transportation surveys, postal codes from donor addresses, and census data to create location models to identify potential locations for the new blood bank. Following analysis, the potential locations were then reviewed for practical constraints not included in the models, such as freeway access and geographical limitations. The sites were then ranked across multiple criteria and recommendations were made to the stakeholders.

Another example motivated by timely access to care is the model developed in Côté et al. (2007) where patients with traumatic brain injuries (TBI) need timely treatment to achieve the best outcomes. A mixed integer programming model is formulated to locate treatment units, which include access to various types of multidisciplinary therapies, within the health system of the Department of Veterans Affairs. To obtain a region's expected TBI incidence, TBI hospitalization incidence rates from the Centers for Disease Control (CDC) were applied by adjusting for a region's age, gender, and veteran population. The model was solved for various scenarios constructed based on managerial preference criteria such as the level of centralization of the treatment units. The authors note that the optimal location-allocation policy is dependent upon the managerial preferences and environment parameters. However, the analysis presented a set of policies to which decision makers could apply their judgment.

The role of location models in developing health systems in developing nations is articulated in Rahman and Smith (2000). The authors explain the implications that access to health services has on a nation's economic development and quality of life. A common trait of location models in a developing nation's health system is the need to include the hierarchical structure that is inherent to the health system organization. For example, each neighborhood or community may have a small number of primary care clinics that serve as the first contact to the health system. As the need for more specialized and coordinated care arises, patients may be referred to more centralized healthcare centers for surgery or more intensive care. The authors go on to explain that although location models are often dismissed as too sophisticated for use in health system planning in developing nations, there are many reported applications of location models in developing nation settings including identifying new hospital locations, locating rural clinics, and locating medical supply centers.

Many location models in healthcare are concerned with timely access to emergency services. A recent review of location models in emergency services is provided in Jia et al. (2007). Besides identifying common emergency service location model formulations, the authors concentrate on services for large-scale emergencies, such as natural disasters or terrorist attacks, and identify the unique aspects that need to be considered when locating medical services for rare, but high impact, events. The authors explain that for large-scale emergency events, medical services need to be redundant and dispersed. Redundancy is necessary due to the extremely large demands being placed on single facilities. However, if such facilities themselves are rendered inefficient as a result of an emergency, other facilities need to be dispersed in order to provide continuing service. Another important aspect that needs to be included in large-scale emergency service location models is accounting for differences in the likelihood of different types of events occurring in different regions based on geographic or population differences. The authors adapt common location model formulations to account for the differences presented by large-scale emergency planning and use several example scenarios for demonstration.

Location models have also been formulated using medical outcomes as an objective, extending the traditionally operations-based decisions in location models to the area of medical decision making. As an example, the problem of organizing transplant regions in the United States was formulated as a mixed integer program (MIP) in Stahl et al. (2005). In the hierarchical organization of liver donations, donor livers are made available first to recipients in the same region due to faster transportation time of the organ and thus higher organ quality. However, a smaller regional structure also creates inefficiencies since it results in a smaller donor pool. The authors formulate a model whose objective is to maximize efficiency and equity through transplant region design. Although the focus is on how to organize the country into transplant regions, the relevance to location models is clear in the decision process of allocating patients to the transplant regions. The authors conclude that their optimal region design offers important insights when compared with the current region design in place, such as organizing regions based on densely populated areas.

Preventive care is another healthcare setting that presents unique challenges for location models. For example, as the authors explain in Verter and Lapierre

(2002), utilization of preventive care service is inversely related to the distance and access that patients have to preventive care locations. The intuitive solution is to develop many preventive care settings dispersed throughout a community. However, the high cost of opening and operating many facilities is not the only constraint involved. For example, each facility must serve a minimum number of patients so providers may maintain their accreditation. In Verter and Lapierre (2002), the authors formulate a location model to maximize participation while ensuring a defined patient threshold. Data from public health centers in Fulton County, Georgia, and breast cancer screening centers in Montreal, Canada, are used to compare two proposed solution methods. The author's also note the potential extension of the work within a hierarchical model framework.

In a more recent example, the authors of Chanta et al. (2011) introduce equity as a criterion in locating emergency medical services. In determining the locations of ambulances at stations, the authors location model seeks to balance customer perceptions of equity in access to services by minimizing customer *envy*. The location model is formulated as an integer program. A heuristic method is introduced as a result of the integer programming formulation's computational challenge. The solution quality of the heuristic results are shown to perform favorably when compared to other common location model formulations. By including equity in their location model formulation, the authors demonstrated how location models motivated by healthcare applications are generating opportunities for developing novel models and methods in the broader scope of location models.

## 3 Types of Location Models

Location models are commonly formulated as MIPs. The decision variables often define the choice of locations for resources within a defined network and how customers are to be assigned to the chosen locations. The objective function defines the criteria by which alternative locations are compared, usually in terms of distance or some measure of access. Criteria can vary among decision makers and may include minimizing cost, maximizing the proportion of customers within a desired distance of the resources, or minimizing the maximum distance of any given patient from a resource. The constraints typically limit the location decisions based on the availability of limited resources. We review the most common MIPs used for making location decisions in practice and discuss the pros and cons of each formulation. We also present an easy-to-implement heuristic that can provide near optimal solutions to large problems.

Many applications of location problems have been reported in the literature. Although facility location problems have been formulated in stochastic and dynamic contexts, we focus on the deterministic models due to their ease of implementation and because they are often reasonably effective even in stochastic environments. For more comprehensive coverage of location models, including stochastic models, we refer the reader to Francis and Goldstein (1974), Handler and Mirchandani (1979), Love et al. (1988), Daskin (1995).

In a location problem there are *customers* that need to be served. Serving the populations are a set of *resources*, often referred to as *facilities*. There are assumed to be a predetermined set of $N$ potential locations for facilities that the decision maker may choose from. Often the number of resources are limited, and the decision maker must choose a subset of size $p$ from the $N$ potential locations. Thus, the total number of possible decisions is $\binom{N}{p}$. Therefore, the number of alternative decisions is often very large. The decision maker is faced with the challenge of choosing the best set of locations among a large number of alternatives.

### 3.1 *p*-Median Formulation

The $p$-median model locates a fixed number, $p$, of resources out of a greater set of preselected possible locations, $N$. Populations to be served by the locations have a size and location, where the location is the centroid of the zip code. The distance a population travels to a possible facility location is weighted by the size of the population. In the $p$-median model and all the models covered in this section, populations are served by the nearest located resources. The $p$-median model is suitable for settings where there are a limited number of resources available, such as the limited number of treatment units in Côté et al. (2007).

In the $p$-median model formulation, populations are indexed by $i$, and locations are indexed by $j$. The objective is to determine the subset of size $p$ that minimizes the weighted travel distance. The MIP formulation can be written as follows:

**Parameters**

$d_{ij}$: Distance from population $i$ to possible facility location $j$
$w_i$: Size of population $i$ (weight)
$p$: Number of facilities to locate

**Decision Variables**

$$y_j = \begin{cases} 1 & \text{if a facility is located at possible location } j \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij} = \begin{cases} 1 & \text{if population } i \text{ is served by a facility at location } j \\ 0 & \text{otherwise} \end{cases}$$

$$\min \sum_{\forall i} \sum_{\forall j} w_i d_{ij} x_{ij} \tag{14.1}$$

$$\text{s.t.} \sum_{\forall j} x_{ij} = 1 \quad \forall i \tag{14.2}$$

$$\sum_{\forall j} y_j = p \tag{14.3}$$

$$x_{ij} \leq y_j \quad \forall i, j \tag{14.4}$$

$$x_{ij}, y_j \in \{0, 1\} \quad \forall i, j \tag{14.5}$$

The objective, (14.1), is to minimize the total demand-weighted distance between populations and facilities. The constraints limit the allowable decisions to feasible decisions. Constraint (14.2) ensures that each population is assigned to exactly one located facility, whereas constraint (14.3) guarantees exactly $p$ facilities will be located. Constraint (14.4) states a population can only be assigned to a location if a facility is located there, and (14.5) defines the decision variables as binary.

## *3.2 Uncapacitated Facility Location Model*

Closely related to the $p$-median formulation is the uncapacitated facility location (UFL) model. In the UFL, constraint (14.4) of the $p$-median formulation is relaxed. That is, there is no longer a limit on the number of facilities that can be located. Therefore, there is no reason not to locate a facility for every element of the population. In order to discourage opening an unnecessary number of facilities, a cost term is added to the objective function. This can be interpreted as the cost of opening an additional facility relative to the cost of the distance traveled. UFL models may be appropriate for identifying flu clinic locations, for example, since the cost for locating flu clinics will likely be the limiting factor as opposed to available resources such as staff or facilities. Similar to the $p$-median formulation, the MIP formulation for the UFL model can be written as follows:

**Parameters**

$d_{ij}$: Distance from population $i$ to possible facility location $j$
$w_i$: Size of population $i$ (weight)
$f_j$: Cost of locating facility $j$

### Decision Variables

$$y_j = \begin{cases} 1 & \text{if a facility is located at possible location } j \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij} = \begin{cases} 1 & \text{if population } i \text{ is served by a facility at location } j \\ 0 & \text{otherwise} \end{cases}$$

### *p*-Median Formation

$$\min \sum_{\forall i} \sum_{\forall j} w_i d_{ij} x_{ij} + \sum_{\forall j} f_j y_j \tag{14.6}$$

$$\text{s.t.} \sum_{\forall j} x_{ij} = 1 \quad \forall i \tag{14.7}$$

$$x_{ij} \leq y_j \quad \forall i, j \tag{14.8}$$

$$x_{ij}, y_j \in \{0, 1\} \quad \forall i, j \tag{14.9}$$

## 3.3 *p*-Center Model

The third model, called the *p*-center model, seeks to minimize the maximum distance traveled by any population. Thus, it differs from the *p*-median and UFL models in that it focuses on limiting the worst case. Such models are more likely to be applicable to public sector settings where reasonable access to resources such as a hospital, or emergency services, is necessary for the entire public. The *p*-center MIP is formulated as follows:

### Parameters

$d_{ij}$: Distance from population $i$ to possible facility location $j$
$p$: Number of facilities to locate

### Decision Variables

$$y_j = \begin{cases} 1 & \text{if a facility is located at possible location } j \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij} = \begin{cases} 1 & \text{if population } i \text{ is served by a facility at location } j \\ 0 & \text{otherwise} \end{cases}$$

$D =$ maximum distance between a population and the nearest facility

**$p$-Center Formation**

$$\min D \tag{14.10}$$

$$\text{s.t.} \sum_{\forall j} x_{ij} = 1 \quad \forall i \tag{14.11}$$

$$\sum_{\forall j} y_j = p \tag{14.12}$$

$$x_{ij} \leq y_j \quad \forall i, j \tag{14.13}$$

$$D \geq \sum_{\forall j} d_{ij} x_{ij} \quad \forall i \tag{14.14}$$

$$x_{ij}, y_j \in \{0, 1\} \quad \forall i, j \tag{14.15}$$

This formulation is similar to the $p$-median formulation, but the objective function (14.10) minimizes the distance, $D$, that is defined as the maximum travel distance any population travels to the nearest facility. As in the $p$-median formulation, the $p$-center formulation requires that every population is assigned to one facility, $p$ facilities are located, and a population can only be served by a facility that has been located. This model favors equity of cost over efficiency (UFL) and average travel distance ($p$-median)

## 3.4 Maximal Covering Model

In the context of allocating emergency resources, it may be the case that there exists a tolerance, or maximum distance, to which a location plan must adhere. For instance, there may be a maximum acceptable response time. For example, in North America a common EMS response time target is 9 min 90% of the time in urban settings and 15 min 90% of the time in rural settings (Fitch 2005). Thus, resources within a specified proximity satisfy this requirement, and resources outside do not. A population is considered *covered* if a facility is located within the required distance. The MIP corresponding the maximal covering model can be written as follows:

**Parameters**

$d_{ij}$: Distance from population $i$ to possible facility location $j$
$w_i$: Weight of population $i$ (size)

$p$: Number of facilities to locate
$A$: Acceptable distance between a population and a located facility

**Decision Variables**

$$y_j = \begin{cases} 1 & \text{if a facility is located at possible location } j \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij} = \begin{cases} 1 & \text{if population } i \text{ is served by a facility at location } j \\ 0 & \text{otherwise} \end{cases}$$

$$z_i = \begin{cases} 1 & \text{if population } i \text{ is covered} \\ 0 & \text{otherwise} \end{cases}$$

**Maximal Covering Formulation**

$$\max \sum_{\forall i} w_i z_i \tag{14.16}$$

$$\text{s.t.} \sum_{\forall j} x_{ij} = z_i \quad \forall i \tag{14.17}$$

$$\sum_{\forall j} y_j \leq p \tag{14.18}$$

$$d_{ij} x_{ij} \leq A y_j \quad \forall i, j \tag{14.19}$$

$$x_{ij}, y_j, z_i \in \{0,1\} \quad \forall i, j \tag{14.20}$$

In the covering formulation the objective (14.16) maximizes the weighted sum of elements that are covered. Similar to the $p$-median and $p$-center models, the maximal covering model limits the number of facilities to $p$, and every population is served by no more than one location. However, unlike the previous models, the covering formulation requires that if a population is served by a located facility, that facility must be within a certain distance as defined by (14.19). As in all of the above formulations a feasible solution is guaranteed to exist; however, it is no longer required that every population be served. That is, the acceptable distance, $A$, and the number of facilities, $p$, may not allow for a location plan that covers all population elements.

## *3.5 Other Models*

While the models discussed in Sects. 3.1–3.4 represent the most commonly used models in research and practice, there are other location models that have implications for location planning in healthcare settings. Other location models that apply to healthcare contexts are briefly discussed in this section.

Similar to the maximal covering problem, the set covering problem is focused on *covering* customers with accessible demand locations. In the set covering problem, however, the objective is to group populations into sets and assign them to resources such that the number of sets is minimized. Potential sets and resource locations are constrained by predefined acceptable distance limitations, similar to the maximal covering problem. If significant costs are associated with the resource locations, the objective of the set covering problem can be adjusted to minimize the total cost rather than the number of resources located.

Another alternative location model is the $p$-dispersion model. The objective of the $p$-dispersion model is to maximize the distance sum between located resources. The result of the $p$-dispersion model could be advantageous when deciding where to locate outpatient clinics within a provider network, for example. In this context, clinics that are located too close to each other may end up competing with each other for a patient population in addition to other provider clinics in the surrounding area. An important note regarding the $p$-dispersion problem is that the objective focuses only on the location of resources relative to each other and not the distance to customer demand.

One final alternative formulation is the hub location problem. The objective of the hub location problem is to minimize the total cost, including resource location costs and travel distance costs, of locating hub networks where large resource hubs are located along with a branching network of resources that service the hub. In healthcare, services such as labs are often described by the hub location problem where patients will have blood drawn and lab work collected at a local clinic, but the blood work and labs are processed at a centralized location for surrounding clinics.

## *3.6 Solution Methods*

Methods for solving MIPs, such as the above models, fall under the scope of the field of mathematical programming, specifically integer programming due to the discrete nature of the decisions. For a discussion and explanation of integer programming and solution methods, the reader is referred to Nemhauser and Wolsey (1988). Many advanced solution methodologies have been developed for common location model formulations that take advantage of each formulation's mathematical structure. For further information on location model solution methods, Daskin (1995) and Drezner and Hamacher (2004) provide an overview of solution methods as well as identify when particular solution methods may be more suitable based on the context.

In many location model settings, the formulations discussed above, and variants thereof, are often computationally challenging and have high solution times, even when using advanced solution methods. Therefore, *heuristics* are often resorted to in such settings. Even though the heuristic solution may not be optimal, the ease of implementation along with a fast and near optimal solution may make heuristics a suitable method.

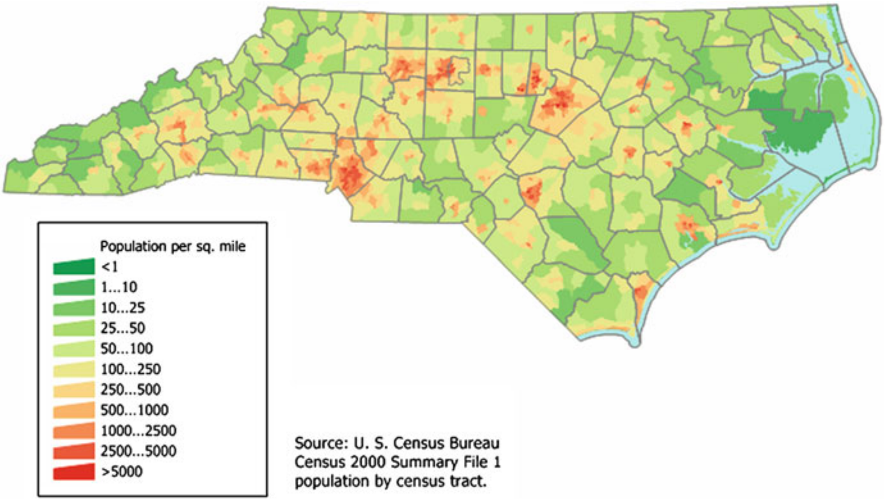
```
Algorithm 1: Greedy Heuristic for $p$-median location problem.
input  : $\mathbf{w}$ = Weighted matrix of populations for each population center $i = 1,\ldots,n$
         $\mathbf{D}$ = Matrix containing the distances from each population center $i$ to each possible
             facility location $j = 1,\ldots,m$
output: Feasible location of facilities and allocation of populations to facilities.
1 Matrix multiply $\mathbf{wD}$. Find the minimum entry in the $1 \times m$ matrix and locate a facility there.
2 If a node was just assigned a facility, set the corresponding value in $\mathbf{wD}$ equal to $\infty$, so it is
  not chosen again. If the number of currently located facilities equals $p$, stop; the current
  location of the $p$ facilities represents the feasible solution. Otherwise, go to Step 3.
3 Determine which unassigned possible facility location has the lowest cost in terms of
  weighted distance by updating $\mathbf{wD}$. Each entry that does not have a value of $\infty$ is
  recalculated by multiplying the population weight, $w_i$, by the smallest of the distances from $i$
  to either that location or a facility already located. Locate a facility at the node with the
  minimum matrix entry and go to Step 2.
```

In Eiselt and Sandblom (2004), the authors present one of the first and simplest heuristics for the $p$-median problem presented in Sect. 3.1. Their heuristic falls into a class of heuristics known as “greedy” or “myopic” heuristics. Such heuristics work by initially assuming only one resource is to be located and then finds the best location possible for that one facility. It then fixes this choice and iteratively repeats the process of assigning one more resource until all $p$ resources have been located. In practice this often results in near optimal decisions. The algorithm adapted from Eiselt and Sandblom (2004) is as follows:

The benefits of the above greedy heuristic are that it is easy to understand, easy to implement, and results in a feasible solution, often in less computation time than is required for traditional integer programming solution methods. The above heuristic is described in the context of the $p$-median problem, but similar heuristics can be adapted for the other location problems described in this section.

## 4 Case Study

In this section we present four of the most common types of location models for a specific example: the $p$-median model, the UFL model, the $p$-center model, and the maximal covering model. We use the optimal location of nerve gas antidote as an example to illustrate the use of the models described. The models are presented in the context of optimally locating ChemPacks within North Carolina. ChemPacks are large containers of stockpiled antidotes that can be used in response to the accidental or intentional release of a nerve agent or organophosphate. The ChemPack program is part of the Strategic National Stockpile (SNS). The purpose of the program is to place repositories of chemical-agent antidotes throughout the USA to be used in response to biological, radiological, or chemical attacks. Inside the ChemPacks are atropine sulfate, pralidoxime chloride, and diazepam along with MARK I auto-injectors loaded with single doses.



Source: U. S. Census Bureau
Census 2000 Summary File 1
population by census tract.

**Fig. 14.1** North Carolina population density

In this hypothetical example we assume the federal government has allocated 50 of these containers to the state of North Carolina. Antidotes must be administered very soon after exposure to be effective. Therefore, the State Health Department's goal is to locate the ChemPacks so that everyone in the state is as close as possible, and preferably within 50 miles, of a ChemPack. However, a finite number of ChemPacks combined with a geographically distributed population density makes this a challenging problem. Thus, the ChemPacks are limited resources to be located across the geographic region. The potential locations are hospital emergency departments in the state. The customers are potential residents in need of access to a ChemPack. As is common in location models, the customers are aggregated into groups within the geographic network. In the results we present "customers" are aggregated populations defined by zip codes in North Carolina.

We show how location models can be employed to find the optimal placement of ChemPacks within North Carolina. We use population data for the state of North Carolina to formulate a location optimization problem which considers all hospital emergency departments across the state as potential locations for one of the ChemPacks. Four alternative MIPs are compared. We compare the optimal solutions to the alternative MIPs as well as a very easy to implement, and often effective, "greedy heuristic."

In North Carolina, there are 112 emergency departments that are possible locations for ChemPacks (NCDETECT 2009). We collected location and population data at the ZIP Code level for North Carolina, assuming the emergency departments and populations are located at the centroid of the ZIP Code area. Population data for the 783 populated ZIP Codes in North Carolina was extracted from the 2000 US Census (United States Census Bureau xxxx). The population density for North Carolina is illustrated in Fig. 14.1.

**Table 14.1** Four location model formulations are compared based on multiple criteria

| Model | Objective | Average travel distance (miles) | Maximum travel distance (miles) | Total cost ($millions) |
|---|---|---|---|---|
| *p*-median | Minimize the weighted sum of travel distance | 10.58 | 62 | $16.28 |
| UFL | Minimize the total cost without a limit of available ChemPacks | 7.18 | 59 | $11.20 |
| *p*-center | Minimize the maximum travel distance | 41.53 | 59 | $63.62 |
| Maximal covering | Maximize covered population ($A = 50$) | % population covered 99.96 | NA | $35.33 |

To compare the location models, we defined a problem instance and compared the resulting criteria among the models. To create the matrix of travel distances between each population element and emergency departments, we use functions written in Matlab to calculate the geographic distance between two ZIP Codes (population center to emergency department) and multiply it by 1.2 in order to estimate the mean road distance. The weighted demand matrix is generated from ZIP Code census data and is defined as the population residing in each ZIP Code. In the UFL model we use the IRS standard mileage rate for medical purposes ($0.19/mile) as a travel cost estimate, and the estimated storage costs for ChemPacks ($2,000–$2,500) as the cost of locating a ChemPack at an emergency department. In the maximal covering model, where there is a limit to the acceptable travel distance for a population to be covered, we report results for a distance of 50 miles. We assume $p = 50$ as a base case, and we explore the sensitivity of performance measures to $p$.

Table 14.1 summarizes the criteria for each of the location models. Using the base case estimates of costs, the UFL model minimizes each of these measures but may be unrealistic as it requires a ChemPack to be located at every potential location. This result is in part due to the assumed low storage cost of a ChemPack relative to the travel cost. The $p$-median, $p$-center, and maximal covering formulations limit the number of ChemPacks to $p = 50$. While the $p$-median formulation does not achieve the minimum maximum travel distance, as in the $p$-center model, the result is very close (3 miles) and offers significant improvement in the average travel distance over the $p$-center formulation. The maximal covering model results show that a vast majority (99.96%) of the population can be within 50 miles of a ChemPack.

In addition to the four models presented as part of the models in Table 14.1, the greedy heuristic was also analyzed for the scenario with 50 available ChemPacks. The greedy heuristic shares a similar objective with the $p$-median formulation in minimizing the weighted sum of travel distance. The results were also close to

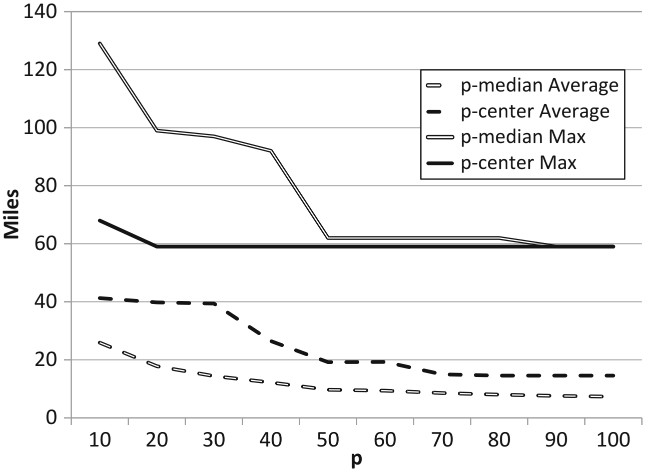


**Fig. 14.2** Sensitivity of average and maximum travel distances to the number of ChemPacks available, *p*

those of the *p*-median formulation with an average travel distance of 9.98 miles, a maximum travel distance of 62 miles, and a total cost of $15.38 million. In general, one can observe that the greedy heuristic provided competitive results through an easy-to-implement algorithm.

Figure 14.2 illustrates the sensitivity of the results to the number of ChemPacks available, *p*, in the *p*-median and *p*-center models. Both the average and maximum travel distance decrease as more ChemPacks are available. The performance measures that are not the primary objectives of each model are more sensitive to *p* than the primary objectives. That is, the maximum travel distance for the *p*-median model and the average travel distance for the *p*-center model are more sensitive to *p* than the average travel distance for the *p*-median model and the maximum travel distance for the *p*-center model. The maximum and average travel distances for both formulations are more sensitive to $p \leq 50$, and the marginal benefits diminish for $p > 50$.

The location models described in this section have a number of limitations. For example, they are deterministic models that assume knowledge of the exact size of the population within the geographic region. The models are uncapacitated in that they assume that a single ChemPack will be sufficient to treat a population in the event of an emergency. Finally, the approximate distribution of the population using aggregation of the population into ZIP Codes may not perfectly represent the true population distribution. In spite of these assumptions, which are common to most applications of location models, these types of models have proven effective in the location of critical resources for emergencies.

## 5 Conclusions

In this chapter we discussed location models in the context of healthcare applications. Location models have the opportunity to play an integral role in healthcare planning in many settings. As evidenced by examples in the literature, new formulations and adaptations of location models allow policy insights to be gleaned in healthcare settings which present unique challenges in the decision making process.

We presented four location model formulations that are commonly employed in location applications. In addition, we discussed the benefits of using heuristics as solution methods and presented the greedy algorithm in the context of the $p$-median location problem formulation. The four model formulations and heuristic were compared using a case study of allocating ChemPacks to hospitals in North Carolina. While each model formulation may be preferred for certain performance criteria, in general, the easy-to-implement greedy heuristic provided a competitive solution across multiple measures.

While location models have an established presence in many application settings, including healthcare, there are many opportunities for future research. One of these future areas is incorporating clinical outcomes and performance measures among the criteria that are used for decision making in location models. For example, in identifying future locations for preventative services, it will likely be important to study the demographics and health status of the populations being considered. If certain communities are expecting to have populations with high utilization levels of such services in the near (or far) future, this weighting needs to be considered in a location model. Including clinical measures in location models will continue to bridge a gap between applying location model methodologies to traditional healthcare planning settings that focus on capital-intensive resources and infrastructure, to medical decision making settings where such models help decision makers in clinical settings. While location models have proven valuable in settings such as prostate cancer treatment planning (Lee et al. 1999), there are many opportunities for new applications in healthcare.

Another direction for further research is including location models in analyses supporting healthcare policy decision making. As focus on preventive care and screening for chronic diseases continues to increase, patient access to such services will continue to play an informative role in identifying cost-effective healthcare policies. Location models can not only aid in the decisions of where preventive care and screening facilities should be located but also help determine the number and size of such facilities in order to provide sufficient capacity required by the demand resulting from policy recommendations.

**Acknowledgements** I would like to thank Brian Denton for his comments and suggestions on this chapter. I would like to thank Lindsay Moomaw and Seth Placke for their contributions to the data analysis for the case study. The research supporting the case study was carried out in part by the North Carolina Preparedness and Emergency Response Research Center (NCPERRC) which is part of the UNC Center for Public Health Preparedness at the University of North Carolina at Chapel Hills Gillings School of Global Public Health and was supported by the Centers for Disease Control and Prevention (CDC) Grant 1PO1 TP 000296.

## References

Chanta S, Mayorga ME, Kurz ME, McLay LA (2011) The minimum p-envy location problem: A new model for equitable distribution of emergency resources. IIE Trans Healthc Syst Eng 1(2):101–115

Côté MJ, Syam SS, Vogel WB, Cowper DC (2007) A mixed integer programming model to locate traumatic brain injury treatment units in the department of veterans affairs: A case study. Health Care Manag Sci 10(3):253–267

Daskin MS (1995) Network and discrete location: Models, algorithms, and applications. Wiley-Interscience, New York

Daskin MS, Dean LK (2005) Location of health care facilities. In: Brandeau ML, Sainfort F, Pierskalla WP, Hillier FS, Price CC (eds) Operations research and health care. International series in operations research & management science, vol 70. Springer, New York, pp 43–76

Drezner Z, Hamacher HW (2004) Facility location: Applications and theory. Springer, Berlin

Eiselt HA, Sandblom CL (2004) Decision analysis, location models, and scheduling problems. Springer, Berlin

Fitch J (2005) Response times: Myths, measurement & management. JEMS: J Emerg Med Serv 30(9):47

Francis RL, Goldstein JM (1974) Location theory: A selective bibliography. Oper Res 22(2): 400–410

Handler GY, Mirchandani PB (1979) Location on networks: theory and algorithms (Vol. 1). The MIT Press

Jia, H, Ordóñez F, Dessouky M (2007) A modeling framework for facility location of medical services for large-scale emergencies. IIE Trans 39(1):41–55

Lapierre SD, Myrick JA, Russel G (1999) The public health care planning problem: A case study using geographic information systems. J Med Syst 23:401–417

Lee EK, Gallagher RJ, Silvern D, Wuu CS, Zaider M (1999) Treatment planning for brachytherapy: An integer programming model, two computational approaches and experiments with permanent prostate implant planning. Phys Med Biol 44:145

Love RF, Morris JG, Wesolowsky GO (1988) Facilities location: Models & methods, vol 7. North-Holland, Amsterdam

NCDETECT (2009) www.ncdetect.org/hospitalstatus.html. Accessed 2 Feb 2009

Nemhauser GL, Wolsey LA (1988) Integer and combinatorial optimization, vol 18. Wiley, New York

Price WL, Turcotte M (1986) Locating a blood bank. Interfaces 16(5):17–26

Rahman S, Smith DK (2000) Use of location-allocation models in health service development planning in developing nations. Eur J Oper Res 123(3):437–452

Revelle C, Bigman D, Schilling D, Cohon J, Church R (1977) Facility location: A review of context-free and ems models. Health Serv Res 12:129–146

Stahl JE, Kong N, Shechter SM, Schaefer AJ, Roberts MS (2005) A methodological framework for optimally reorganizing liver transplant regions. Med Decis Making 25(1):35–46

Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. Oper Res 19:1363–1373

United States Census Bureau. www.census.gov. Accessed 2 Feb 2009

Verter V, Lapierre SD (2002) Location of preventive health care facilities. Ann Oper Res 110(1):123–132

# Chapter 15
# Models and Methods for Improving Patient Access

**Jonathan Patrick and Anisa Aubin**

## 1 Introduction

The goal of most health systems is to ensure access to quality healthcare for all its members through the implementation of policies designed to further that goal. Laudable as this goal is, it is often the case that access to quality healthcare has not been as readily available as we might hope. For example, the Canada Health Act, adopted in 1984, mandated healthcare provision in Canada to be comprehensive, universal, portable, and accessible. While there is debate about what constitutes a "comprehensive" healthcare system, the major cornerstone where the Canadian system is seen to have fallen short of its ideal is in accessibility. While patients have access to the right care, it is often only after a significant period of waiting. This problem is not unique to Canada as many countries report similar struggles with providing timely access to care. For example, in countries such as the USA, where universal healthcare is not provided, there are still issues surrounding lengthy wait times for treatment—especially for those without insurance.

There are now numerous studies that demonstrate long and growing wait lists in a variety of OECD (Organization for Economic Cooperation and Development) countries. Australia, Canada, Denmark, Finland, Ireland, Italy, Netherlands, New Zealand, Norway, Spain, Sweden, and the UK have all reported significant wait time challenges within their healthcare systems (Siciliani and Hurst 2004). One report stated that 38% of patients wait at least 4 months for elective surgery in the UK, 27% in Canada, 26% in New Zealand, and 23% in Australia (Hurst and Siciliani 2003). Another showed that 89% of patients waited more than 3 months for cardiovascular

J. Patrick (✉) • A. Aubin
Telfer School of Management, University of Ottawa,
55 Laurier Avenue, Ottawa, ON, Canada K2G 3A6
e-mail: patrick@telfer.uottawa.ca; rcteam03@aol.com

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_15, © Springer Science+Business Media New York 2013

procedures in the UK, 47% in Canada, and 18% in Sweden (Blendon 2002). The excessive nature of these wait times is highlighted by a Canadian study that surveyed physicians across the provinces and across specialties. Each physician was asked to give both the length of time a patient should expect to wait for treatment as well as their own assessment as to what would be the maximum recommended waiting time. In 88% of the cases (each case representing a different specialty and province), the expected waiting time was significantly longer (Esmail 2005). The same study suggested that overall average waiting time was 92% higher in 2004 than it was in 1993. In a separate study, 19% of patients who were surveyed after having visited a specialist reported having been adversely affected by the length of the waiting time (Sanmartin 2004). Both these studies face the obvious criticism that they are based on surveys and not concrete waiting time data, but they nonetheless highlight the fact that, in the opinion of both the professionals in the system and the patients using it, wait times have reached unacceptable levels.

Basic queuing theory demonstrates that wherever demand and/or service times are stochastic, there is the potential for the development of a queue (see Chapter 2 for a review of queuing theory). It has also demonstrated that small changes in the distributions of the demand and/or the service times can have significant impacts on the nature of the resulting queue, suggesting that health systems are well-served ensuring that they are running in an efficient manner. Of course, if the system is seriously under-capacitated, then improvements in the management of the wait list will fail to have any major impact. Conversely, if the system is seriously over-capacitated, then multiple wait list management policies will likely perform equally well. Thus, the management of wait lists becomes crucial only in that narrow range where capacity is sufficient but only just so. Of course, it is within this narrow range that a health system must function if it is run most efficiently.

Few systems face more variability than healthcare. Demand and/or supply are often highly stochastic. Even for planned procedures such as elective surgeries, it is not the demand that is fixed but the number of scheduled surgeries for a given day. The daily arrivals of new requests for surgery remain a stochastic process. Similarly, very few medical procedures are deterministic in their actual service times. For example, a computed tomography (CT) scan scheduled for 15 min may take anywhere from 10 min to half an hour. This, statistically, is the nature of healthcare where unexpected complications and delays are common place. The result is that queues are ubiquitous throughout the system. Of course, variability can be managed if capacity is increased sufficiently and one is willing to allow for a significant amount of idle time. However, many health systems, on top of dealing with significant variability, are chronically short on capacity as well. Rural areas for instance are most often the hardest hit in terms of a dearth of providers. Thus, the chronic lack of capacity combined with significant variability means that health systems are generally dealing with significant patient access challenges.

Wait lists arise because demand cannot be met immediately, and thus a queue forms. For the purposes of this chapter, the management of the wait list will refer to the process of determining when and in what order those patients in the queue are going to receive treatment. There are, of course, other aspects of wait list management that could be discussed, but the purposes of keeping this chapter succinct we will narrow the scope as outlined.

Operations researchers have applied an impressive array of methodologies in an effort to improve resource management and, therefore, improve accessibility and quality of care. This chapter provides the reader with an overview of the work to date as well as some insight into general policies that have proven useful and further challenges that remain to be tackled. For many health services, the primary method for managing queues is through the appointment schedule, and thus we will often alternate between talking about patient access and determining a scheduling policy.

Patient access decisions can be divided into two stages. The first stage involves the day-to-day decisions as to how many patients to assign for a given procedure each day within a booking window (e.g., the number of days in advance an appointment can be set). This is often referred to as the *advanced scheduling problem*. It is complicated by the common reality that there exist multiple priority classes of patients (differentiated by the urgency of the need for a health service) all competing for a single resource. Thus, the decision facing the resource manager is to look at the current booking slate (number of patients of each priority class booked into each day for the duration of the scheduling window) and determine where to book each patient from each priority class. This problem is even further complicated by the reality that not all patients will necessarily show up for their appointment, and different patients may in fact have different service time distributions and vary in the type and amount of resources they require.

Surgical scheduling is a classic example where there are multiple priority classes within each type of surgery competing for time in the operating rooms as well as time in the hospital. Thus, the optimal policy needs to take into account the stochastic nature of both service times (operating room time and length of stay in the wards) in order to determine the best schedule. Moreover, outside constraints (e.g., surgeon availability) may impose further restrictions on the potential scheduling strategy. This is just one example of how the human element often plays a significant role—whether it be the patient who cannot make the appointed time or the provider who influences the potential form of the scheduling policy. The importance of each potential complication may vary depending on the context. For instance, in outpatient clinic scheduling, no-shows are a major factor, whereas in surgical scheduling, they are not. The form of the objective function may also vary depending on the context. Is it to minimize wait times, maximize profit through ensuring high throughput, or ensuring that the majority of patients in each priority class meet a prespecified wait time target? All of these, as well as others, may be of interest to the various stake holders in the decision-making process that drives patient access.

But even once the decision regarding how many patients to book into a given day has been determined, there remains a second stage of the decision-making—a further scheduling problem regarding how much resource time to allocate to each patient and in what order. This is often referred to as the *appointment scheduling* *problem*. Most formulations of this problem have attempted to set the appointment time of each patient so that some combination of patient wait time, resource idle time, and overtime is minimized. Again, there are complications to the standard problem, including no-shows, resource breakdown or interruption, and the question of how one weights the three components in the objective function.

The combination of an advanced schedule and an appointment schedule provides a complete policy for managing patient access from *decision-to-treat* date to the date of service. In other words, the combination of the two outlines when and in what order patients in the queue will receive service. For those scenarios where multiple resources are used in sequence, a potential third consideration for patient access is patient flow. Poor downstream capacity planning can disrupt patient flow leading to unforeseen cancelations and backlogs disrupting even the most efficient schedules. However, for the purposes of this chapter, we will concentrate on the scheduling issues and not capacity planning. Nonetheless, it is worth noting that in reality, the implemented scheduling policy clearly impacts on the required capacity, and the available capacity will clearly impact on the optimal scheduling policy. We refer the reader to Chap. 7 for a more detailed analysis.

A major stumbling block often faced by those seeking to manage patient access is the inaccurate or incomplete nature of demand data. Demand arrives once a decision to treat has been made, but this date is often not captured, leading to wait times that are usually underestimates. Oftentimes, a patient's total wait time consists of a number of wait times (e.g., time to physician consult, time from physician to specialist consult, time from specialist to actual treatment), all of which are captured in different systems further exacerbating the challenge. Furthermore, data about demand for health services is often censored since most systems record actual appointment dates and times, and not requested dates and times. Finally, it is often problematic to base conclusions on collected wait times simply because organizations differ on what they consider to be the start of the wait time.

Both the appointment scheduling problem and the advanced scheduling problem have been tackled in a wide variety of applications and using a wide variety of methodologies. Applications include surgeries, outpatient clinics, diagnostic imaging, and hospital ward management. Methodologies include stochastic linear programs, approximate dynamic programming (ADP), Markov decision processes, simulation, and queuing theory. We provide an overview of all these attempts at tackling what has proved to be a very challenging and fruitful area of healthcare management for the operations research community.

The remainder of this chapter is organized as follows. Section 2 will provide an overview of the research done to date on appointment scheduling problems detailing the methodologies used, the variations of the problem studied, and the policy insights that can be gleaned. Section 3 does the same for the advanced scheduling problem. Section 4 provides a brief summary of research dealing with patient flow. Finally, we conclude with future challenges and insights in Sect. 6.

## 2 Appointment Scheduling

The appointment scheduling problem seeks to determine the optimal start times for *n* patients who are scheduled for service on a particular day. The goal is often to minimize some combination of server idle time, patient wait time, and overtime. Server idle time occurs whenever a service finishes before the start time of the next patient. Patient wait time occurs whenever the accumulated service times and idle times of previous appointments extend beyond the start time of the current patient. Overtime occurs whenever the prespecified length of day is exceeded. Here, the management of patient access is reduced solely to the problem of scheduling a set number of patients for a particular day.

One of the earliest papers on appointment scheduling is the work of Charnetski (1984) in the context of surgical scheduling. In this chapter, the author looks solely at waiting time and idle-time costs and seeks to minimize a ratio of the two costs by intelligently choosing the start times for each surgery. A simulation model is used to test various scheduling policies that seek to determine the optimal allocation of time for each surgery providing an efficient frontier for the best scheduling policies among the ones tested in the paper.

Ho and Lau (1992) similarly attempt to minimize a combination of physician idle time and patient wait time by testing a number of scheduling rules across a variety of scenarios characterized by the probability of a no-show, the coefficient of variation of service times, and the number of patients per clinical session. They demonstrate that the Bailey–Welch (Ho and Lau 1992) scheduling rule (schedule *n* patients at the beginning of the day and then separate all subsequent appointments by the average service time) does reasonably well in a number of the scenarios.

More recently, Denton and Gupta (2003) provide a stochastic linear programming approach that also solves this classic appointment scheduling problem. Based on a large set of sample paths of potential service times, Denton and Gupta develop a two-stage stochastic linear program to set the start times for a fixed number of services based on a weighted average of idle time, wait time, and overtime costs. One result demonstrated by their model and also earlier by Wang et al. (Wang 1993), is that, provided services times are i.i.d., and wait time and idle time costs are the same for all appointments; the start times exhibit a “dome” shape. That is, start times are more tightly bunched at the beginning of the day and at the end and are more spread out in the middle of the day. The dome structure would seem to accomplish the same goal as the Bailey–Welch scheduling rule (avoiding idle time at the beginning of the day) without imposing such a wait time burden on patients. They also demonstrate numerically instances where scheduling based on mean length of service time is reasonable and highlight the importance of the variability of service time in setting an optimal schedule (Denton and Gupta 2003). Finally, they demonstrate that a heuristic that books patients for service from least to most variable performs very well in a range of scenarios. More recent work by Denton extends the model to a situation where the number of patients scheduled in a day is uncertain due to the arrival of unscheduled appointments.

Gul et al. (2011) look at the impact of variable surgical times. They use a discrete event simulation to evaluate 12 different appointment schedules in an effort to minimize a weighted combination of wait times and overtime. The authors then seek to improve these appointment schedules through a bi-criteria genetic algorithm that takes into account both total patient waiting time and surgical suite overtime. Finally, they expand into the area of advanced scheduling by allowing surgeries to be bumped to another day. They test four different sequencing schemes: *increasing mean of procedure time, decreasing mean of procedure time, increasing variance of procedure time*, and *increasing coefficient of variation of procedure time*. Patients are separated by the estimated procedure time of the previous patient (with that estimation varied over various percentiles of the distribution). They conclude that the *longest processing time first* leads to high expected overtime while the *shortest procedure time first* appears to perform well in a number of settings.

Begen and Queyranne (xxxx) provide an integer programming approach to the appointment scheduling problem. Initially they assume that the distribution of processing times is integer and follows a known discrete probability distribution. They demonstrate that an optimal solution can be found in polynomial time. They then relax the assumption of a known discrete probability distribution and determine the necessary sample size of surgical times to obtain provably near-optimal solutions with high probability. While they provide an analytical model that can be solved for specific instances, they do not provide any generalizable policy insights.

Green et al. (2006) look at a scheduling problem for diagnostic imaging that is somewhat of a departure from the above-cited papers on appointment scheduling. Here, the appointment schedule is already set (a fixed number of outpatients are scheduled uniformly throughout the day) so that the issue instead is whether to serve a scheduled outpatient or whether to give the current service slot to a waiting emergency patient or inpatient. Decisions are made based on the number of waiting outpatients, inpatients, and emergency patients, and the optimal policy is found through formulating this as a stochastic dynamic program. They demonstrate that the optimal policy lies in the set of *monotone switching curve policies* where outpatients are served as long as the number of waiting outpatients exceeds a certain threshold and where that threshold is a function of the number of waiting inpatients. As the number of waiting inpatients increases, the threshold naturally increases as well.

Since the form of the optimal policy in their model is complex, Green et al. further tested the sensitivity of a number of heuristic appointment policies using a simulation model. They looked at service (i.e., the ordering of patients) and appointment policies separately. The service policies include critical-first and a linear approximation heuristic based on their model. (This heuristic essentially simplifies the original policy to allow for easier implementation.) Appointment policies include *fill all slots*, *balanced*, *news-vendor*, and *fill alternate slots*. The research suggests that if a threshold appointment schedule is used, then *fill all slots* and critical-first priority rules result in the best capacity management for the majority of scenarios. Waiting costs as an end-of-day penalty refer to the costs applied when an inpatient is not served by the end of the day, or in the case of an outpatient, the overtime associated with examining a patient after hours. When the waiting costs and revenue are close, then the linear approximation heuristic is the better priority rule. If patient service takes priority, then the balanced heuristic is preferable to the *fill all slots*. If waiting costs are high, the *fill alternate slots* policy appears to be the most profitable (Green et al. 2006).

Another important aspect of appointment scheduling is the sequencing of patients on the day of service. Sequencing heuristics, in the context of surgical scheduling, such as shortest case first (SCF), longest case first (LCF) and first-come-first-served (FCFS) or first in first out (FIFO) were investigated by Dexter (Dexter and Marcon 2006). Although LCF is most commonly used in practice, Dexter suggests that it performs very poorly. Random sequencing although trivial to implement yields poor results in some cases (Gul et al. 2011). However, other research suggests that it is not the length of service that should determine the sequence of patients but rather the variability in service. Scheduling patients from the least variable to the most variable is a good heuristic in some cases (Denton and Gupta 2003). Of course, this may be difficult to implement in practice as it often equates to leaving the more complex cases to later in the day which may not be advisable—depending on the particular type of health service under investigation. Another possibility is to schedule patients based on the ratio of variation in service time to wait time cost. However, as wait time costs are difficult to quantify, such a policy is less easily justified.

The advantage of the appointment scheduling problem is that it has a neatly contained scope, and yet, nonetheless, analytical solutions and policy insights have largely been contained to simplified models that ignore key components such as the prioritization of patients and the variability in service distributions between patient classes. Such models are motivated by the goal of improving efficiency on a particular day of service. This may ultimately translate in an increase in the number of patients that can be scheduled on a particular day and therefore a decrease in the waiting time from the requested date of service to the scheduled date of service. However, most appointment scheduling models do not explicitly consider this.

## 3 Advanced Scheduling

This section breaks down advance scheduling problems into those composed of scheduling a single day, and problems of scheduling multiple days in advance.

### *3.1 The Single-Day Scheduling Problem*

The advanced scheduling problem is broader in scope than the appointment scheduling problem. We will first discuss the restricted problem of scheduling a single day (or period) before discussing the multi-period scheduling problem. The single-day advanced scheduling problem somewhat blurs the line between appointment scheduling and advanced scheduling (since it does not deal with the issues arising from day-to-day scheduling) but more aptly fits under advanced scheduling as these models assume the appointment times have been set while the number of patients booked into a day is flexible.

Green and Savin (2008) use a queuing approach to investigate the scheduling of patients for physician appointments. They assume equal appointment lengths and seek to determine the panel size (e.g., number of patients per physician) that can reasonably be handled given a fixed rate of demand associated with each patient on the panel. The model is quite stylized reflecting its high-level goal of capacity planning. They take into account the reality that patients do not always show up to a scheduled appointment thus allowing physicians to carry a larger panel than they might otherwise. Their model includes backlog-dependent cancelation rates and provides upper and lower bounds on the size of the backlog for large patient panels.

Another paper that relaxes the assumption of a fixed number of appointments for a given day is the stochastic model developed by Muthuraman and Lawley (2008). They look at a single day's bookings and seek to determine the necessary overbooking in order to offset the probability of no-shows. Here, the appointment times are fixed, but it is up to the clinic manager to determine when to stop accepting patients for that day. They seek to minimize an objective that takes into account revenue generated by each patient seen as well as wait times. This work is extended to the case where patients may have different no-shows rates in Zeng et al. (2009).

Kim and Giachetti (2006) solve a very similar problem by developing a stochastic overbooking model that seeks to maximize revenue by using overbooking to compensate for no-shows while also incorporating walk-ins. They too do not seek to determine actual start times for each appointment but rather to set the optimal number of appointments that should be accepted each day. They demonstrate that their model can do significantly better than one that simply adds the mean number of no-shows minus the mean number of walk-ins to the number of appointments to accept.

Gupta and Wang (2008) incorporate into their model the additional complication of patient choice. They build a Markov decision process (MDP) model that seeks to determine the number of advanced appointments to accept for a specific day given a known distribution of same-day demand. Patients call in with specific requests for a particular time of day and physician. The clinic's decision is simply to look at the current slate for that day and determine whether to accept or reject that request. They demonstrate that for a single physician's clinic the optimal policy is to allow patients to book any one of the unreserved slots up to a booking limit.

One final paper in this vein is the work of LaGanga and Lawrence (2007) who use a simulation model to demonstrate the value of overbooking in a variety of settings for an outpatient clinic where the trade-off is between patient wait times and overtime. They again ignore the issue of start times by assuming that appointments are evenly spread out throughout the day. Their model demonstrates that the utility of overbooking is dependent on the size of the clinic and the no-show rate.

## 3.2 The Multi-day Advanced Scheduling Problem

All the above research considers a single day and seeks either to determine the start times for a fixed number of appointments or else seeks to determine the optimal number of appointments to schedule into fixed appointment lengths. Concentrating on objectives that deal only with rewards or costs associated with a single day (e.g., idle time, overtime, and wait time) neglects the fact that for most healthcare operations, demand that is unmet on one day must be met on another.

The issue of scheduling patients in a multiday context has received much less attention in the literature with one notable exception—the scheduling policy for outpatient clinics proposed by Murray and Tantau (1999) called Advanced Access or open access. The idea behind this policy is to combat the real issue of no-shows by attempting to serve *today's demand today* thus negating the need for any day-to-day wait list management as demand does not spill over from one day to the next. (This is a bit of an extreme description of open access as in reality some appointments, such as follow-up appointments, have to be scheduled in advance, but the goal of the policy is to keep these to a small percentage of the whole.) Under the open access policy, clinics are largely left unscheduled, and clients call in at the beginning of the day to book an appointment for later that same day. This simple scheduling policy has generated a surprising amount of research that has sought either to validate it or suggest alternatives.

One such attempt to validate open access is the work of Kopach et al. (2007) where they develop a simulation model to determine the impact of four clinic characteristics on the successful implementation of open access—namely (1) the fraction of patients being served on open access, (2) the scheduling horizon for patients scheduled in advance, (3) the existence of provider care groups, and (4) the permissibility of overbooking. They demonstrate the potential for open access to improve throughput but caution that too aggressive an implementation of open access may harm the continuity of care.

Robinson and Chen (2009) use a linear programming model that seeks to minimize a combination of patient waiting time, physician idle time, and overtime to demonstrate that, for clinics experiencing a significant no-show rate, same-day booking outperforms a traditional advanced booking policy with no overbooking. Moreover, they show that the added flexibility of being able to defer some demand till tomorrow improves on a strict adherence to open access.

Patrick (2012) uses a MDP model to demonstrate that, for a clinic primarily interested in maximizing revenue, a short booking window with the potential for overbooking can do as well as open access and with much more predictable throughput thus requiring less capacity. Moreover, for a clinic primarily concerned with resource efficiency, the policy developed by the MDP model outperforms open access in a wide variety of scenarios. The policy developed is described as *delayed open access*. It fills remaining capacity in a given day's schedule with any new demand but resists overbooking unless the queue size reaches a derived threshold(s). At that point, it will begin to overbook. Depending on the scenario, the policy may then simply revert to open access and book any additional demand into today or it may book only one additional client today and then delay again until another threshold triggers a second overbook. The model is a simplification of reality in that it ignores stochastic service times and thus treats overtime as simply the number of appointments above the capacity limit.

Perhaps the most realistic model for outpatient scheduling that has been developed to date is the work of Liu et al. (2009). They build a MDP model that tracks the number of days in advance each patient is booked and uses a no-show distribution that is dependent on the number of days between the booking date and the scheduled date. They also allow for advance cancelations that free up reusable capacity. The downside of such a complete model is that it is no longer solvable. The authors therefore develop heuristic policies based on applying a single step of the policy iteration algorithm starting with a “good” initial solution. In line with other work, they demonstrate that a 2-day booking window outperforms open access and that they can improve on an initial solution by applying a single step of policy iteration.

What the majority of the above research suggests is that a short booking window with overbooking will outperform open access. Liu et al. (2009) as well as Robinson and Chen (2009) demonstrate the value of a 2-day booking window, while Patrick (2012) demonstrates that the optimal booking window may stretch out to 5 days depending on the variability in demand, the ratio of demand to capacity, and the severity with which the no-show rate increases as the appointment lead time lengthens.

The only paper described above that explicitly captures patient wait times over days is the Liu et al. (2009) paper. This is unsurprising as tracking patient wait times creates challenges of problem size that make many methodologies intractable. As a result, Liu et al. (2009) resorted to a heuristic approach. The challenge is that in the advance scheduling problem, the decision depends on the number of appointments already booked into each day of the booking horizon. This yields a state space that is computationally prohibitive either for a linear program or a MDP model.

One methodology that has been applied to the multi-period advanced scheduling problem and that seeks to overcome this issue of problem size is that of ADP. Schutz and Kolisch (2010) provide an advanced scheduling model for a single future day for a diagnostic imaging department that seeks to maximize revenue by selectively choosing to accept or reject requests for service over the course of the booking window. Similar to Liu et al. (2009), they allow for no-shows and advance cancelations but also incorporate variable services times. To derive a solution, they implement a simulation-based approach to ADP developed by Gosavi (2004) that allows them to formulate a near-optimal policy. They then extend their work to a multiday problem where booking occurs over a period of days instead of just one.

Patrick et al. (2008) study a similar problem where patients of varying priority classes are scheduled into a booking window of available treatment days. The goal is to ensure that the majority of patients are seen within priority-specific wait time targets. The authors develop a MDP model translated into the equivalent linear programming form. They then solve the linear program under the restriction that the value function is assumed to take a specific form. This restriction is required in order to make the linear program tractable. The chosen form of the approximation is, of course, crucial to the success of the resulting policy and is an area of active research in ADP. The authors provide a solution that is easily translated into a readily implementable policy and derive bounds on the necessary capacity for this policy to function well. The derived policy books the highest priority class into the first available slot, while lower priority classes are given booking windows (with the upper bound of the window being the wait time target) within which a patient of that priority class can be booked. If no space is available in the booking window for a given patient, then overtime is used to alleviate the stress on the system. Overtime is also used if there is no available space for the highest priority class within its wait time target. This policy is coupled with a capacity plan (how much overtime capacity to anticipate given a certain base capacity) in order for the policy to function well.

Saure et al. (2011) extend the work of Patrick et al. (2008) to the scenario of radiation treatment scheduling. This scheduling problem is complicated by the fact that scheduling a single patient does not equate to one appointment but rather to a series of appointments, all of which are not necessarily the same length. They too implement the linear programming approach to ADP in order to derive a near-optimal scheduling policy. Conforti et al. (2010) also look at radiation treatment scheduling. Rather than resort to an approximation scheme, they simplify the problem by developing a linear program that seeks to minimize mean wait time or maximize throughput. This allows them to avoid tracking actual wait times.

It is clear from the above literature review that the multiday advanced scheduling problem is typically more challenging than the single-day advanced scheduling problem. As a result, research in this area has resulted in advances in the area of heuristics and approximation methods.

In this subsection, we have concentrated on applications to outpatient services primarily. There are also a number of papers that deal with the advanced scheduling of surgeries. We refer the reader to the review by Gupta (2007) and Chap. 5 of this book.

## 4 Patient Flow

Systems where patients move through a number of services from first entry into the health system to exit can present significant challenges to patient access. A lack of downstream capacity can result in bottlenecks that create a ripple effect on all upstream services. Examples within the acute care setting include the cancelation of surgeries due to a lack of beds and the escalation of wait times in the emergency department due to the inability to readily move admitted patients to the wards. This last example is in turn the result of the inability to readily move patients out of the acute setting and into more appropriate facilities in the community once the acute phase of their treatment is done.

Problems where multiple resources are consumed in sequence are understandably quite complex and therefore difficult to model. Often, the only feasible methodology is discrete event simulation, or a very stylized queuing model. Still, some attempts have been made of which we provide a sample as illustrative of the work done in this area.

Oddoye et al. (2009) formulate a goal programming model with simulation in an attempt to eliminate bottlenecks impeding optimal clinical work flow. The model was successful in determining the impact of disruptions such as staff sick days or an increase/decrease in beds. The goal programming model sought to balance optimization and equity of allocation (Oddoye et al. 2009). The model provides users with a capacity estimate given predetermined costs associated with each additional bed and the penalty associated with increased wait time for patients.

Koizumi et al. (2005) develop a queuing network model that tracks the flow of psychiatric patients through a series of treatment units depending on the severity of their condition. They develop a methodology for extending service times in order to account for the reality that patients may remain in a facility longer than necessary due to a lack of resources at the more appropriate facility. Their model helps identify where increases in capacity might be most beneficial.

Thompson et al. (2009) built a MDP model that seeks to determine when it is optimal to transfer patients between wards in an attempt to anticipate demand surges and thus prevent blockages. Due to the complexity of the model, they resort to random sampling in order to generate an approximate solution that they demonstrate in practice results in a more efficient use of resources. The authors demonstrate that a policy that transfers patients in anticipation of demand surges can both improve revenue and reduce the time for an admitted patient to receive a bed.

Patrick (2011) develops a MDP model that determines the necessary access for a hospital to maintain the census of patients waiting for long-term care below a given threshold. This model is augmented with a simulation model that can be used for capacity planning in long-term care in order to meet both the threshold census levels for multiple hospitals in a region as well as wait time targets for those clients who apply for long-term care directly from the community. Again, the link to wait list management is that such capacity planning downstream is a pre-requisite for the success of the scheduling policies that more immediately impact on the management of the wait list.

As these articles demonstrate, the complexity of multiple stages of service means that methodologies tend to depend on approximations and/or heuristics to generate solutions. Patient flow is thus one of the areas in wait list management that remains most challenging.

## 5 Case Study: Managing Patient Access to Diagnostic Imaging

In this section, we provide a case study from one of the authors past work on patient access to diagnostic imaging (Patrick et al. 2008) and specifically for scheduling patients for CT scans. Diagnostic imaging is complicated by the presence of multiple priority classes competing for a single resource. Lower priority demand is scheduled in advance, but if that is not done well, it can lead to either unacceptable wait times for later-arriving higher priority demand or else high levels of overtime. Thus, the balancing act is between insuring that wait time targets are met for all priority classes while minimizing, as much as possible, the use of overtime.

The model developed in Patrick et al. (2008) assumes that there are $I$ different priority classes and that service times are deterministic and of equal length. This is a simplification as, in reality, patients are assigned to 15, 30, 45, or 60 min slots. However, since all assigned lengths are multiples of 15, demand can be viewed as the number of requests for 15 min slots with a patient requiring a 30 min scan being viewed as demand for 2 slots. The model has actually been adapted to incorporate different scan lengths with predictable results that will be discussed below. Here, we present the simpler model.

Since this is a natural sequential decision problem, it was modeled as a infinite horizon, discounted MDP model. A MDP model bases a periodic decision or action on a set of information called the state. Demand arrives according to a known distribution and is collected each day, prioritized by a radiologist and returned to the booking clerk for scheduling. The state of the system, denoted as $\mathbf{s} = (\mathbf{x}, \mathbf{y})$, (i.e., the information upon which a decision is based) consists of the current booking slate, $\mathbf{x} = (x_1, \ldots, x_N)$, where $x_n$ represents the number of scanning slots already booked on day $n$ as well as the waiting demand, $\mathbf{y} = (y_1, \ldots, y_I)$, where $y_i$ represents the number of patients of priority $i$ waiting to be booked. Based on this information, the booking clerk must determine an action, $(\mathbf{a}, \mathbf{z})$, where the components of $\mathbf{a}$, $a_{in}$, represent how many of the waiting priority $i$ patients to book on day $n$ and the components of $\mathbf{z}$, $z_i$, represent how many priority $i$ patients to serve through overtime. Costs, represented by $c(\mathbf{a}, \mathbf{z})$, are incurred if demand is not booked (delay cost), if demand is booked further out than a priority-specific wait time target, or if overtime is used to satisfy demand.

An MDP model is solved by determining the value function, $V^{\pi}(\mathbf{s})$, which is a policy-specific vector that represents the total discounted cost from starting in state $\mathbf{s}$ and using a policy $\pi$ to govern future actions. The policy $\pi$ defines what action to take in every state. The cornerstone of the MDP approach is the Bellman equation written as:

$$\max_{\mathbf{a}} \left\{ c(\mathbf{s}, \mathbf{a}) + \lambda \sum_{\mathbf{s}' \in S} p(\mathbf{s}'|\mathbf{s}, \mathbf{a}) V(\mathbf{s}') \right\} \quad \forall \mathbf{s} \in S \tag{15.1}$$

This equation balances the cost, $c(\mathbf{s}, \mathbf{a})$, of taking a given action, $\mathbf{a}$, in a given state, $\mathbf{s}$, against the expected future costs associated with that action. The major challenge in solving an MDP is that, for applications such as the one described above, the computational costs of solving the Bellman equations and finding the optimal policy are prohibitive due to the large size of the state space. For instance, if the booking clerk is booking over a 30-day window and there are only 10 scanning slots in a day, then there are $30^{10}$ different configurations of $\mathbf{x}$. It thus becomes impossible to determine the value function for every possible state for any policy. This is referred to as the "curse of dimensionality."

One possible solution to this dilemma is to assume that the value function has a certain parametric form and solve over that reduced space instead. This approach is often referred to as approximate dynamic programming or ADP. For instance, in the case study presented here, we assumed that the value function had the form:

$$V(\mathbf{s}) = V_0 + \sum_{(i,n)} V_{in}x_{in} + \sum_{i} W_i y_i \tag{15.2}$$

where $V_0, \mathbf{V}, \mathbf{W}$ are tunable parameters. Thus, the value function, $V(\mathbf{s})$, is assumed to be a linear function of the number of appointment slots already booked and the number of patients waiting to be booked. There are a number of methods available for choosing the best parameter values for the value function approximation with the two broad streams being a simulation-based approach or a linear programming approach (Bertsekas and Tsitsiklis 1996; Powell 2011). There is, however, no methodology for determining what is the best form for the value function approximation. Quite often, an approximation that is linear in the state vector, as above, works well, but it is certainly likely that there are other more sophisticated approximation schemes that would work better. In this work, we used a linear programming approach to ADP to determine the optimal linear approximation to the value function. The full details of the formulation can be found in Patrick et al. (2008). Here, we present some results to illustrate some important aspects of patient access.

For each priority class, the policy gives a booking window in which clients of that priority can be booked. The upper bound on that window is the wait time target, while the lower bound depends on a number of factors including the demand rates for higher priority classes. Clients from each priority class are booked into the appropriate window starting from the lower bound to the upper bound for the highest priority class but starting from the upper bound and moving forward to the lower bound for all other priority classes. This gives the booking clerk the greatest flexibility in terms of insuring capacity is available for the highest priority class while still meeting the targets of the lower priority classes. The solution to the model also gives a threshold priority class such that if a patient is from a priority class that is classified as more urgent than the threshold priority class and there is no scanning slot available in the appropriate booking window, then that client is serviced through overtime. (Should there not be sufficient overtime to service all eligible clients, then the lower priority clients booking decision is delayed.) All other clients simply have their booking decision delayed should there be no capacity available in the appropriate booking window.

In addition, the ADP approach yielded a capacity constraint that specified the necessary available overtime capacity for a given regular-hour capacity in order for the above policy to function well. This is a surprising result in that there is no guarantee that an ADP model for scheduling would give any insight into the required capacity. However, the nature of the conditions required for proving the

**Table 15.1** A comparison of AOP and the booking limit policy for a small outpatient clinic with 95% confidence intervals

| Criteria | Priority class | Approximate optimal policy | Booking limit policy BL = (1,7,9) |
|---|---|---|---|
| Percent late | P1 | (0.18,0.26) | 0 |
| | P2 | 0 | (0.25, 0.59) |
| | P3 | 0 | (47.4, 48.16) |
| | Overall | (0.09,0.13) | (9.56,9.82) |
| Percent Served through overtime | P1 | (1.49,1.63) | 0 |
| | P2 | 0 | 0 |
| | P3 | 0 | (20.19, 21.75) |
| | Overall | (0.71,0.85) | (4.04,4.36) |
| Utilization percentage | | (98.97,99.13) | (95.59,95.87) |

Note that P1 refers to priority class 1, P2 to priority class 2, and P3 to priority class 3

optimal form of the approximation provided a lower bound on overtime capacity that nicely links the capacity planning problem and the scheduling problem together. Both problems clearly impact on patient access and are generally treated in isolation when in reality they are very much related.

The paper in Patrick et al. (2008) provides simulation results for running the derived policy from the ADP model in four different settings—small and large outpatient clinics and two different-sized hospitals including one based on data from a hospital in Vancouver. The largest-sized problem consisted of a hospital with the capacity for 126 15 min slots per day. Current practices are rather difficult to simulate. For comparison purposes, we consider a strict booking-limit policy that will not book a lower priority patient into a given day unless the number already booked is below a threshold. For a small-sized clinic, such a booking-limit policy can be determined by enumeration. Table 15.1 provides the comparison between the approximate optimal policy (AOP) and the optimal booking-limit policy for a small outpatient clinic with three priority classes. The results demonstrate how the AOP trades off a small increase in overtime use for the highest priority class in order to greatly improve the timely access of the lower priority classes. The intuition is that by using overtime in a judicious fashion to service the highest priority class, the booking manager can avoid the congestion that leads to lengthy wait times for the lower priority classes. Thus, the policy acts proactively to avoid congestion rather than reacting to congestion that has already occurred.

The above case study demonstrates the significant impact on patient access of an intelligent advanced scheduling policy. Poor scheduling can lead to unnecessarily lengthy delays in treatment and significant variation in wait times even within the same priority class. In contrast, the kind of policy derived from the above model provides equitable and timely access while using resources as efficiently as possible.

## 6 Conclusion and Future Research

While there now exist some excellent models for the appointment scheduling problem, there remain some serious computational challenges that make solving realistic instances challenging. Exact solutions are often too time-consuming, and thus approximation methods come into play. However, with these methods, we need some means of determining a bound on the optimality gap between our approximate solution and the true optimal solution. This is a serious issue in the ADP literature where the bounds that have been developed to date are too loose to be of much practical value. There are also complications to the appointment scheduling problem that need to be addressed. What, for instance, is the impact of interruptions on the form of the optimal policy? Does the dome shape still hold if interruptions are known to follow a nonstationary distribution? Surgical scheduling, as an example, is often interrupted by the arrival of an emergency surgery, and these arrivals are much more likely at certain times of the day. Surely, this may impact on whether a dome-shaped policy remains optimal? Another issue is that much of the literature assumes i.i.d. service times which is not true in a number of settings and would impact on the form of the optimal policy.

In advanced scheduling there are perhaps even more open challenges. While the model in Patrick et al. provides an easily implementable policy, it is unlikely to be optimal due to its simple approximation architecture. Can a better approximation architecture lead to a better policy? Schutz et al. (Schutz and Kolisch 2010) improve on the Patrick (Patrick et al. 2008) model by incorporating no-shows and stochastic service times, but their policy loses any easily implementable form and in fact is counterintuitive at times perhaps because of the limitations of the simulation-based approach. Thus, there would seem to be plenty of opportunity for improving both policies. In addition, neither policy has yet been implemented, and thus an observational study of the impact of such policies in practice would be an interesting avenue of research.

Finally, perhaps the most obvious area where continued challenges lie is in patient flow. The movement of patients within an acute hospital is extremely complex and driven by any number of uncertain events. Unexpected complications, bed blockages downstream, and staffing shortages are just some of the factors that come into play. While queuing models can help, they are necessarily high level in order to maintain tractability. Simulation models, on the other hand, have tended to get bogged down in details that make generalizable insights next to impossible. Moreover, neither methodology optimizes patient flow leaving it to the ingenuity of the modeler to determine potentially useful policy changes.

It is encouraging that many health systems have begun to recognize the potential for operations research to improve healthcare management. The challenge that patient access management presents to the operations research community provides an abundant opportunity for researchers and participants alike to have an impact on patient access thus serving to improve the quality of care that lies at the heart of well-functioning health systems.

## References

Begen M, Queyranne M Appointment scheduling with discrete random durations. Math Oper Res, 36(2):240–257

Bertsekas D, Tsitsiklis J (1996) NeuroDynamic programming. Athena Scientific, Belmont

Blendon R (2002) Inequities in health care: A five country survey. Health Aff 21:182–191

Charnetski J (1984) Scheduling operating room times with early and late completing penalty cost. J Oper Manag 5:91–102

Conforti D, Guerriero F, Guido R (2010) Non-block scheduling with priority for radiotherapy treatments. Eur J Oper Res 201:289–296

Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. IIE Transactions 35:1003–1016 (2003)

Dexter F, Marcon E (2006) Impact of surgical sequencing on post anesthesia care unit staffing. Health Care Manag Sci 9:87–98

Esmail N (2005) Waiting your turn: hospital waiting lists in Canada. Fraser Institute, Vancouver. Available at www.fraserinstitute.ca

Gosavi A (2004) Reinforcement learning for long-run average cost. Eur J Oper Res 155:654

Green L, Savin S (2008) Reducing delays for medical appointments: A queuing approach. Oper Res 56:1526–1538

Green L, Savin S, Wang B (2006) Managing patient service in a diagnostic medical facility. Oper Res 54:11–25

Gul S, Denton B, Fowler J, Huschka T (2011) Bi-criteria scheduling of surgical services for an outpatient procedure center. Prod Oper Manag 20:406

Gupta D (2007) Surgical suites operations management. Prod Oper Manag 16:689–700

Gupta D, Wang L (2008) Revenue management for a primary-care clinic in the presence of patient choice. Oper Res 56:576–592

Ho C, Lau H (1992) Minimizing total cost in scheduling outpatient appointments. Manag Sci 38:1750–1764

Hurst J, Siciliani L (2003) Tackling excessive waiting times for elective surgery: A comparison of policies in twelve oecd countries. OECD Health Working Papers, 6

Kim S, Giachetti R (2006) A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. IEEE Trans Syst Man Cybern-Part A: Syst Hum 36:1211–1219

Koizumi N, Kuno E, Smith T (2005) Modeling patients flows using a queuing network with blocking. Health Care Manag Sci 8:49–60

Kopach R et al (2007) Effects of clinical characteristics on successful open access scheduling. Health Care Manag Sci 10:111–124

LaGanga L, Lawrence S (2007) Clinic overbooking to improve patient access and increase provider productivity. Decis Sci 38:251–276

Liu N, Ziya S, Kulkarni V (2009) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. Manuf Serv Oper Manag 12:347–364

Murray M, Tantau C (1999) Redefining open access to primary care. Manag Care Q 7:45–51

Muthuraman K, Lawley M (2008) A stochastic overbooking model for outpatient clinical scheduling with no-shows. IIE Trans 40:820–837

Oddoye J, Jones D, Tamiz M, Schmidt P (2009) Combining simulation and goal programming for healthcare planning in a medical assessment unit. Eur J Oper Res 193:250–261

Patrick J (2011) Access to long term care: The true cause of hospital congestion? Prod Oper Manag 20:347–358

Patrick J (2012) Clinic scheduling: Balancing arrival certainty and booking flexibility. Health Care Manag Sci 15:91–102

Patrick J, Puterman M, Queyranne M (2008) Dynamic multi-priority patient scheduling for a diagnostic resource. Oper Res 56:1507–1525

Powell W (2011) Approximate dynamic programming. Wiley, Hoboken

Robinson L, Chen R (2009) Traditional and open-access appointment scheduling policies: The effects of patient no-shows. Manuf Serv Oper Manag 12:330–346

Sanmartin C et al (2004) Access to health care services in Canada, 2003 (2004). Statistics Canada, Catalogue 82-575-XIE

Saure A, Patrick J, Tyldesley S, Puterman M (2011) Dynamic multi-appointment patient scheduling for radiation therapy. Eur J Oper Res 223(2):573–584

Schutz H, Kolisch R (2010) Capacity allocation for magnetic resonance imaging scanners. Working Paper

Siciliani L, Hurst J (2004) Explaining waiting-time variations for elective surgery across oecd countries. OECD Econ Stud 38 (2004)

Thompson S, Nunez M, Garfinkel R, Dean M (2009) Efficient short term allocation and reallocation of patients to floors of a hospital during demand surges. Oper Res 57:261–273

Wang P (1993) Static and dynamic scheduling of customer arrivals to a single-server system. Nav Res Logist 40:345–360

Zeng B, Turkcan A, Lin J, Lawley M (2009) Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. Ann Oper Res 178:121–144

# Chapter 16
# Coordinating Health Services: An Operations Management Perspective

**Thomas R. Rohleder, David Cooke, Paul Rogers, and Jason Egginton**

## 1 Introduction

The rising costs of healthcare in the USA and around the world are well documented. A recent report from the Commonwealth Fund (Davis et al. 2010) showed that costs throughout the world have been rapidly increasing in recent decades. Driving these cost increases are new technologies, tests, patient access, inefficiencies, and myriad other factors. Such cost increases might be justified if the healthcare delivered was of higher quality. However, this is highly questionable. For example, in the previously referenced report by Davis et al., the USA was last of seven major developed countries in overall health system performance (see Fig. 16.1). Yet, the cost of healthcare in the USA is nearly double the cost of the top performer, the Netherlands.

The Commonwealth Fund Report has several different categories that make up health system performance. One of the components of the Quality Care category is Coordinated Care. Based on the rankings in Fig. 16.1, it appears that this component may have some influence on both the quality of care and healthcare expenditures. In the context of the Commonwealth Report, Coordinated Care focuses on how well care is managed among health services from a medical perspective. However, from an operations research perspective, coordination is also about having the right

T.R. Rohleder (✉) • J. Egginton
Division of Healthcare Policy and Research, Department of Health Sciences Research,
Mayo Clinic, Rochester, MN, USA
e-mail: Rohleder@mayo.edu

D. Cooke
Cooke Research & Consulting Inc., Calgary, AB, Canada

P. Rogers
Department of Mechanical Engineering, University of Calgary, Calgary, AB, Canada

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_16, © Springer Science+Business Media New York 2013

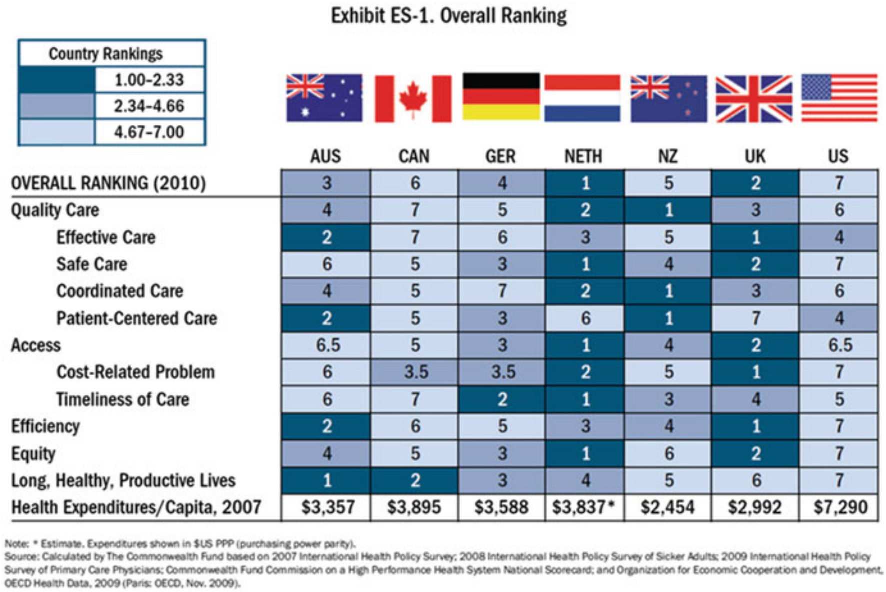


Exhibit ES-1. Overall Ranking

| Country Rankings | |
|---|---|
| ■ | 1.00–2.33 |
| ▨ | 2.34–4.66 |
| □ | 4.67–7.00 |

| | AUS | CAN | GER | NETH | NZ | UK | US |
|---|---|---|---|---|---|---|---|
| OVERALL RANKING (2010) | 3 | 6 | 4 | 1 | 5 | 2 | 7 |
| Quality Care | 4 | 7 | 5 | 2 | 1 | 3 | 6 |
| Effective Care | 2 | 7 | 6 | 3 | 5 | 1 | 4 |
| Safe Care | 6 | 5 | 3 | 1 | 4 | 2 | 7 |
| Coordinated Care | 4 | 5 | 7 | 2 | 1 | 3 | 6 |
| Patient-Centered Care | 2 | 5 | 3 | 6 | 1 | 7 | 4 |
| Access | 6.5 | 5 | 3 | 1 | 4 | 2 | 6.5 |
| Cost-Related Problem | 6 | 3.5 | 3.5 | 2 | 5 | 1 | 7 |
| Timeliness of Care | 6 | 7 | 2 | 1 | 3 | 4 | 5 |
| Efficiency | 2 | 6 | 5 | 3 | 4 | 1 | 7 |
| Equity | 4 | 5 | 3 | 1 | 6 | 2 | 7 |
| Long, Healthy, Productive Lives | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Health Expenditures/Capita, 2007 | $3,357 | $3,895 | $3,588 | $3,837* | $2,454 | $2,992 | $7,290 |

Note: * Estimate. Expenditures shown in $US PPP (purchasing power parity).
Source: Calculated by The Commonwealth Fund based on 2007 International Health Policy Survey; 2008 International Health Policy Survey of Sicker Adults; 2009 International Health Policy Survey of Primary Care Physicians; Commonwealth Fund Commission on a High Performance Health System National Scorecard; and Organization for Economic Cooperation and Development, OECD Health Data, 2009 (Paris: OECD, Nov. 2009).

**Fig. 16.1** Report on healthcare system performance: Commonwealth Fund (2010)

quantities of healthcare resources and how well they are managed together. This latter perspective of coordination affects the former by ensuring patients have access to the needed healthcare resources and flow through them effectively and efficiently.

Why is this type of coordination important? There are many reasons, including avoiding duplication of services. Without coordination, the same information from patients may be gathered at several points, tests may be repeated, and redundant health services may be provided, resulting in higher cost and burden to patients. In addition, a lack of coordination may lead to delays in patient treatment. Chronic shortages of key resources may lead to bottlenecks and long waiting times. As a way to avoid waiting, patients may use (or be referred to) inappropriate and more costly resources. For example, shortages of primary care capacity often lead to increased use of emergency departments and similar more costly and inappropriate resources (Cheung et al. 2011). Thus, poorly coordinated healthcare resources may lead to poor quality and higher costs.

The various symptoms of a lack of coordination may lead to lower satisfaction for patients and result in less than optimal treatment. Potentially, deadly errors could occur if patients receive conflicting medications or treatments. Poor coordination is one of the issues discussed in the report "To Err is Human: Building a Safer Health System" (IOM 1999) that notes "One oft-cited problem arises from the decentralized and fragmented nature of the healthcare delivery system–or 'nonsystem,' to some observers."

Lack of coordination of healthcare resources has been considered from many perspectives. Clinicians look at the issue from the perspective of the frontline healthcare provider and consider improved communication and information sharing as a key to improvement (Sucliffe et al. 2004). Related to better communication is the push for better patient record sharing and management, particularly better electronic medical records (Hillestad, et al. 2005). However, we propose to look at coordination from an operations management (OM) perspective and focus on methods that simultaneously increase effectiveness and reduce costs.

In this chapter we will focus on how operations research can improve service coordination in health systems. An overview is provided to show how operations research contributes to value in healthcare systems. Next, we highlight previous research that looks specifically at health services. Two case studies will focus on different aspects of coordination:

- Case one uses system dynamics modeling to show the importance of coordinating the right capacity levels of resources and the need for long-term planning models in a regional health system.
- Case two uses discrete-event simulation to show the value of coordinating health services within a hospital to ensure timely patient access.

Finally, we will discuss future challenges in coordinating health services that are being driven by changes in healthcare delivery models such as the medical home, shared decision making, and centralization of health services.

## 2 Background on Health Systems

Operations management within healthcare can be viewed from the traditional transformation process perspective (Meredith and Shafer 2007), as shown in Fig. 16.2. Various staff, facilities, technology, and patients are inputs to the healthcare transformation process. The healthcare delivery process is the value-adding stage and includes how the inputs are managed. The effectiveness of delivery strategies and policies and how these are executed in the healthcare operations are key in creating valuable outputs. Treating patients, medical and healthcare service research, and training of healthcare staff are all common outputs for healthcare organizations. In the end, value is defined by the patient and healthcare market. Effective coordination of the inputs is a key to achieving high-value healthcare. High value can be defined as achieving high system performance and low per capita costs (as highlighted in Fig. 16.1).

A complication with understanding value in healthcare is that it entails multiple dimensions. Certainly improved patient health or condition is one dimension of value. However, patient perceptions of what treatment was received and how this treatment was delivered are another dimension and may be at odds with the health outcome. An arduous treatment process or treatment that does not use the inputs desired by the patient may cause dissatisfaction even though the health outcome

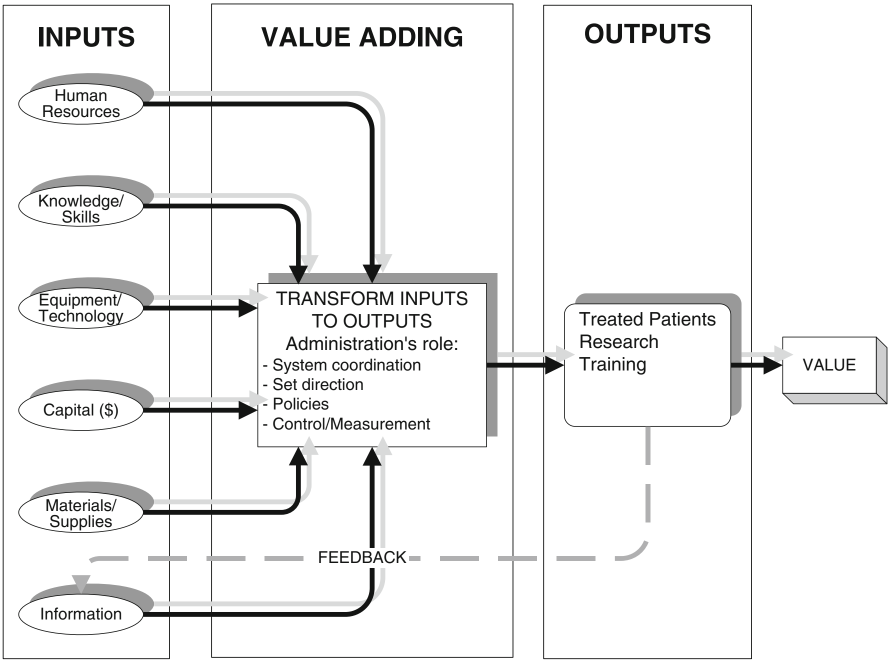


**Fig. 16.2** The operations management transformation process

is positive. In this chapter we will consider value in a multidimensional context that includes service quality, timeliness, price/cost, as well as treatment outcomes. However, we will assume that patient outcomes are highly correlated with patient satisfaction even though for individual patients this may not always be true.

Tension between the individual versus population-based orientation in healthcare is another complication to understanding value. This is particularly relevant when a health system has significant resource constraints. If the capacity for a technology like positron emission tomography (PET) is scarce, then it makes sense from a population perspective to allow only the most appropriate patients access to this diagnostic/treatment. This may exclude some patients from receiving the treatment, even though they could benefit. Thus, individual value may be sacrificed to maximize population or system-wide value. We will consider value from both perspectives, but focus on the system or population perspective.

## *2.1 Literature Review*

In "Building a Better Delivery System: A New Engineering/Healthcare Partnership," the Institute of Medicine and the National Academy of Engineering (2005) partnered with the goal "…to transform the U.S. healthcare sector from an underperforming conglomerate of independent entities (individual practitioners, small group practices, clinics, hospitals, pharmacies, community health centers et. al.) into a high performance 'system' in which every participating unit recognizes its dependence and influence on every other unit." The report notes the importance of systems engineering, operations research, and operations management as a means to achieve health systems coordination at both the inter- and intraorganizational levels. The report is US based; however, we believe the basic philosophy is globally relevant. To emphasize the IOM/NAE's suggested direction, this review will focus on operations research-oriented literature that emphasizes resource coordination.

### 2.1.1 Macrolevel Health Systems Modeling

In this section we discuss macrolevel modeling that considers flows and decision making across organizations and/or major health services functions. The focus is on policy level decisions such as overall capacities of the organizations and functions (e.g., hospitals, primary care). The following section will focus on microlevel modeling that emphasizes flows within organizations or patient pathways. To some extent, this delineation is artificial, and there will be overlaps between problem domains and the tools used for analysis. However, there is likely affinity for certain methods in each modeling level, and therefore, we believe separating them provides a useful structure.

Considering the integration of healthcare systems requires conceptual and quantitative methods that incorporate their inherent breadth and complexity. While detailed operations problems in healthcare can rely on available or collectable data, often healthcare systems entail elements where data are not available. Thus, research and planning methods need the capability to include relationships without data support.

System dynamics (SD) is a methodology that meets these requirements. Its simple set of constructs made up (primarily) of stocks and flows are designed for high-level model building (Sterman 2000). Stocks are accumulations in systems and in healthcare can represent tangible elements such as patients waiting for treatment or intangible elements such as physician's commitment to safety. The flows control the dynamics of the changes in the stocks by specifying the rate of transition from one stock to another. In healthcare, systems flows can represent service rates controlled by the capacity of resources such as physicians or diagnostic equipment. Flows could also be the motivational forces that cause an increase or decrease of intangible stocks such as commitment to safety. As such, the tool is effective for considering policies that address both healthcare system structure and infrastructure. Further, the method is designed to incorporate the effects of time lags and feedback (Morecroft 2007). All of these aspects are prevalent when considering a systems perspective of healthcare.

An area where the SD approach has been applied to healthcare systems is coordinating urgent and elective (or scheduled) care. Brailsford et al. (2004) used an SD model focusing on the accident and emergency department (ED) of a hospital in England. The model showed that if growth in ED patients continued, it would require the eventual reduction of elective admissions to the hospital and missing government prescribed patient access performance targets. The model was also used to evaluate scenarios where some patients were seen in community-based diagnostic centers rather than the hospital ED. Thus, this example shows some of the strengths of SD modeling: analyzing the changes in systems over an extended period of time and exploring broad-based policy options.

Lane et al. (2000) used SD to consider a similar environment to that of Brailsford et al.; however, their model focused on the system effects of hospital bed reductions on the ED. The authors noted that because of patient priority policies, decreasing inpatient beds did not have the expected effect of delaying ED patients; rather it caused elective surgery cancellations. Thus, an important, generalizable conclusion from this work is that healthcare systems should be considered from a holistic perspective that assesses performance of all system components.

The conceptual system model developed by Cooke et al. (2010) considered a similar environment to those of Brailsford et al. and Lane et al.; however, it emphasized the role of changing patient demographics in explaining healthcare system capacity problems. An aging population placed increasing demands on the EDs in the health system studied. Poor planning and coordination of the capacities of primary care, hospital beds, and medical specialties created surprising delays in the EDs given the new demands on the healthcare system. Using causal loop diagrams that showed the complex interrelationships and feedback effects among the health system elements and supporting health system data, the Cooke article hypothesized the many reinforcing factors that created waiting time and access issues at EDs in a Canadian health system. A more complete quantitative model developed for this environment will be discussed in Sect. 3.

Wolstenholme (1993) developed a similar model with a focus on how elderly patients flowed in a healthcare system in the UK. This study showed the value of system dynamics modeling tools to enhance systems thinking. The model identified that as posthospital care of elderly in nursing homes was passed on to budget constrained community care entities, the indirect effect would be to transfer even higher costs back to the National Health System (NHS) via a need for more hospital beds and associated resources. This article, again, reinforces the theme of coordinating multiple elements of systems for achieving effective health systems.

### 2.1.2 Microlevel Health Systems Modeling

System dynamics is often a good choice of methodology for the macrolevel and where its simple constructs have sufficient explanatory capability. However, more detailed modeling methods such as queuing analysis and discrete-event simulation

(DES) are often required for studying integration of healthcare services within organizations, healthcare functions, and patient pathways. Brailsford et al. (2004) used DES to evaluate the efficacy of *streaming* patients with minor issues to separate resources within EDs. This model addressed the ED as a system within a hospital that cares for patients with significantly different levels of acuity. Streaming these different patient types is increasingly considered a practical way to improve performance (Kelly et al. 2007). This type of detailed policy is generally not easy or appropriately modeled using SD.

Levin et al. (2011) presented a case study in which they considered the effect of increasing heart surgery volumes on ED patient access to cardiology services. They used regression analysis to model the relationship of patient characteristics and length of stay (LOS) at various care resources. This analysis was incorporated into a DES model that explored options of increasing capacity and reducing patient LOS on patient access. A unique aspect of this study is that it considered how the resources and operations issues in one department (surgery) affect the service received in another (ED).

Another way that detailed modeling has been applied to healthcare coordination is looking at how upstream decisions affect downstream performance. The article by Bekker and de Bruin (2010) used queuing analysis to report on a number of analysis scenarios, including how operating room scheduling influences the number of ward beds required for recovering patients. Using typical operating room schedules (i.e., no elective surgeries on weekends), the authors show how changing these schedules affects the number of ward beds required. In particular, they show that *front loading* surgeries on Mondays and Tuesdays lead to a more balanced load on the wards and less overall beds are required. Haraden and Resar (2004) discussed the various issues integrating hospital services to improve patient flow, in particular the need for better management of variability in volumes that are created by hospital staff and decision makers. One example from the article noted that in most hospitals, elective surgery schedules are based on individual preferences and do not consider the downstream patient demands in the ICU and other recovery resources. Without coordination, the downstream recovery resources simply have to react to the highly variable upstream decisions.

Rohleder et al. (2007) discussed facility design and implementation decisions associated with building and locating patient service centers (PSCs) for laboratory testing. Opening new service centers in a Canadian city where existing service centers were being phased out led to initial underutilization and very high initial patient satisfaction at the new PSCs. However, the dynamics of patient visitation eventually led to lower satisfaction and patient complaints. By considering the implementation and facility design decisions together rather than sequentially, some of the issue could have been anticipated and the dissatisfaction mitigated. Assisting decision making for this kind of coordination may also require several methodologies. Like Brailsford et al. (2004), this study used both SD and DES methodologies.

# 3 Case Studies

This section will discuss two case settings showing the value integration of health services and the potential of operations management concepts and techniques to assist decision makers. The first case will discuss the use of system modeling to diagnose and provide policy decision support for a large Canadian urban healthcare system. The second case will look at coordination of patient pathways for downstream planning of patient services.

## *3.1 A System Dynamics Model of the Calgary Health Region Emergency Departments*

Rohleder et al. (2009) expanded on the qualitative model described in Cooke et al. (2010) to create a macrolevel system dynamics model of the Calgary Health Region (CHR). The original purpose of the model was to explain the causes for delays in treating patients in the city's EDs; however, as the study progressed, it was evident that the scope needed to consider the whole of the major health services in the region. Therefore, this case shows the importance of looking at healthcare from a systems perspective and the importance of coordinating capacity levels across all health services.

At the start of the study, the average LOS of patients in the largest ED was 8 h, and the average time for initial treatment by an ED doctor was about 1.5 h. Due to the effects of variability, sometimes the wait time before being seen by a physician was even longer and led to some unfortunate outcomes in the ED waiting rooms (Lang 2006). An original hypothesis of the causes for delays is shown in Fig. 16.3.

An aging demographic was believed to be driving up the overall severity level of patients in Calgary EDs. The proportion of patients that had higher acuity and greater complexity increased. This increased the load on the ED resources and therefore the number of patients in the waiting rooms.

Initial data on patient ages and acuity were collected, and they supported the model in Fig. 16.3; however, the full scope of the problem quickly expanded to include components of the healthcare system outside of the ED. As the model evolved, the focus remained on the congestion in the EDs. It also showed the interconnectedness within health systems and the need to plan and implement integrated services.

While decision makers continued to try and resolve the problems by improving ED operations, it became clear from our model that over time the ED became the *safety net* to accommodate the demographic effect on healthcare capacity from a growing population of older and sicker patients. The model helped explain why the EDs' patient access performance was declining rapidly when year-over-year patient volume growth was almost flat. Essentially, the performance problems in the EDs were a symptom of capacity and coordination problems throughout the healthcare system.

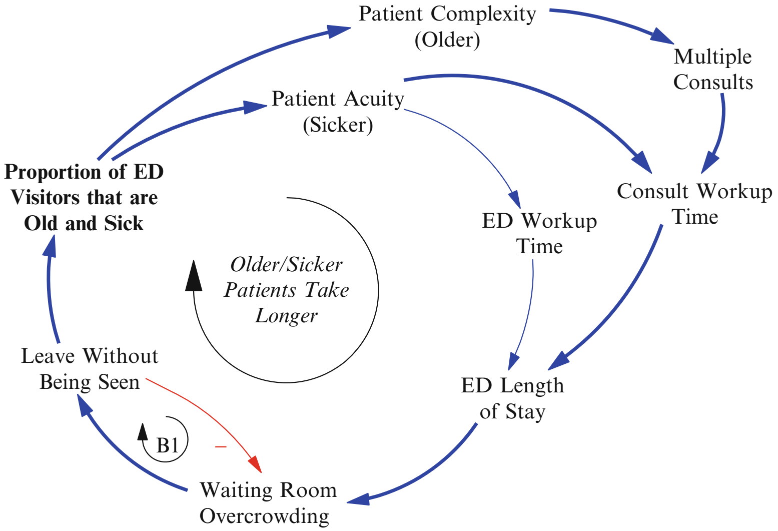

Patient Complexity (Older)
Multiple Consults
Patient Acuity (Sicker)
Proportion of ED Visitors that are Old and Sick
Consult Workup Time
ED Workup Time
Older/Sicker Patients Take Longer
Leave Without Being Seen
ED Length of Stay
B1
–
Waiting Room Overcrowding

**Fig. 16.3** Causal loop diagram to explain delays in EDs

The purpose of the system model was to examine the impact on the EDs of patient flows in the overall CHR system. For example, insufficient inpatient capacity for elective surgeries increased the possibility that patients on the hospital waiting list had to go to an ED. Similarly, if a patient was unable to access a family physician, then they were more likely to go to an ED or urgent care center. In this study we were interested in long-term impacts of changes in population demographics and available health services. Thus, we were not concerned with the hour-to-hour fluctuations in demand for health services, but rather with the month-to-month and year-to-year ability of the system to cope with the average demand for services from the regional patient population. The following subsections will briefly discuss the model building process and results from the model related to coordination and operations management.

The qualitative model that served as the basis for the SD model was developed using an iterative group process that required many interviews with people involved in the healthcare system and bringing groups of healthcare staff and administrators together to understand the patient flows and factors influencing patient access. Figure 16.4 shows a highly simplified version of the model of the patient flows that evolved from this process and became the basis for the quantitative SD model. Three *sub-models* emerged from the overall systems view. The ED care sub-model is the set of stocks and flows along the bottom of Fig. 16.4; the acute care sub-model included the stocks and flows associated with hospital services (right side of figure); and the primary care sub-model includes the primary care, urgent care (and other clinics), and specialist consulting services (top left of figure). Splitting the model

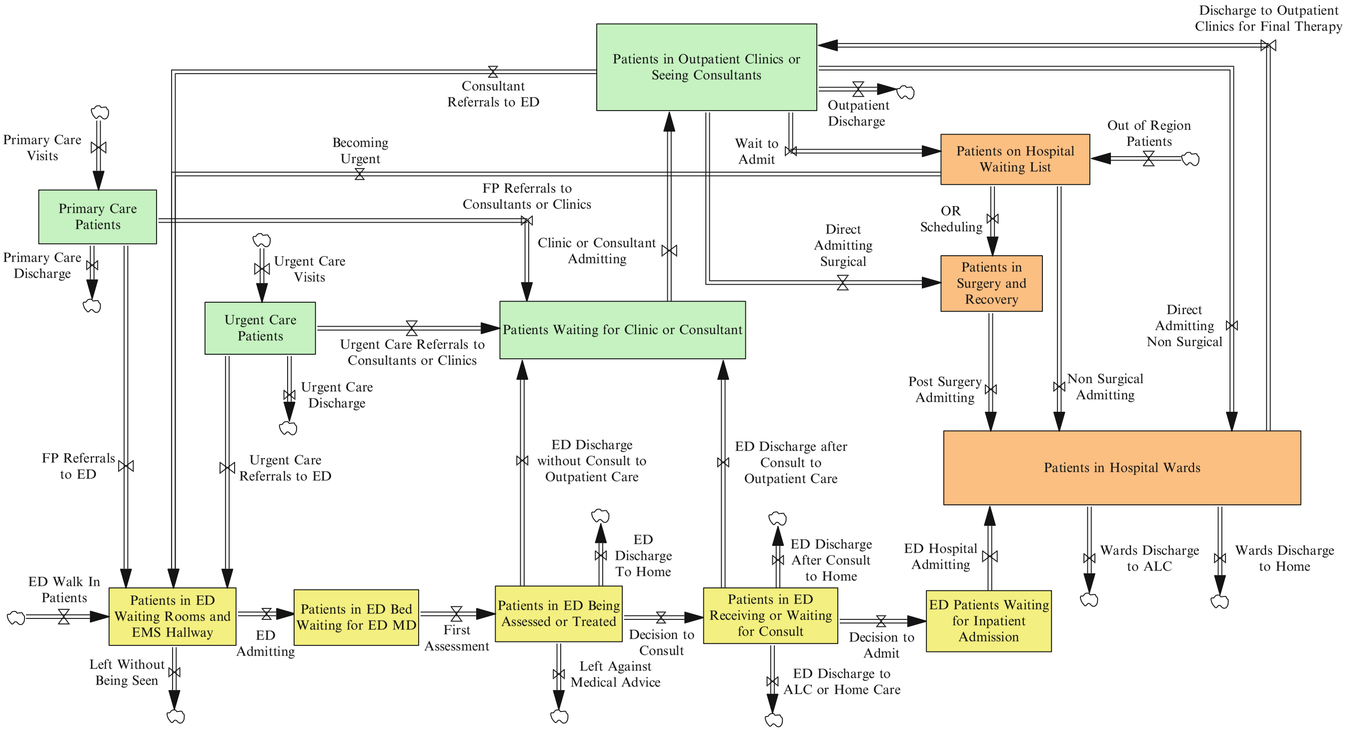

Discharge to Outpatient
Clinics for Final Therapy
Patients in Outpatient Clinics or Seeing Consultants
Consultant Referrals to ED
Outpatient Discharge
Primary Care Visits
Becoming Urgent
Wait to Admit
Patients on Hospital Waiting List
Out of Region Patients
Primary Care Patients
FP Referrals to Consultants or Clinics
OR Scheduling
Direct Admitting Surgical
Primary Care Discharge
Urgent Care Visits
Clinic or Consultant Admitting
Patients in Surgery and Recovery
Urgent Care Patients
Patients Waiting for Clinic or Consultant
Direct Admitting Non Surgical
Urgent Care Referrals to Consultants or Clinics
Urgent Care Discharge
Post Surgery Admitting
Non Surgical Admitting
ED Discharge without Consult to Outpatient Care
ED Discharge after Consult to Outpatient Care
FP Referrals to ED
Urgent Care Referrals to ED
Patients in Hospital Wards
ED Discharge To Home
ED Discharge After Consult to Home
ED Hospital Admitting
Wards Discharge to ALC
Wards Discharge to Home
ED Walk In Patients
Patients in ED Waiting Rooms and EMS Hallway
Patients in ED Bed Waiting for ED MD
Patients in ED Being Assessed or Treated
Patients in ED Receiving or Waiting for Consult
ED Patients Waiting for Inpatient Admission
ED Admitting
First Assessment
Decision to Consult
Decision to Admit
Left Without Being Seen
Left Against Medical Advice
ED Discharge to ALC or Home Care

**Fig. 16.4** Systems model of the Calgary Health Region (with detail of the emergency department)

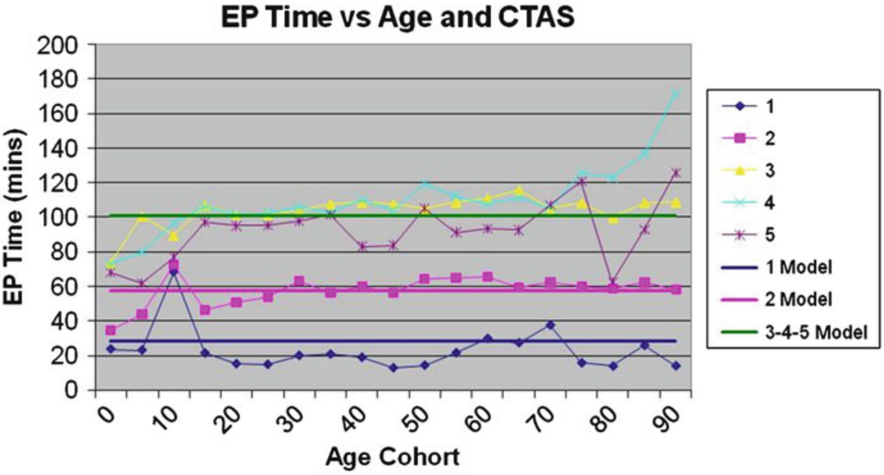


**Fig. 16.5** Data for emergency physician time by CTAS (1–5) and age

into sub-models made it easier to handle the complexity of the entire system. The ED sub-model has greater detail due to the initial focus on this area. Nonetheless, the important flows from each service sector are included.

With the model components and boundaries established, sources of available data were identified to populate the model. To incorporate the urgency of a patient's condition, we included patient age and an urgency score based on the Canadian Triage Assessment Scale (CTAS) as patient attributes. Together these account for much of patients' usage of healthcare resources. The CTAS score of a patient has a particular meaning in the ED where it is used to prioritize patients; however, we used CTAS throughout the model to determine when patients required urgent or acute care.

As an example of how we used data in the model, we will use the time required for an ED physician to perform a first assessment on a patient. The time spent waiting for an emergency physician (EP) is called mean EP time. The model uses EP time data that were collected from an ED during the April–September 2006 time period, involving treatment of over 31,000 patients. Figure 16.5 shows that the mean EP time is about 23 min for CTAS 1, about 58 min for CTAS 2, and about 100 min for CTAS 3–5 patients. We assume no correlation with age in this situation; however, for other flows, age was an important factor. Although there appears to be some correlation with age for CTAS 4 patients, the data is skewed by just a few very old patients with long wait times. For CTAS 4 patients aged 80 and under, there is little or no correlation between patient age and EP time. Probabilitydistribution functions were used in the model to represent CTAS 1–5 patients.

The overall model was built in this manner, incorporating relationships between age and patient acuity as appropriate. As identified in Fig. 16.4, the stocks in the system are where patients accumulate and wait for access, and the flows control

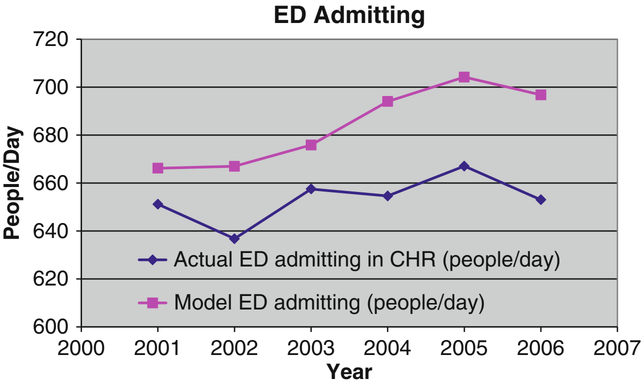


**Fig. 16.6** Comparison of ED admitting rates for model validation

the movement of patients between stocks. Together, these stock and flow constructs required hundreds of data elements.

Toensure the model created was valid, we compared model results to actual results for several measures including ED admission rate, inpatient admission rate, and ED and inpatient LOS, among others. Figure 16.6 shows the graphical comparison for the average number of patients admitted to the ED per day. The model reports slightly better performance than occurred in practice. In part this is due to the fact that the SD model does not incorporate the effects of short-term variability where within-day peaks and valleys have a negative impact on resource utilization due to congestion. Nonetheless, the differences were deemed acceptable to decision makers.

Having a valid model is not enough to ensure its usefulness. A key concern with use of the model for making system-wide decisions was, would it do a better job of explaining system behavior than the simple, standard measures in current use? Typically, health system planners use metrics, such as number of beds per unit population, to plan the capacity of the healthcare system. A common expectation in practice is that if the growth in the number of staffed beds grows with the population, and support services such as laboratory testing and diagnostic imaging keep pace with bed growth, then capacity needs should be met. To show why the SD model may be more useful than such simple rules of thumb, Fig. 16.7 reports a comparison of various key measures related to population changes and healthcare system performance.

Bed growth kept pace with general population growth over the 2001–2006 period, but the ED admitting rate did not. A good measure of system capacity is *throughput*, and in the case of hospital EDs, this is measured by the ED admitting rate per day. While the SD model results are not exact, Fig. 16.7 suggests that they do a better job of predicting actual ED admitting rates than would be expected from a bed growth projection. As noted in Fig. 16.7, part of the reason for the inability of bed growth projections to satisfy future demographic demands is that

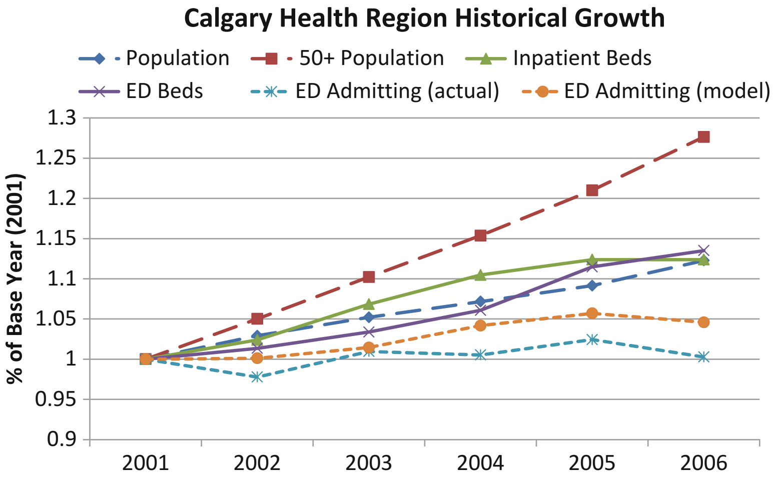


**Fig. 16.7** ED admitting compared to population and bed growth

**Table 16.1** Bed policy scenarios

| Scenario number | Scenario name | Brief description |
|---|---|---|
| 1 | Base 2008 | No change in ED or hospital capacity (beds and staffing) from 2008 to 2016 |
| 2 | ED expansion only | ED capacity is expanded by ∼20% to match 2008–2016 population growth |
| 3 | Hospital expansion only | Hospital capacity expansion to 2011 per CHR plan |
| 4 | ED and hospital expansion | Combination of cases 2 and 3 |
| 5 | Further hospital expansion | As case 4, but further 10% hospital capacity expansion 2014–2016 |

the aging population is growing at a much faster rate than the general population. The system dynamics model is able to take these demographic changes and system-wide capacity levels that influence ED patient demand into account.

Our focus in the analysis was on the effect of bed capacity on health system performance; however, as described later, the model could be used to explore many facets of resource coordination including staffing. We considered five scenarios, summarized in Table 16.1, which represent different possible alternatives for staffed bed expansion. Examination of the results will illustrate the use of the system dynamics model for policy evaluation.

Scenario 1 provides a baseline projection of what would happen if no capacity expansion took place beyond what existed at the end of 2007. Scenario 2 assumes that ED bed capacity is expanded at the same rate as population growth. As shown in Fig. 16.7, this essentially replicates historical practice. Scenario 3 assumes that inpatient bed capacity is expanded according to projections provided by senior-level decision makers. This projection takes into account the planned hospital expansions.

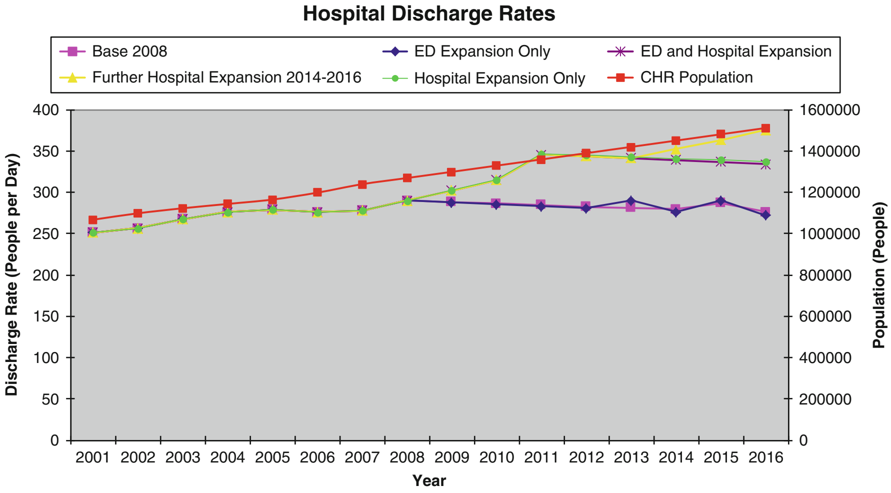


**Fig. 16.8** Hospital discharge rates for bed expansion scenarios

Scenario 4 is a combination of scenarios 2 and 3, and these three cases will let us answer the question, is it better to expand ED beds, inpatient beds, or expand both together? Finally, scenario 5 was added because the results of scenario 4 revealed that capacity limitations will again be encountered in 2014 unless further inpatient bed capacity is added. Scenario 5 assumes the same ED bed capacity as in cases 2 and 4.

In all scenarios we use the actual CHR population figures for 2001–2006 and the CHR population projections for 2008–2016. We assume that there is sufficient support staff and services to operate the beds and all five scenarios have the same assumptions for capacity increases in family physicians, urgent care centers, consultants, and outpatient clinics.

An important issue is the effect of bed expansion on hospital discharge rates, a measure of patient throughput. In Fig. 16.8 we compare the hospital discharge rates for each of the five cases to the growth in population rate. The main conclusions that can be drawn are as follows:

- Without inpatient bed expansion, throughput will actually *drop*. This is because in a capacity constrained system, the sickest people will be given priority and they will consume more resources and stay longer than the less sick people who are displaced. This point emphasizes the need to look at the healthcare services as a coordinated system.
- Adding ED beds does nothing to change the throughput of hospital inpatients.
- The planned 2008–2011 expansions (inpatient beds) will close the gap that has developed between population growth and hospital throughput during the 2001–2007 period.

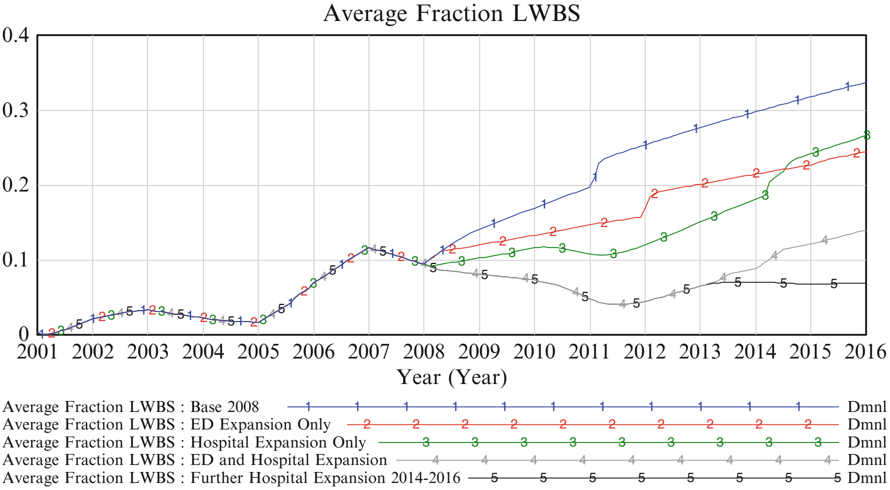


**Fig. 16.9** Patients who leave the ED without being seen

- While capacity from the 2008–2011 planned expansions will be sufficient to support population growth to 2013, further capacity expansion in the 2014–2016 time frame will be required to prevent future system bottlenecks.

Another major concern for ED performance is the proportion of patients who leave the ED without being seen (LWBS). Figure 16.9 shows the results for this measure for the five scenarios. The model picks up the degradation in performance in 2006 when several highly publicized adverse events occurred (including a death that was partly attributed to a patient who left an ED, Lang 2006). Again, it is clear that a coordinated planning perspective that considers increasing both ED and hospital beds together is required. Further, additional expansion to that which is planned is needed to avoid running into severe access problems in the future.

Our model emphasized the role that bed capacities and their coordination play in the access of integrated health services. However, we could have easily used the model to explore a variety of strategic level decisions beyond bed capacities. For instance, there was a prevailing perception that there was a chronic shortage of primary care services in the CHR. Our model could explore the effect of increasing primary care capacity on patient access. The effect of adding any type of the resources in the model versus any other could help shape recruiting and future system design decisions of a long-term strategic nature.

Overall, this case helps to show the importance of understanding the system-wide flow of patients when considering a single service. System dynamics is a good tool for this purpose due to its ability to incorporate the relationships between many system elements, including feedback from other system elements. Further, SD's ability to incorporate dynamic changes of key inputs like the demographics of a patient population makes it useful for high-level healthcare planning over a long horizon.

## 3.2 A Simulation Model to Improve Coordination of Health Services in a Hospital

The previous section described an SD model that considered the potential value of improving capacity decisions at a long-term, strategic level. In contrast, the discrete-event simulation model described in this section will focus on some shorter-term decisions that are more operational in nature. It will focus on a care pathway for patients who are admitted to the Hospital Medicine Service (staffed by Hospitalists who provide delivery of comprehensive medical care to hospitalized patients). This service typically handles patients with multiple comorbidities or with complex diagnoses that do not fit well into another specialized service. Therefore, Hospitalist patients were frequently elderly patients or those that require significant care services upon discharge from the hospital.

The primary purpose of this model is to show how the downstream operations of an admitting service affect patient flow in the ED. In this model we consider some of the specific mechanisms causing the boarding delay of ED patients in the Hospitalist Service at a Calgary hospital. Thus, we are looking at coordinating the decision making of healthcare services at the process level. Using the model, our major objectives were:

1. To show the key factors in the Hospitalist admitting service that cause boarding delays in the ED.
2. Identify opportunities to improve patient flow for both the ED and the Hospitalist Service.

As the system model demonstrated in Sect. 3.1, there are interlinkages among many health services that contribute to delays and impediments to smooth patient flow. The model in this section considers one of the links in the system in more detail, so as to increase understanding of these interactions. This will help understand how the system structure can affect performance at the operational level and vice versa. As is common in the OR discipline, our objective was to find improvement solutions that worked well within the whole system and benefited all of the services involved.

The high-level patient flow for the Hospitalist Service is shown in Fig. 16.10. Some of the key highlights of the model are:

- The targeted capacity of the Hospitalist Service was 180 patients. However, up to 203 patients could be handled depending on circumstances.
- Patients under consideration for admission from the ED could be diverted to other services when the target capacity limit was reached or passed. From a patient level of 180–203, there was a specified probability that patients would not be admitted. This probability increased with the number of patients. Once the *hard* limit of 203 patients under care was reached, admissions via the ED and transfers from other services were not accepted. The exception to this was ICU patients who were always accepted, given the importance of maintaining ICU availability.

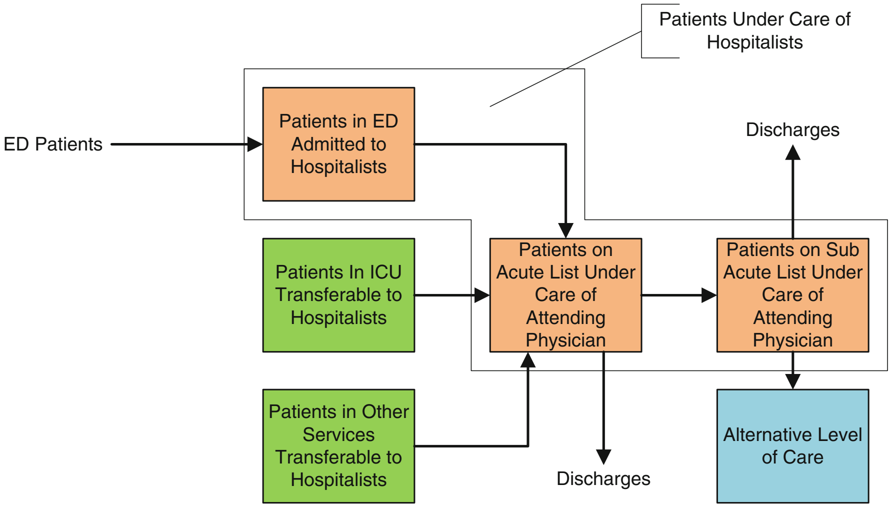

Patients Under Care of Hospitalists
ED Patients
Patients in ED Admitted to Hospitalists
Discharges
Patients In ICU Transferable to Hospitalists
Patients on Acute List Under Care of Attending Physician
Patients on Sub Acute List Under Care of Attending Physician
Patients in Other Services Transferable to Hospitalists
Discharges
Alternative Level of Care

**Fig. 16.10** Diagram of high-level patient flows within Hospitalist Service

For this reason, ICU patients are given priority when multiple patients are waiting for a bed.

- Two classifications of patients were considered: acute and subacute. Acute patients required greater care and attention than those considered subacute. A higher level of staff were required to handle acute patients. Subacute patients were often waiting to get into an alternative level of care (ALC) space such as a nursing home bed.

Figures 16.11 and 16.12 show a couple of the key data elements incorporated in the simulation model of the Hospitalist Service. Figure 16.11 shows the hourly patterns for the ED and OS (other, non-ED hospitalized) patients that are admitted or transferred to the Hospitalist Service. The figure shows that ED patients tend to be spread more evenly across all hours of the day with several small peaks. On the other hand, the transfers from other services tend to be concentrated during the hours of 9 am to 6 pm. ICU patient transfers are also concentrated during standard working hours.

Figure 16.12 shows that patient volumes also varied by day of the week. The figure shows the pattern for each of the three incoming patient streams to the Hospitalists. Note the peaks on Fridays for the transfer of patients form other services (OS) in the hospital. This was, in part, due to a desire for these services to pass their patients on to another service before the weekend.

Along with the patient arrival data, we incorporated the utilization level of various hospital resources into the simulation model, in particular, LOS in hospital beds. Table 16.2 shows the average LOS for patients coming from the ED, ICU, and other medical services as well as for patients who are eventually classified as subacute within the Hospitalist Service. Various mathematical functions were

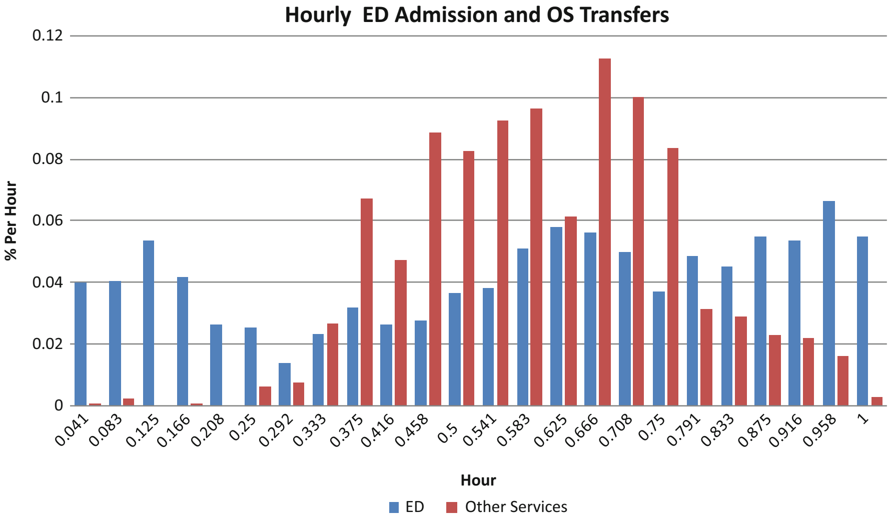


**Fig. 16.11** Admission rates to the Hospitalist Service

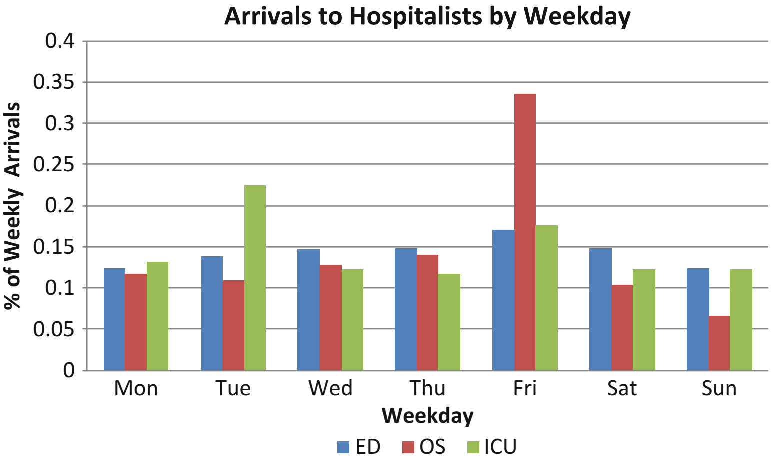


**Fig. 16.12** Weekly pattern of arrivals to the hospital service

**Table 16.2** LOS means for Hospitalist patients (days)

| From | Acute ED | ICU | OS | Subacute All |
|---|---|---|---|---|
| Mean LOS | 11.13 | 13.34 | 16.27 | 31.82 |

evaluated to determine how well they fit the data. The gamma distribution was found to be a good fit because the LOS data all had a significant right skew and was therefore used in the simulation model.

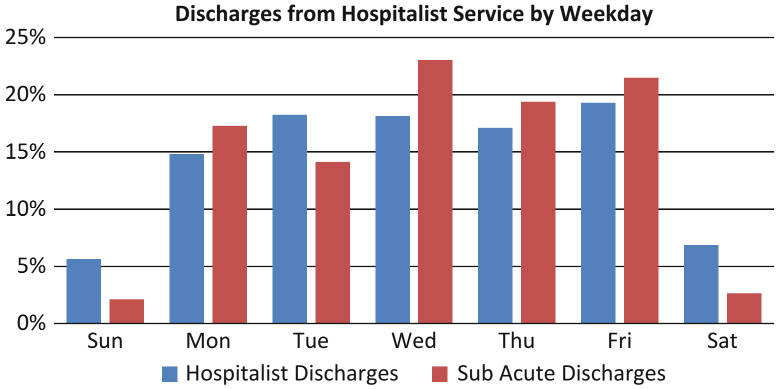


**Fig. 16.13** Hospitalist discharge pattern

**Table 16.3** Discharge patterns evaluated

| Discharge pattern | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Sunday |
|---|---|---|---|---|---|---|---|
| Current (%) | 14.8 | 18.2 | 18.1 | 17.1 | 19.3 | 6.8 | 5.6 |
| Smoothed (%) | 14.3 | 14.3 | 14.3 | 14.3 | 14.3 | 14.3 | 14.3 |

Finally, the discharge process also had to be built into the model. Figure 16.13 shows the empirical discharge patterns for acute and subacute patients. The process for modeling the discharge patterns is fairly complex, but requires adding additional time onto the raw LOS obtained from the data. To avoid biasing the model with longer lengths of stays than in reality, we adjusted the means and variance levels of the treatment LOS so that when the discharge adjustment was made, the overall LOS times in the model would be approximately the same as in the real system. We tested the discharge modeling approach and were able to achieve the desired discharge patterns and average LOS times.

A key focus in looking at the decision making for the Hospitalist Service is how it affected the waiting time for patients in the ED who were eventually admitted to the Hospitalist Service. We evaluated three options to address these integrated effects: smoothing discharges, smoothing transfers from other services, and reducing the LOS for subacute patients.

Table 16.3 shows the current and proposed average discharge rates we tested. The hourly patterns were not adjusted since we viewed these as more difficult to change given the various parties that need to come together to complete a discharge.

Table 16.4 reports the results for the alternative discharge patterns. Smoothing the discharges more evenly across the week makes a significant difference, reducing boarding time by over a half an hour on average. Extreme delays are reduced even more, with the smoother pattern resulting in a 17 h reduction. Significant reductions in the number boarding and the % of ED patients who are diverted also result from a more even discharge pattern.

**Table 16.4** Results for ED waiting from smoothing discharge pattern in Hospitalist Service

| Scenario | Avg. boarding time (h) | Max boarding time (days) | Avg. # boarding | % ED patients diverted |
|---|---|---|---|---|
| Base | 6.95 | 4.75 | 3.14 | 2.21 |
| Smoothed discharge | 6.43 ± 0.2[a] (7.5%)[b] | 4.04 (15.0%)[b] | 2.89 ± 0.1[a] (8.0%)[b] | 2.00 (9.5%)[b] |

[a] 95% confidence interval halfwidth

[b] Percentage improvement in parentheses

The smoothed discharges across all days are not necessarily optimal—it is likely we could identify an even better weekday pattern of discharges, particularly given the uneven arrivals to the Hospitalist Service. However, the constant average values across the weekdays were viewed as a reasonably implementable target.

Figure 16.12 showed that Fridays have significantly more transfers from other services than other days (nearly one third on Friday alone). However, all things being equal, smoothing these transfers did not have a positive effect on performance. Because transfers are somewhat controllable, the current pattern already adapts to the flow of patients from other sources.

Finally, the results of interviews with Hospitalist Service staff led us to believe that one of the major causes of delays and poor patient flow from the ED through the Hospitalist Service was the inability to move patients from subacute care to ALC options. The primary impact of such a situation in the system would be lengthy stays in subacute care. This was confirmed by the data once we combined all stays after a patient was initially classified as subacute. The average LOS was nearly 33 days for subacute patients. Thus, our final analysis shows the effects of reducing subacute LOS.

Figure 16.14 shows the results of reducing the average subacute LOS by 5% increments down to 50% of the current (or base) value (or about 16 days). The combined effect of discharge smoothing is also shown in the figure. The results show that ED boarding time is significantly reduced for patients admitted to the Hospitalist Service as the subacute LOS is reduced. Even a relatively modest 10% reduction in LOS leads to over a 1.5 h reduction in boarding time for patients waiting to get into Hospitalist Service beds. If subacute LOS could be reduced by 50%, the impact on the flow of ED patients is dramatic. Boarding times are projected to be reduced to less than an hour, and on average, nearly two bed spaces in the ED are freed compared to the base case results.

It is interesting to note that the discharge smoothing essentially has the constant effect of reducing boarding times by an average of about a half hour regardless of subacute LOS. Therefore, since there is little or no interaction effect between the two improvement options, it appears they could be implemented independently and both significantly improve ED boarding time performance.

In summary, this case used discrete-event simulation to show the value of coordinating health services for a specific patient pathway in a hospital. This level of coordination is just as important to creating high-value outcomes as the previous strategic level coordination model described in Sect. 3.1. While strategic capacity

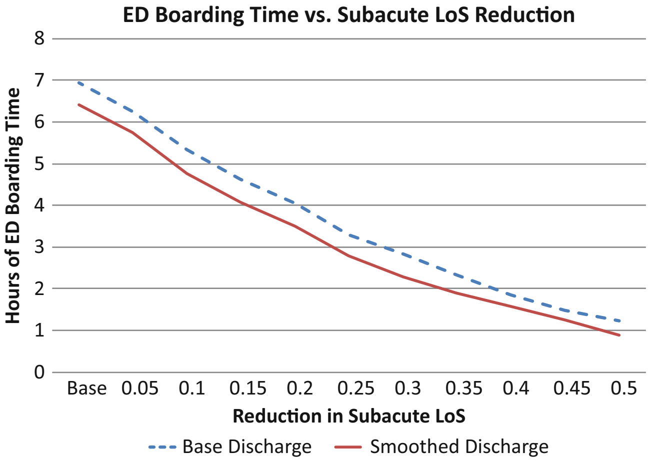


**Fig. 16.14** Effect of reducing subacute LOS

decisions are certainly constraints for tactical and operating level decisions, if effective operations management that coordinates the higher level capacity is not applied at these lower level decision tiers, system performance will be suboptimal.

## 4 Future Research Challenges

Some emerging trends in healthcare delivery have significant implications for the coordination of health services. In this section we will discuss two of these trends: the patient-centered medical home and the centralization of health services at medical centers.

An emerging movement in healthcare coordination is the concept of a PCMH, a term first coined by the American Academy of Pediatrics in 1967. Proponents of PCMH in healthcare share many of the same objectives as health systems engineers/operations managers, that is, to improve the efficiency and effectiveness of delivering medical care and, by extension, the health and wellness of the whole patient.

A PCMH is a team-based model of care led by a personal physician who provides continuous and coordinated care throughout a patient’s lifetime to maximize health outcomes (American College of Physicians 2011). Further, a PCMH is where the *responsibility* for coordinating all of their healthcare is intentionally focused, and serves as the repository for a patient’s entire medical record (Hughes and Stiles 1977). A PCMH is, essentially, part of an *engineered health system* that incrementally reorganizes and leverages existing resources to serve the patient.

One area where the PCMH concept will require OM expertise is the design and management of large groups of patients with chronic conditions. Appropriate teams and resourcing will be need to be structured in a coordinated and systems focused manner. Instead of small groups of primary care doctors caring for chronically ill patients, teams that include doctors and care managers will likely be used. Homer et al. (2004) discuss the use of an SD model to help in the transformation in chronic care services for diabetes and cardiac patients at the county level in the USA. They highlight that new, focused teams of resources can provide better care at lower cost; however, it requires a reconfiguration of the healthcare system away from traditional primary care practice. OM may help in making appropriate trade-offs among health quality performance and cost to improve the level of healthcare value.

Another emerging trend in healthcare is for greater centralization of healthcare services. Major medical centers offer the opportunity to provide inpatient and outpatient services together using the same facilities (or campus) under a single organizational structure. This allows for unique advantages to care, most importantly the ability to quickly bring together a multidisciplinary group of physicians and services to diagnose and treat patients. One such center, Mayo Clinic, was founded on the concept of integrated health services by housing physicians of all disciplines in concentrated locations at their three campuses (Jacksonville FL, Rochester MN, and Scottsdale AZ). This allows for coordination in treatment and use of resources in ways that standard healthcare systems cannot achieve.

Mayo Clinic and similar centers are considered *destination* medical because patients will travel significant distances, including internationally, to receive treatment. High-quality health services are required to attract patients from long distances; however, operational performance must also be excellent. In particular receipt of required health services in a short period of time is of prime importance. This allows patients to plan stays of reasonable length for travel and expense purposes.

While destination medical centers like Mayo Clinic have many potential advantages, they also have operations challenges related to coordination. To avoid many patients being scheduled or seeking to use the same services simultaneously, careful management of physician schedules and capacities is required. To this end, Mayo Clinic uses advanced decision support systems to help with scheduling for certain key diagnostic services. Nonetheless, bottlenecks still occur and may cause delays for patients. To alleviate such circumstances, visits of patients of different types may be coordinated. For example, national or international patients coming to Mayo Clinic with similar schedules may cause peaks of demand for particular services. Local patients can be scheduled to smooth out these peaks since they have more flexibility with respect to travel.

Another issue in managing medical centers where patients make appointments well in advance is no-shows, reschedules, and cancellations (NRC). A significant proportion of appointments end up in these categories and create additional load on staff to resolve schedules as well as unused capacity due to unfilled appointment slots. Figure 16.15 shows a causal loop diagram describing the reinforcing system behavior associated with NRCs. As patient volumes increase, appointment availability decreases. Less availability causes delays in finding appointments (increasing Appointment Latency) which causes more NRCs. This type of behavior is also

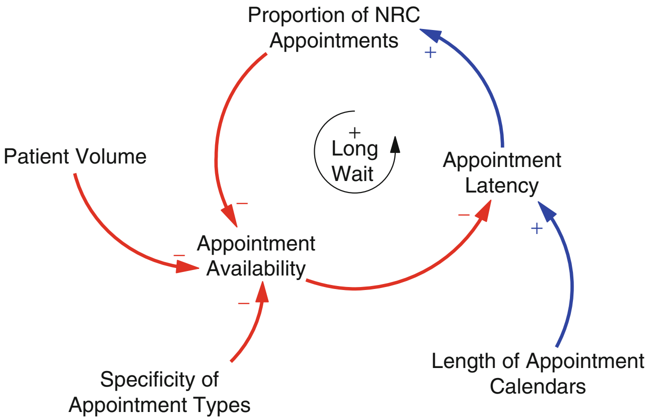

Proportion of NRC Appointments
+
+
Long Wait
Patient Volume
Appointment Latency
–
–
+
–
Appointment Availability
–
Specificity of Appointment Types
Length of Appointment Calendars

**Fig. 16.15** Causal loop diagram of no-show, reschedules, and cancelation behavior

described in Gallucci et al. (2005). The NRCs take up *false capacity* and hence serve to further reduce appointment availability.

This negative cycle may be alleviated by shortening the length of time that appointments can be scheduled and by creating more flexibility in physician calendars. However, patients who are traveling from a long distance often want to make travel plans well in advance of their visit, and physicians prefer to specialize in treating particular types of patients for research and quality purposes. Thus, while Open Access scheduling, where little or no time occurs between the demand and delivery of the health service (see Kopach et al. 2007 for a discussion), may alleviate NRCs, it may not be easily implementable at large medical centers. Using advanced scheduling approaches like that discussed in Helm et al. (2011) that consider the interaction of hospital resources may be more appropriate for advanced scheduling of patients moving through several healthcare services.

In summary, emerging trends in healthcare like the PCMH and centralization of healthcare services into medical centers create new challenges for healthcare operations management. Operations management, systems engineering, and systems level modeling can play important roles in how effectively these new delivery approaches create healthcare value to patients.

## 5 Conclusions

In this chapter we focused on the value of coordinating healthcare services. With costs of healthcare rapidly increasing and changes in demographics and policies that are expanding patient populations, the value of both high-quality and efficient healthcare systems is paramount. The discipline of operations research will play a key role in the success of healthcare service integration because of its focus on coordinating resources to achieve desired outcomes. In the past, this coordination tended to focus on individual health services. For the future, the coordination of several services or a system of services will become increasingly important.

It is clear that operations management can play a valuable role in improving the coordination of existing healthcare services. However, it is likely that new healthcare systems will evolve in the future, often pushed by the need for radical change due to increasing costs and policy changes. Our section on research challenges provided two examples of emerging trends in healthcare delivery important to operations management. With the patient-centered medical home, new team structures and care management approaches will be necessary to success. Another trend toward the growing role of medical centers that patients travel to creates unique opportunities for integrated care, but significant challenges in ensuring timely and efficient delivery.

Thus, operations management has much to offer health services in managing current healthcare process and designing new ones to maximize the delivery of value to patients.

**Acknowledgements** The authors wish to thank the following staff from the Calgary Health Region: Karen Abbott, Rob Abernethy, Rick Anderson, Maria Bacchus, Simon Chi, Janet Cohen, Gil Curry, Stafford Dean, Jim Eisner, Ward Flemons, Jennifer Foster, Toni Furber, Gavin Greenfield, Nancy Guebert, Beth Harris, Caroline Hatcher, Kathy Howe, Lucy Hyndman, Grant Innes, Peter Jamieson, Edward Kukec, Linda Lutz, Sharleen Luzny, Lester Mercuur, Micheline Nimmock, Charlotte Parkinson, Judy Pederson, Lucy Reyes, Tom Rich, Karen Ruggles, Tanya Sakamoto, Jason Scarlett, Lisa Schaffer, Karalee Schideler, Rick Schorn, Liselotte Simeon, Dongmei Wang, and Chris Wilkes. In addition, we acknowledge the contribution of Serene Chen who contributed to this work as a summer intern student at Mayo Clinic.

## References

American College of Physicians (2011). http://www.acponline.org/advocacy/where_we_stand/medical_home/. Accessed 18 Apr 2011

Bekker R, de Bruin AM (2010) Time-dependent analysis for refused admissions in clinical wards. Ann Oper Res 178:45–65

Brailsford SC, Lattimer VA et al (2004) Emergency and on-demand health care: modelling a large complex system. J Oper Res Soc 55:34–42

Cheung PT, Wiler JL, Ginde AA (2011) Changes in barriers to primary care and emergency department utilization. Arch Intern Med 171(15):1397–1398

Cooke D, Rohleder T et al (2010) A dynamic model of the systemic causes for patient treatment delays in emergency departments. J Model Manag 5(3):287–301

Davis, K, Schoen, C, Stremikis, K (2010) Mirror, Mirror on the Wall: How the Performance of the U.S. Health Care System Compares Internationally, 2010 Update, Commonwealth Fund Report, June 2010

Gallucci G, Swartz W, Hackerman F (2005) Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. Psychiatr Serv 56:344–346

Haraden C, Resar R (2004) Patient flow in hospitals. Frontiers of Health Serv Mgmt 20(4): 3–15

Helm JE, AhmadBeygi S, Van Oyen M (2011) Design and analysis of hospital admission control for operational effectiveness. Prod Oper Manag 20(3):359–374

Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, Taylor R (2005) Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Aff 24(5):1103–1117

Homer J, Hirsch G et al (2004) Models for collaboration: how system dynamics helped a community organize cost-effective care for chronic illness. Sys Dyn Rev 20(3):199–222

Hughes JR, Stiles F (1977) Fragmentation of care and the medical home. Pediatrics 60:559

Institute of Medicine (1999) To err is human: building a safer health system. National Academy Press

Kelly A-M, Bryant M, Cox L, Jolley D (2007) Improving emergency department efficiency by patient streaming to outcomes-based teams. Aust Health Rev 31(1):16–21

Kopach R, DeLaurentis P-C, Lawley M, Muthuraman K et al (2007) Effects of clinical characteristics on successful open access scheduling. Health Care Manag Sci 10(2):111

Lane DC, Monefeldt C et al (2000) Looking in the wrong place for health care improvements: a system dynamics study of an accident and emergency department. J Oper Res Soc 51:518–531

Lang M (2006) City emergency care under review: probe follows high-profile problems. Calgary Herald, Calgary

Levin S, Dittus R et al (2011) Evaluating the effects of increasing surgical volume on emergency department patient access. BMJ Qual Saf 20(2):146–152

Meredith JR,Shafer SM (2007) Operations management for MBAs, 3rd edn. Wiley, New Jersey

Morecroft J (2007) Strategic modelling and business dynamics: a feedback systems approach. Wiley, West Sussex, England

Sterman JD (2000) Business dynamics: systems thinking and modeling for a complex world. McGraw-Hill, New York

Rohleder TR, Bischak DP et al (2007) Modeling patient service centers with simulation and system dynamics. Health Care Manag Sci 10(1):1–12

Rohleder TR, Rogers P, Cooke DL, Xu S (2009) Emergency department simulation models: a report for the Calgary Health Region. Report commissioned by the Calgary Health Region

Wolstenholme EJ (1993) A case study in community care using systems thinking. J Oper Res Soc 44:925–934

# Chapter 17
# Managing Supply Critical to Patient Care: An Introduction to Hospital Inventory Management for Pharmaceuticals

**Anita R. Vila-Parrish and Julie Simmons Ivy**

## 1 Introduction

Hospital operations and patient care are inextricably linked to the supply chain. Like procurement personnel in other industries, hospital buyers are challenged to develop inventory policies in light of changing demand, limited suppliers, manufacturing issues, and regulatory rulings that affect drug supply (Choudhary et al. 2011). While on the surface these challenges are similar to those faced by other industries, the impact of shortages in critical hospital supplies can have detrimental impacts to patient care, patient outcomes, and the cost of care. This chapter focuses on inventory management of critical materials and supplies that directly impact patient care. We will focus on one category of hospital supplies: pharmaceuticals. Pharmaceuticals can be categorized by the form of medication: raw drug (i.e., a powder or solid that requires further processing prior to administration) and prepared medications (e.g., an intravenous fluid, IV).

A unique challenge in the inventory management of pharmaceutical supplies that are critical for providing patient care is the impact of suboptimal control on both a patient's treatment plan and the resources involved in ordering, producing, and administering medications. As Choudhary et al. (2011) described, drug shortages can force patients to alternative treatments or result in the denial of treatment. The impact on healthcare providers must also be considered as time spent developing and executing contingency plans results in time away from providing care. Drug shortages have been increasing over the past decade, tripling since 2006, as shown by data collected by the University of Utah at the Drug Shortages Summit in 2010 (ASHP 2010). Data from the University of Utah Hospitals and Clinics revealed that

A.R. Vila-Parrish • J.S. Ivy (✉)
Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC, USA
e-mail: arvila@ncsu.edu; jsivy@ncsu.edu

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_17, © Springer Science+Business Media New York 2013

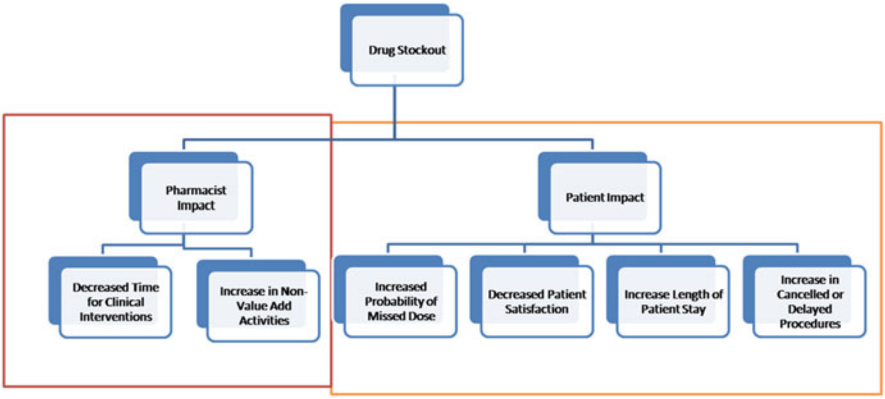

Drug Stockout
Pharmacist Impact
Patient Impact
Decreased Time for Clinical Interventions
Increase in Non-Value Add Activities
Increased Probability of Missed Dose
Decreased Patient Satisfaction
Increase Length of Patient Stay
Increase in Cancelled or Delayed Procedures

**Fig. 17.1** Impact on pharmacists and patients as a result of a drug stockout

25% of buyers' time was spent managing shortages (Choudhary et al. 2011). Finally, it is estimated that medical supplies account for 30–40% of a hospital's budget (Neil 2005). In the last decade it has been estimated that a hospital could reduce its total expenses by approximately 2% through improved inventory and distribution processes (Schneller and Smeltzer 2006).

Inventory management of pharmaceuticals within the hospital pharmacy is a complex task. There may be as many as 2,000 drugs inventoried, many drugs are perishable (i.e., they have a short shelf life), and demand for the drugs changes from day to day. The impact of increased time spent on managing stockouts goes far beyond the pharmacy inventory manager. Figure 17.1 characterizes some of the impacts to pharmacists and patients in the event of drug shortages.

As depicted by the figure, the impact of mismanagement extends beyond healthcare costs (due to waste and shortages) to patient care and utilization of resources (Baumer et al. 2004). Drug availability is a critical factor in a hospital's ability to provide effective, timely, and safe patient outcomes. These issues are so pervasive that in a national survey of 374 U.S. pharmacy directors, 75% of respondents indicated they were forced to purchase the drug off-contract from their current vendor, borrow the drug from another institution, or purchase the drug from an alternative vendor at an increased price when confronted with a drug shortage (Baumer et al. 2004). In addition to the increase in purchasing costs, two-thirds reported delayed or canceled medical procedures due to drug shortages resulting in an interquartile range of $33–300 million dollars in additional costs to the US healthcare system (Baumer et al. 2004). Further, suboptimal inventory systems cause hospital pharmacies to average a low 10.2 inventory turns per year, lose contract compliance opportunities, and continue costly process inefficiencies (Alverson 2003). In fact, most pharmacies average between 8 and 10 turns per year (Blackburn 2010). Blackburn (2010) provides a compelling example regarding the impact of this low number of inventory turns, suggesting that a pharmacy purchasing

$100K per month could save $20,000 per year in on-hand investment dollars (or cash flow savings) with each single-digit increase in the inventory turnover rate. While some drug shortages are uncontrollable (e.g., due to a natural disaster), improper inventory management can result from the lack of procurement expertise of the pharmacists managing the inventory (Alverson 2003).

This chapter will describe the unique modeling challenges associated with inventory decisions for pharmaceuticals, such as their multi-echelon nature, the potential for perishability and/or obsolescence, and internal production/preparation lead times. Further, we will discuss the broader implications of inventory management including the impact of resource constraints and risk, demand uncertainty, and unique outcome considerations such as the impact of supply availability on patient outcomes.

We have organized this chapter by first presenting an introduction to current inventory systems in healthcare, followed by a discussion of supply chain structures, in Sects. 2 and 3, respectively. In Sect. 4, we discuss the current state of research focused on optimization of hospital inventory and conclude with a discussion of areas for future research in Sect. 5. Before we present the aforementioned research, it is important to put the discussion into the context of the current state of hospital inventory management.

## 2 Introduction to Inventory Systems in Hospitals

Simple inventory management strategies are often used to focus the efforts of pharmacy materials managers. For example, classification systems have been developed and implemented in order to create a simple method for improving inventory costs and service levels. Hospitals often use ABC inventory classification strategies, which identify the small fraction of items that account for the highest percentage of expenditures, to include the criticality of the item. In Table 17.1 we describe the types of control procedures commonly used depending on the ABC classification category.

In order to accurately categorize medication inventory, Srinivasan (2008) suggests that substitutable drugs should be considered as a single item in order to simplify their management. Gupta et al. (2007) used the ABC method and the criticality classification analysis called VED, where “V” is for vital items that the hospital relies on to function, “E” is for essential items which can impact the quality of service, and “D” stands for desirable items that do not hinder hospital functionality, and classified nearly 500 drugs procured at the Armed Forces Medical Services in India. They found that if ABC analysis were used in isolation, there would be approximately 22 drugs that would have fallen into the B and C group but were vital to the functionality of their hospital. Al-Qatawneh and Hafeez (2011) took this analysis a step further and combined these classification methods with a continuous replenishment inventory and order-based production control system.

**Table 17.1** Connection between ABC classification category and levels of hospital supply inventory management control and effort (adapted from Reddy 2008)

| Control procedure | A items: high consumption value | B items: moderate consumption value | C items: low consumption value |
|---|---|---|---|
| Degree of safety stock | Very low or stockless strategy, combined with frequent ordering | Low, ordering done on a less frequent basis | High, bulk ordering on an infrequent basis |
| Demand | Should be regularly occurring (e.g., daily) | Monthly | Quarterly |
| Material planning | Accurate and updated regularly | Can rely on historical data | Estimates are sufficient |
| Optimization effort | High, focus on reducing waste, obsolescence, and surplus | Moderate | Review policies annually |
| # of suppliers | High, short lead times required, centralized control | 2–4 sources, shorter lead time preferred | 1–2 sources, decentralized control |

Given that a hospital pharmacy will stock hundreds if not thousands of materials, these classification methods help focus the hospital's limited resources. However, these methods have limitations. After the subset of materials is chosen through ABC/VED analysis, it is still necessary to establish inventory and medication preparation policies.

A number of pharmaceuticals used in hospitals are produced on site. The consideration of the production strategy, make-to-order verses make-to-stock, is complementary to the classification and prioritization of pharmaceutical supplies as well as any inventory management strategy. There has been research regarding make-to-order versus make-to-stock policies for hospital-administered pharmaceuticals, in particular, strategies for premixing doses of intravenous (IV) medications, and the impact of these strategies on waste have been studied. Chiu (2010) examined various IV batching schedules focusing on the resulting waste cost. The IVs were produced and stored as finished goods inventory in a cart which was then sent directly to a patient care unit (PCU). A unique challenge in planning both inventory and production/scheduling strategies was due to the fact that there was a no reuse policy for IVs. Once the IVs left the pharmacy, they could not be returned for reuse in part because of their perishability and stability. Further, Chiu (2010) found that changes in patient condition, discharge, or regimen contributed to an order being discontinued thus becoming waste. These occurrences prompted the author to model the impact of increasing the frequency of batching operations from one time per day to three times per day as well as shifting production/delivery times on waste costs. Model results suggested that preparing finished goods three times per day reduced waste by ~80% or $15,000 per month in the pediatric unit while a less disruptive change, shifting the preparation time and delivery in a one time per day model, yielded a reduction of 47–60%. A close examination of the data showed a strong correlation between time of day and the probability of an order being discontinued. An interesting area for future research would be to develop models to truly understand the linkage between time of day, physician rounds, and discharge times in order to develop more intelligent models and inventory ordering and preparation strategies.

Given the prevalence of inventory challenges, there is a need for modeling and analysis of critical medical supplies, such as pharmaceuticals. Questions such as how often to perform periodic replenishment or how to set adequate periodic replenishment policies or critical levels have not been adequately addressed in the literature (Rosales 2011). Next, we discuss some examples of supply chain strategies that have been instituted in hospital organizations in an attempt to improve their inventory systems.

## 3 Characteristics of Hospital Inventory Systems

As suggested by the above discussion, the unavailability of critical supplies can endanger the health of patients, making the reliability of supply chains within hospitals important. This is a challenging problem because demand is uncertain. Demand for medication is based on the patient population in the hospital which is uncertain and varies through time due to seasonal factors, e.g., flu season. Due to this stochastic demand, inventory levels are set artificially high in order to hedge against uncertainty. Many have described specific impacts of drug shortages and overages on hospital costs and efficiencies. For example, in a large scale study spanning 1 year in a large tertiary hospital, Gillerman and Browning (2000) found of six anesthesia drugs that they tracked, waste costs from five of them contributed to 26% of the total cost of departmental pharmaceutical budget and 2% of the total annual hospital expenditure. Because of the high waste costs for certain drugs, preparation of doses in the dispensed form (i.e., syringe or IV) in advance of demand has decreased (Fraind et al. 2002). However, in a study of IV preparation in European hospitals, many medication errors, such as the use of the wrong diluents, incorrect administration rate, and products that were not mixed, were found when doses were prepared on demand (Cousin et al. 2005). Fraind et al. (2002) and Cousins et al. (2005) suggested having pharmacy departments or suppliers prepare IVs in advance (internally or by a supplier) in order to mitigate some of the mistakes. Though these studies discuss the challenges of managing medications, quantitative models or methods to aid decision making have not been developed or well studied.

In conventional hospital inventory systems, inventory is stored in a central pharmacy. The satellite pharmacies (located in the PCU) place their orders with the central pharmacy, while the central pharmacy orders products from the distributors/manufacturers. In the healthcare context, the term stockless inventory system implies that products are stocked at the distributor or manufacturer. Orders are placed from the hospital's units directly to these upstream supply chain partners.

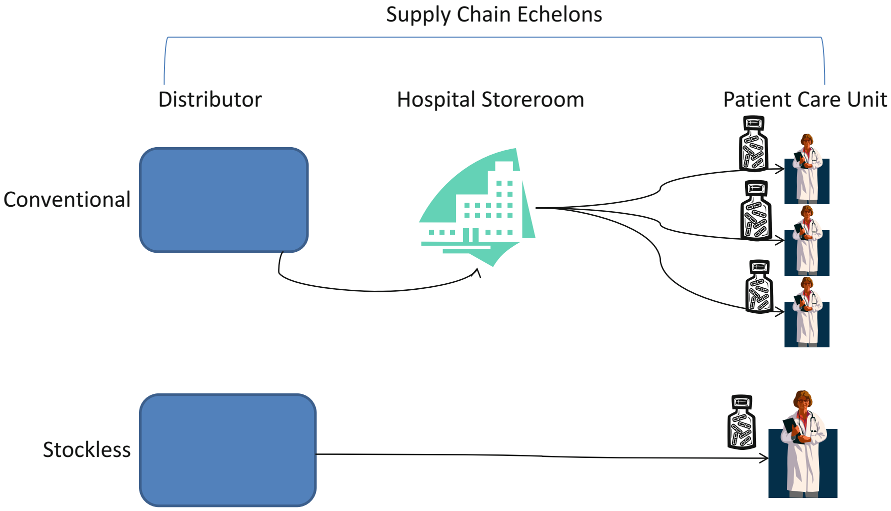


**Fig. 17.2** Comparison of conventional versus stockless healthcare distribution network (adapted from Rivard-Royer et al. 2002)

Stockless inventory systems have been explored in healthcare primarily through case studies and trade journals (see Rivard-Royer et al. 2002; Wade 2011, respectively). Figure 17.2 depicts the conventional (distributor–hospital–PCU) distribution system and a stockless system which bypass the hospital level. The remainder of this section discusses research associated with each of these types of systems. Our discussion of these distribution systems will focus on the movement of drug within the hospital, i.e., from the hospital storeroom to the PCU, in the context of a conventional system and from the distributor directly to the PCU in the stockless system. Similar to other industries, inventory management of hospital supplies is being outsourced to a variety of entities, such as the manufacturers of automated dispensing equipment.

## *3.1 Conventional Inventory Systems: Automated Systems*

The focus of our discussion of conventional inventory systems will be on the intermediary distribution process, i.e., the movement and management of drug inventory from the hospital storeroom to the PCU, and systems implemented to enhance this process. Automated drug delivery systems describe a class of technologies that facilitate this aspect of conventional distribution systems commonly found in hospitals. In an attempt to improve their drug delivery system, many hospitals have implemented automated systems such as Pyxis machines (Handfield 2007). Pyxis is an automated medication management system that dispenses prescribed patient medication to clinicians for administration to the patient. While these machines allow for greater efficiency in comparison to drug delivery in patient-specific trays, they are labor intensive and inefficient without the implementation of proper inventory management policies. Inaccurate inventory policies necessitate refilling the machine several times per day due to machine stockouts. Handfield (2007) found that many hospitals have not experienced the anticipated benefits of optimizing inventory by simply deploying Pyxis; in fact, some have reported decreased performance. Similar to Pyxis machines, perpetual inventory systems while providing real-time information to hospital management regarding supply levels do not provide guidance regarding inventory policy. The potential of these types of systems cannot be realized without a robust inventory policy to guide decision makers to make the best use of the real-time information. While there have been efforts to automate the dispensing of drugs, computerize ordering transactions, and electronically track medical supplies and goods, there has been limited quantitative analysis or optimization of inventory management.

### *3.2 Stockless Inventory Systems*

Stockless inventory systems take the automated drug delivery systems one step further by completing removing the drug distribution process from the hospital to the distributor. Rivard-Royer et al. (2002) cited many studies that discuss the staff reductions and higher service levels achieved when stockless inventory systems are employed. However, Wade (2011) discussed the complexities involved in the outsourcing decision-making process citing factors such as cost, preparation time, forecasting, and shelf life as drivers in the decision to outsource medication compounding.

As a result of the challenges associated with a pure stockless strategy, Rivard-Royer et al. (2002) proposed a hybrid approach "combining the stockless method with a conventional approach to PCU replenishment." This hybrid strategy entailed distribution of products with case quantities by distributors while bulk purchases of low-volume products were broken down into point-of-use increments at the hospital's central pharmacy (Rivard-Royer et al. refer to the central pharmacy as a storeroom). While their study showed some cost savings, the authors highlighted the need for further integration between all supply chain partners (including the manufacturer) and the development IT infrastructure (e.g., HL7) to communicate effectively between these groups.

## 4 Pharmaceutical Inventory Policy Optimization

The optimization of inventory systems in healthcare has been explored under various assumptions and supply chain structures such as conventional distribution systems and centralized stockless systems described above. The majority of the

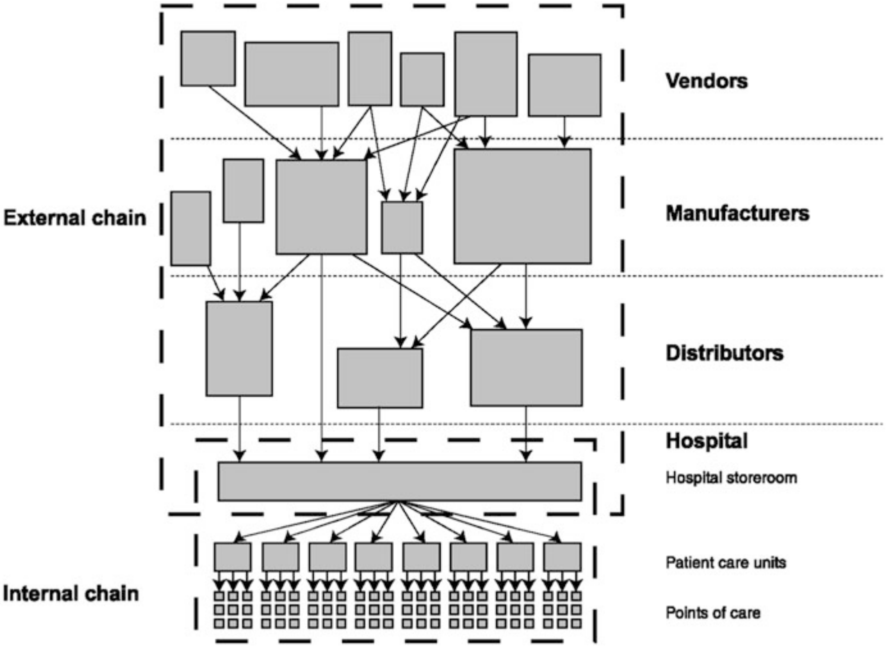


**Fig. 17.3** Healthcare sector supply chain. *Source*: Rivard-Royer et al. (2002)

literature that discusses multiple echelon supply chains defines the echelons as various physical locations. For example, inventory may be held in various locations throughout the supply chain: manufacturer, distributer, hospital's central pharmacy, and PCU pharmacy. Rivard-Royer et al. (2002) represented the healthcare supply chain echelons focusing on the internal and external supply chain partners as shown in Fig. 17.3.

Within the healthcare supply chain, particularly at the hospital level, fulfilling patient needs in a timely and accurate manner is the top priority, and hence inventory management is critical. As motivated in Fig. 17.3, inventory management in hospitals involves a number of interacting entities. Lapierre and Ruiz (2007) is an example of a study that has taken an integrative approach to managing hospital supply decisions by considering factors beyond costs that influence hospital operations. Lapierre and Ruiz (2007) explicitly modeled the impact of inventory management on resource balancing and scheduling of hospital supply management activities. Two models, one with and one without (i.e., inventory cost focused) workload balancing in the objective function were developed in order to compare inventory quantities by location, time, and supplier. They found that taking these factors into account resulted in more evenly balanced workload schedules with little increase in inventory costs.

The hospital inventory system and the supply chain within which it operates is complex. Demand originates at the patient level and is a defining factor in inventory optimization modeling. The majority of the medical supply chain research to date has assumed that demand is stationary (i.e., static) and independent of the system dynamics. However, a number of factors cause demand to be nonstationary. For example, patient condition dynamics change over time affecting their drug needs; there is seasonality in demand (i.e., the number of patients in the hospital) due to events such as flu season; and dosing requirements vary as a function of dynamic patient characteristics such as age, weight, blood pressure, and platelet count, as such demand is a function of the patient population mix as well as the individual patient dynamics. In the following subsection we discuss these operations research models with stationary demand and multiple echelons.

## *4.1 Stationary Demand*

Nicholson et al. (2004) considered the impact of supply chain structure on inventory policy. They developed two inventory models in order to compare a 3-echelon (distributor–hospital–patient unit) distribution system, termed Scenario A (similar to the conventional inventory system shown in Fig. 17.2), to a 2-echelon (distributor–patient unit) distribution model, termed Scenario B (similar to the stockless system shown in Fig. 17.2), for non-critical items. The decision variables were the par levels (i.e., the amount of material needed on hand to satisfy demand) for each of the echelons depending on the specified scenario for a single review period where shipments are assumed to arrive instantaneously. Any back orders were assumed to be fulfilled via an emergency shipment, and the minimum service level was set to 90% which the authors' state is indicative of current practices in hospitals for non-critical items. Since the problem has a non-convex constraint set representing hospital unit specific service levels, it has been shown to be NP-hard, and thus, greedy heuristics were developed to obtain feasible solutions. Their results showed an inventory cost savings (without loss of service) when outsourcing occurs (i.e., from the distributor/supplier directly to the patient unit). Additionally, based on observations there was a benefit to the workload of healthcare providers when Scenario B was implemented—increasing the staff's time with patient care. However, the authors also noted the need for coordinated communications if such a distribution network is to succeed.

Little and Coughlan (2008) developed a constraint programming optimization model for determining optimal stock levels for hospital supplies considering storage space restrictions, item criticality, and delivery frequency. Their proposed inventory control policy was a standard base stock policy whose parameters were found by optimizing the minimum (or average) service level under these constraints where the demand distribution for each item was assumed to be known and stationary. Their model attempted to capture the various objectives of the many stakeholders invested in hospital inventory management by incorporating these constraints and a non-cost-based service level objective in a bounded knapsack problem where there are different types of items (i.e., drugs), each with a weight and value. Although weight corresponds to volume in this context, Little and Coughlan argue that value is not simply a summation but that the objective relates to a measure of service level across all products. The resulting inventory policies were characterized according to the percentage of items at each service level, the average service level, and total amount of space used. Little and Coughlan (2008) showed that the same service levels can be reached by delivering every day with low space usage compared with delivering every three or five days with a high space usage. They suggest the need for good quality data that can provide real-time information between inventory systems and complex demand patterns.

Bijvank (2009) compared the performance of a fixed order size inventory policy (FOSP) to an ($s$, $S$) policy; both consider service level and capacity restrictions for hospital inventory systems. In an ($s$, $S$) policy, the firm places an order when the inventory level reaches $s$, and they order up to the quantity $S$. The focus of this work was on inventory policies for disposable items that are distributed throughout the hospital such as gloves, needles and sutures, and readily available (i.e., negligible lead time) items. These items are commonly stored in bins, and thus, capacity ($C$) is dictated by the bin size. An initial single-item model was derived and then extended to consider a multi-item system. In this multi-item system, the decision was how many bins to allocate to which product—becoming a knapsack-type problem. The two types of replenishment policies are evaluated for infusion liquids (i.e., fluids used in an injection or IV) at three point-of-use locations. The best performing policy was an ($s$, $S$), where ($s = C - 1$), policy; however, the high frequency of orders was prohibitive to implementation. Thus a new reorder level was found that resulted in a similar order frequency to the FOSP. An inventory rule consisting of three steps was developed to simplify the search for the near-optimal parameters, making the problem-solving process easy for practitioners to implement.

Rosales (2011) compared ordering medications out-of-cycle (e.g., ad hoc) and periodic/continuous review replenishment policies. Rosales (2011) developed a hybrid policy where inventory is replenished periodically at the beginning of every period (e.g., shift) using an ($s$, $S$) inventory policy. However, in contrast to a typical ($s$, $S$) policy, if the inventory position of a particular item reached a threshold level $R$, an out-of-cycle replenishment was placed which followed an ($R$, $Q$) policy. The key finding was that a hybrid policy outperformed both pure periodic or continuous review policies. Extending this work to a multi-item context, Rosales (2011) also characterized the benefit attained if other items were included for restocking once an out-of-cycle order was triggered for a single item. The evaluation of four hybrid-joint replenishment strategies revealed that most of the benefits came from joint replenishment during an out-of-cycle order.

DeScioli (2005) evaluated the performance of various inventory policies in automated point-of-use systems (such as Pyxis machines) under conditions of intermittent demand modeled using Croston's method. DeScioli found that an ($s$, $Q$) policy, where $Q$ was found using an EOQ model, performed best under these conditions. The consideration of sporadic demand takes a step toward a dynamic demand model that incorporates nonstationary demand.

## *4.2 Nonstationary Demand*

As discussed earlier, demand originates at the patient level. Demand for the medication depends upon not only the number of patients but also the condition of those patients and their length of stay. The patient condition, length of stay, and the patient mix are each dynamic, stochastic, and change over time. In a departure from the above literature which assumes stationary demand, Duclos (1993) developed a simulation model of a hospital inventory system in order to explore the impact of operations under normal and emergency demand situations (e.g., system shocks). They considered policies that evaluate the cost of operating under various inventory review intervals for both the central store and point-of-use. Duclos (1993) found that changing the review frequency was an important factor in reducing the number of stockouts and expedited fill requirements under these demand conditions.

Vila-Parrish et al. (2008) studied a dynamic inventory model that explicitly considers the link between patient condition, patient admissions, and the resulting demand for perishable medications. In their study, the echelons were representative of the physical state of a pharmaceutical drug product (e.g., raw material and finished good). These products, many of which must be prepared to be administered intravenously, become highly perishable as a result of the preparation of the “raw materials.” The simulation–optimization model developed demonstrated that incorporating knowledge of the changes in patient condition that impact medication demand improved the inventory costs.

Vila-Parrish (2010) developed one of the first analytical models to consider nonstationary demand in the context of a multi-echelon, perishable inventory management. Extending the simulation model developed in Vila-Parrish et al. (2008), Vila-Parrish (2010) and Vila-Parrish et al. (2012) developed a Markov decision process model where the state space was defined by not only the inventory on hand but also the number of patients in a PCU. The items of interest were perishable intravenous medications that expire within 24 h of preparation. This model included decision variables for ordering as well as preparation of medications in advance and in response to patient demand for a conventional inventory system. Two demand fulfillment strategies were explored: (1) external fulfillment—shortages are expedited from an external source (e.g., another hospital) only and (2) hybrid fulfillment—shortages are fulfilled internally via an additional production run, and if there is insufficient inventory in-house, the balance is fulfilled externally. Optimal policy structures were derived that were nonstationary and showed the dependence of ordering and preparation quantities on the patients in the system and the available fulfillment strategies. The findings in Vila-Parrish (2010) suggested that a hybrid policy that includes a combination of make-to-order and make-to-stock product may best fit the hospital environment requirements to minimize waste while satisfying dynamic patient needs in a timely fashion. Lastly, the value of investing in resources or infrastructure to execute a hybrid fulfillment strategy can be found by comparing the costs from each model.

In Vila-Parrish et al. (2012) this modeling framework was applied to the demand for an antibiotic from patients in a particular PCU of a large community hospital. The data analysis from 1,427 patient encounters for 154 distinct patients revealed a statistically significant correlation between the number of admitted patients and number of doses. Two representations of patient demand were modeled; in the first, patient demand is represented by a stationary aggregate demand distribution, and in the second, a Markov chain was developed to represent the change in the number of patients in the each unit period. The number of patients (i.e., the state of the Markov chain) corresponded to a unique demand distribution derived from the dataset and was shown to be a statistically significant predictor of demand for the antibiotic. The results of this analysis showed a cost advantage to using the nonstationary Markov chain-based demand model, and further research to include more patient information that may predict demand in the state definition was suggested. In the following section we present a small example motivated by this application presented in Vila-Parrish et al. (2012).

### *4.3 Meropenem Case Study*

We present an example of a nonstationary demand process driven by patient dynamics within a hospital. As discussed in Sect. 4.2, Vila-Parrish et al. (2012) analyzed data from a large urban hospital in order to model the patient demand process for meropenem and proposed a nonstationary model for managing the drug. Meropenem is an antibiotic used for complex bacterial infections such as bacterial meningitis which costs approximately $30 per 1-g vial. In addition to being fairly costly, meropenem becomes highly perishable when it is converted into an intravenous form—depending on the diluents and storage; its shelf life varies from 1 to 24 h.

The focus of this study was to characterize the meropenem usage for patients in the progressive intensive care unit (PICU) during 2008 and use this information for improving the understanding of the drug demand dynamics. There were over 1,400 patient encounters (where an encounter was defined as each time a patient received the drug) for over 150 patients. The analysis of the drug demand for these patients over time suggested a dynamic and nonstationary demand pattern that corresponded nonlinearly to the number of the patients receiving meropenem. The demand distributions derived from the data as a function of the number of patients are shown in Fig. 17.4. Note the distributions for the number of patients are not simple convolutions of the one patient demand distribution. Figure 17.5 shows the state transition diagram corresponding to the patient-based Markov chain developed where the state was the number of patients and each state corresponds to a demand distribution shown in Fig. 17.4.

The analysis in Vila-Parrish et al. (2012) shows the number of patients in the hospital per day is stochastic, and hence, the meropenem demand is nonstationary. The inventory and production policies for this case were state-dependent (i.e.,

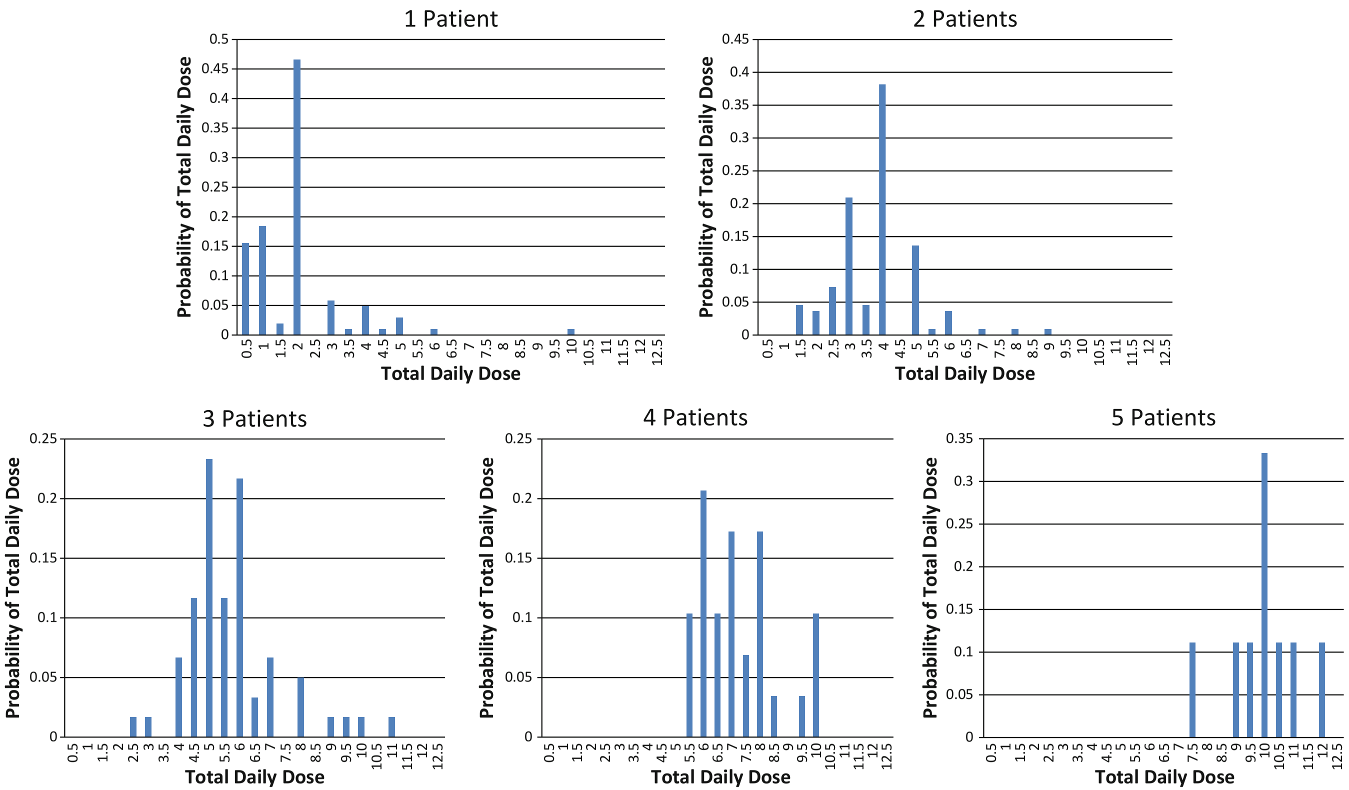


**Fig. 17.4** Total daily meropenem demand distributions as a function of the number of patients in the PICU. The *y*-axis corresponds to the probability of the total daily dose, and the *x*-axis corresponds to the total daily dose in grams (Vila-Parrish et al. 2012)

**Fig. 17.5** Example of a Markov chain in which each state is the number of patients receiving meropenem per day, up to a maximum of 5

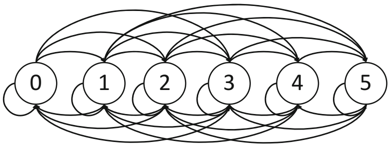


**Table 17.2** Percent change in total expected cost of the dynamic model evaluated with the policies derived by assuming that the demand model is stationary

| Ratio of internal expediting cost to external expediting cost | 1/25 | 1/10 | 1/5 | 2/5 | 3/5 | 4/5 |
|---|---|---|---|---|---|---|
| %Change in total cost | 2.59% | 7.08% | 10.86% | 15.08% | 17.23% | 18.12% |

dependent on the number of patients in the system). The state-dependent policies are then compared to policies derived from an aggregate demand model that ignores the number of patients in the system. Using an internal to external expediting cost ratio of 1:10, the percent change in total expected cost of the dynamic model evaluated with the policies derived by the stationary demand model is shown in Table 17.2. The state-dependent policies outperform all stationary policies. Further, the more similar the internal and external expediting costs are to each other, the worse the stationary policy performs.

The key take-away from this example is that using a patient-driven inventory and production model yields a total cost improvement over a stationary policy that uses an aggregate demand distribution.

## 5 Open Research Challenges

While there have been efforts to incorporate technological advances for managing the pharmaceutical supply chain, e.g., electronic data collection and equipment tracking, the information provided by these innovations is often not used to inform supply management decision making and improve the inventory policy. One critical area for future research will be to consider how to optimize the entire supply network (shown in Fig. 17.3). Improved hospital-based inventory management will result in more accurate ordering by hospitals as well as more informed relationships with suppliers and distributors. While many hospitals belong to group purchasing organizations (GPOs) to improve their purchasing contracts, they have not linked the entire supply chain. GPOs have primarily been focused on purchasing and contracting decisions but have not been utilized fully to influence the pharmaceutical supply chain. As hospitals communicate with each other through GPOs, they could create their own supply chain—sharing inventory across hospitals—and, as a result, could achieve greater efficiency and cost savings by jointly managing inventory across the group.

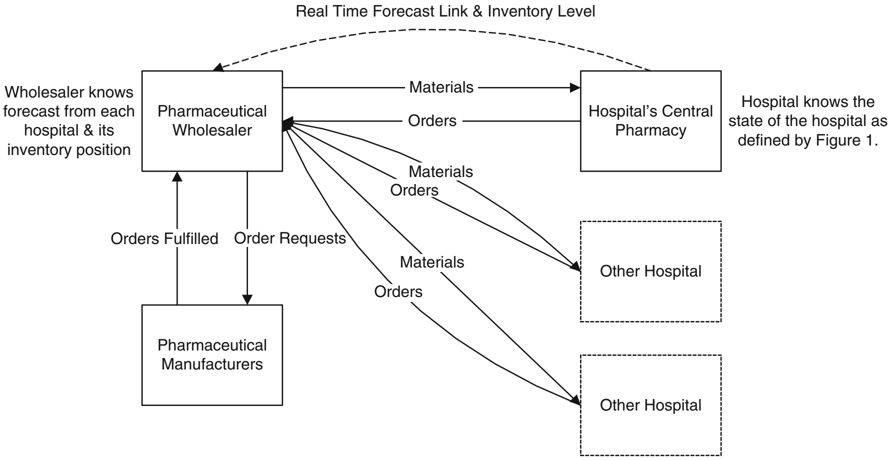

Real Time Forecast Link & Inventory Level
Wholesaler knows forecast from each hospital & its inventory position
Pharmaceutical Wholesaler
Materials
Orders
Hospital's Central Pharmacy
Hospital knows the state of the hospital as defined by Figure 1.
Materials
Orders
Other Hospital
Orders Fulfilled
Order Requests
Materials
Orders
Pharmaceutical Manufacturers
Other Hospital

**Fig. 17.6** High-level material and information flow considering hospital, wholesaler, and manufacturer network

Figure 17.6 shows a potential future healthcare supply network with increased visibility that may result in part from improved inventory management by hospitals and hospital networks. Increased inventory visibility among supply chain partners due to real-time information links between the central pharmacy and the pharmaceutical wholesaler can improve the overall management and allocation of a finite supply of pharmaceuticals over time.

Further, at the hospital level, there is a need to use improved information to understand the link between patient arrivals and condition and demand for medical supplies. The development of real-time forecasting, ordering, and order fulfillment systems that consider patient information as an input is an opportunity for research.

As this discussion of the current literature presented in this chapter implies, healthcare supply networks present a rich field for future research. There have been few analytical studies that consider the complexities that are inherent in pharmaceutical inventory control such as dynamic nonstationary demand, multiple stakeholders, and the fact that the key customer, the patient, is not the payer. The development of models that integrate these dynamics will be necessary for hospitals to understand and control their pharmaceutical supply costs, improve patient care, and efficiently use resources.

## References

Al-Qatawneh L, Hafeez K (2011) Healthcare logistics cost optimization using a multi-criteria inventory classification. In: Proceedings of the 2011 International Conference on Industrial Engineering and operations management, Kuala Lumpur, January 2011

Alverson C (2003). Beyond purchasing—managing hospital inventory. Managed healthcare executive. Retrieved on 01 January 2008: http://www.managedhealthcareexecutive.com/mhe/article/articleDetail.jsp?id=75802

American Society of Health-System Pharmacists (ASHP) (2010) Drug summit summary report, November 5, 2010. www.ashp.org/drugshortages/summitreport. Accessed 29 Sep 2011

Baumer AM, Clark AM, Witmer DR, Geize SB, Vermeulen LC, Deffenbaugh JH (2004) National survey of the impact of drug shortages in acute care hospitals. Am J Health Syst Pharm 61(19):2015–2022

Bijvank M (2009) Service inventory management solution techniques for inventory systems without backorders. Ph.D. Dissertation, University of Amsterdam

Blackburn J (2010) Fundamentals of purchasing and inventory control for certified pharmacy technicians: a knowledge based course. ACPE No. 0096-9999-10-051-H04-T. Release date 21 June 2010. Accessed from https://secure.jdeducation.com/JDCourseMaterial/FundPurch.pdf

Chiu C (2010) The effects of intravenous admixture batching schedules on waste—a computer simulation approach. Master's Thesis, University of Cincinnati

Choudhary K, Fox ER, Wheeler M (2011) Proactive strategies for managing drug shortages. Pharm Purchasing Prod 8(2):8–13

Cousin DH, Sabatier B, Begue D, Schmitt C, Hoppe-Tichy T (2005) Medication errors in intravenous drug preparation and administration: a multicentre audit in the UK, Germany, and France. Qual Saf Health Care 14:190–195

Descioli D (2005) Differentiating the hospital supply chain for enhance performance. Master's Thesis, Massachusetts Institute of Technology

Duclos LK (1993) Hospital inventory management for emergency demand. J Supply Chain Manag 29(4):30–37

Fraind DB, Slagle JM, Tubbesing VA, Hughes SA, Weinger MB (2002) Reengineering intravenous drug and fluid administration processes in the operating room: step one: task analysis of existing processes. Anesthesiology 97(1):139–147

Gillerman RG, Browning RA (2000) Drug use inefficiency: a hidden source of wasted health care dollars. Anesth Analg 91(4):921–924

Gupta R, Gupta KK, Jain BR, Garg RK (2007) ABC and VED analysis in medical stores inventory control. Med J Armed Forces India 63:325–327

Handfield R (2007) New trends in medical dispensing technology: reducing the total cost of patient care, white paper, supply chain resource cooperative. College of Management, North Carolina State University

Lapierre SD, Ruiz AB (2007) Scheduling logistic activities to improve hospital supply systems. Comput Oper Res 34:624–641

Little J, Coughlan B (2008) Optimal inventory policy within hospital space constraints. Health Care Manag Sci 11:177–183

Neil R (2005) Managing costs and building consensus CEOs add strong link to supply chain. Mater Manag Health Care 18–21. Accessed 7 Dec 2005. http://www.soundingboard4life.com/pdf/MatlMngmntNov2005Flynn.pdf

Nicholson L, Vakharia A, Erenguc S (2004) Outsourcing inventory management decisions in healthcare: models and application. Eur J Oper Res 154:271–290

Reddy VV (2008) Managing a modern hospital. Sage, New Delhi (Chapter 6)

Rivard-Royer H, Landry S, Beaulieu M (2002) Hybrid stockless: a case study: lessons for health-care supply chain integration. Int J Oper Prod Manag 22(4):412

Rosales CR (2011) Technology enabled new inventory control policies in hospitals. Ph.D. Dissertation, University of Cincinnati

Schneller ES, Smeltzer LR (2006) Strategic management of the health care supply chain. Jossey-Bass, San Francisco

Srinivasan AV (2008) Managing a modern hospital, 2nd edn. Sage, New Delhi, Thousand Oaks, Calif

Vila-Parrish AR (2010) Dynamic inventory management policies for perishable and short lifecycle products under demand uncertainty. Ph.D. Dissertation, North Carolina State University

Vila-Parrish AR, Ivy JS, King RE (2008) A simulation-based approach for inventory modeling of perishable pharmaceuticals. In: Proceedings of the 2008 Winter Simulation Conference, Miami

Vila-Parrish AR, Ivy JS, King RE, Abel S (2012) Patient-based pharmaceutical inventory management: a two-stage inventory and production model for perishable products with Markovian demand. Health Syst 1:69–83

Wade J (2011) Outsource compounding preparations to reduce waste. Pharm Purchasing Prod 8–13:24–25

# Chapter 18
# The Challenges of Hospital Supply Chain Management, from Central Stores to Nursing Units

**Sylvain Landry and Martin Beaulieu**

## 1 Introduction

In the vast majority of countries, the healthcare sector is the focus of a great deal of attention from public decision makers and media alike. Although healthcare is by definition a clinically driven environment, the practice of patient care is supported by a range of activities that notably include purchasing, inventory management, and the distribution of supplies to the point of care. These activities are associated with healthcare supply chain management, also referred to by many as healthcare logistics. Improving the efficiency of such logistics can provide opportunities for healthcare institutions and health systems to increase the quality of care and reduce costs.

This chapter will describe the challenges of healthcare supply chain management with a focus on the hospital's internal supply chain and more specifically on the distribution of medical supplies from the central storeroom to nursing units (point of care). Section 2 provides background information on healthcare supply chain management, with particular reference made to the efficient healthcare consumer response (EHCR) report, the first industry-wide report on healthcare supply chain integration. Section 2 goes on to discuss the complexities of the internal hospital supply chain and addresses how the materials management function overseeing this activity is structured. Section 3 covers the challenges and methods of distributing medical supplies to nursing units. Sections 4 and 5 identify best practices and future research opportunities. Section 6 concludes the chapter.

S. Landry (✉) • M. Beaulieu
HEC Montréal, Montréal, Canada
e-mail: sylvain.landry@hec.ca

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_18, © Springer Science+Business Media New York 2013

## 2 Background on Healthcare Supply Chain Management

Hospitals are much more than simply a link in the healthcare supply chain. They are on the receiving end of a wide range of supplies that support the delivery of care. This section will address the challenges of the internal and external supply chain and will examine how these activities and processes within the hospital are structured.

### *2.1 Efficient Healthcare Consumer Response*

Publication in the USA of the EHCR report in 1996 marked a turning point in healthcare supply chain management. This analysis presented a global vision of the supply chain in the sector by placing particular focus on medical and surgical supplies and pharmaceutical products (CSC Consulting 1996). The document followed on the heels of studies conducted previously in other sectors, such as those in the apparel industry supply chain in 1986, which led to the Quick Response movement (Blackburn 1991; Hunter and Valentino 1995) and in the (nonperishable) food industry in 1993, which prompted the Efficient Consumer Response (ECR) report (Kurt Salmon Associates Inc. 1993). The EHCR report was largely inspired by the ECR report; however, healthcare issues are much different from those faced by retail businesses, if only in the identification of the consumer, who may be the patient (to whom the supplies are directed and in certain cases charged), the healthcare professional (who prescribes or uses the products or supplies), or the taxpayers, employers, government programs, or insurance companies (who pay).

The vision put forward in the EHCR report brought to light supply chain inefficiencies shared by manufacturers, distributors, and healthcare providers, including duplication of tasks, multiple storage areas, a fragmented information flow, delays of all types, and substandard service. The study itself led to the creation of a number of committees and expectations for changes in the strategies deployed by the stakeholders [providers, distributors, manufacturers, group purchasing organizations (GPOs), etc.]. In turn, substantial savings were expected—savings that at the time were estimated at $11 billion across the US health system or almost half of the costs associated with documented logistics process. However, once the new strategies had been drafted, the corporate priorities of the stakeholders involved, many of them competitors, led a large number of decision makers to take isolated action and dissolve the EHCR committees a few years after their creation (Landry and Beaulieu 2008).

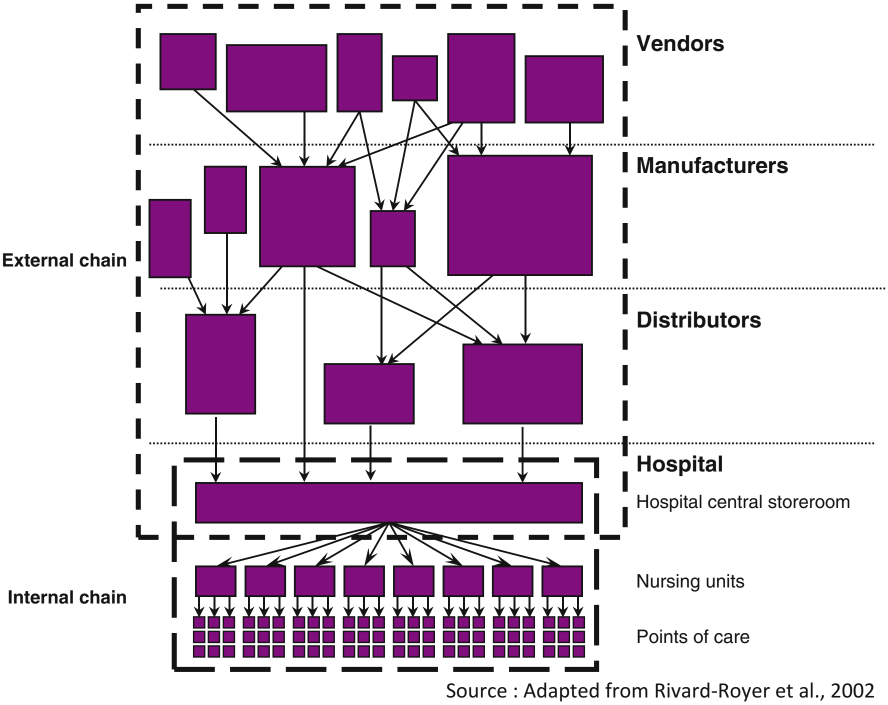
In 2009, a team of researchers from the University of Arkansas Center for Innovation in Healthcare Logistics (CIHL) published a report on the advancement of healthcare logistics practices since the EHCR. The report stated that "despite this effort, a lack of clear and measurable cost and quality improvements is evident within the industry" (Nachtmann and Pohl 2009). Almost half of the respondents, most of whom were employed by hospitals or health systems in director-level positions, indicated that their organization's supply chain was at a low level of maturity (i.e., neither linked nor integrated/extended). Moreover, the report concluded that the healthcare supply chain was starved for accurate and accessible data, the lack of data standards being a major hurdle. However, there was evidence of implementing the strategies recommended in the 1996 report. Indeed, according to this report, 41% of respondents had attempted at least half of the suggested EHCR strategic initiatives, such as e-commerce implementation and supply chain automation, and had achieved performance improvements.

## *2.2 Hospital Supply Chain Management Challenges*

Although the trend is moving toward the establishment of a continuum of care among multiple healthcare providers, such as outpatient clinics, acute care hospitals, and nursing homes within a health system or an integrated delivery network (IDN), hospitals remain the backbone of these systems and present multifaceted supply chain management challenges. In 2009, for example, the just under 6,000 hospitals in the USA (AHA 2010) accounted for close to one third of the national health budget and were the primary expenditure item (California Healthcare Foundation 2011). The challenges these institutions face pertain to the integration of the external supply chain as defined in the EHCR report, that is, manufacturers, distributors, and healthcare providers, as well as the integration of hospital's internal supply chain (Fig. 18.1).

The hospital is much more than simply a link in the supply chain (Landry and Beaulieu 2007), and its internal supply chain is highly complex. Hospitals are generally structured around clinical departments such as emergency, intensive care, oncology, cardiology or coronary care, the catheterization laboratory or cath lab, and surgery [performed in operating rooms (ORs)], with inpatient beds organized in wards or nursing units averaging two dozen beds each. These departments and nursing units must have on hand pharmaceutical products and medical supplies to support patient care, and these products and supplies go through a series of steps before they reach the end user (clinical staff or patient) for consumption. Medical supplies tend to come under the responsibility of the materials management department and most often must be processed by receiving and central stores before being delivered to end users. (An exception is pharmaceutical Thitchie et al. 2000 products, which are processed and managed by the pharmacy and are discussed in Vila-Parish and Ive and therefore fall outside the scope of this chapter.)

Nursing units generally have a main storeroom where medical supplies are kept. However, this room is rarely the final storage position, as secondary storage points located closer to the point of use throughout the unit cater to the specific needs of clinical staff. These points, which are replenished with supplies drawn from the main storage room, may take several forms, from mobile carts that transport supplies from patient to patient to stationary storage units in patient rooms.



Source : Adapted from Rivard-Royer et al., 2002

**Fig. 18.1** Healthcare sector supply chain

In addition, certain supplies commonly known as nonstock items (or direct purchases) are, as their name implies, not stored in central stores but rather delivered directly to a specific nursing unit shortly after being received, often because the unit is the sole user of these supplies. Other supplies, such as food, linens, and surgical instruments, must go through a "transformation" process before being delivered to users (e.g., cooking, washing, or sterilization). After use, certain supplies must go through what is called reverse logistics to be transformed once again (e.g., linens), while other supplies become a waste management issue (e.g., cardboard). Hospitals produce a variety of waste matter (biomedical, chemical, metal, etc.) (Tudor et al. 2009), which often must be managed according to regulatory standards (Tyagi et al. 2010). For ecological reasons, hospitals also develop practices to reduce environmental impacts, such as unpacking and recycling cardboard boxes in central stores to promote paper recuperation and prevent boxes from accumulating in nursing units (Tudor et al. 2008).

Hospitals therefore receive a wide range of supplies that support the delivery of care, either directly (medical supplies, pharmaceutical products) or indirectly (linens, meals, stationery, cleaning products). In most cases these supplies carry a high level of awareness as to the risks of stocking out. Moreover, the many different flows of information and material in a hospital have resulted in a range of clinical staff contributing to the logistics activities associated with the various supplies used. This means that within a hospital, almost everyone is involved in the supply chain, although few realize it (Landry and Beaulieu 2002). Among the professionals involved, clinical staff often have neither the expertise nor resources to efficiently manage logistics activities. And, considering that most industrialized countries are facing nursing shortages, it is vital to find ways to ensure that all of the efforts of clinical staff are channeled toward patient care. Instead, many of these employees currently spend more than 10% of their time on logistics tasks (Chow and Heaver 1994; Rivard-Royer et al. 2002; Fereng 2010). Added to this is the fact that nurses are often interrupted in their work because of supply shortages and other logistics problems (Tucker and Edmondson 2003).

The diversity of flows and the dispersal of logistics activities among the various departments and nursing units in a hospital tend to inflate the costs associated with these activities. In fact, North American studies have found that more than 40% of a hospital's expenses are related to supply chain activities (Chow and Heaver 1994; Nachtmann and Pohl 2009; AHRMM 2010). Similar studies conducted in France and Holland have revealed that 30–35% of a hospital's operating budget is spent on logistics (Bourgeon et al. 2001).

Not all products are alike; their cost or the impact of a shortage will vary (Schneller and Smeltzer 2006). In some cases, the material manager must change strategies to take advantage of any savings that may be available. Tendering strategies and the consolidation of suppliers are primarily used for commodity products (Pedersen 1996). It remains that physicians' preference items (PPI), that is, supplies and expensive disposable items used during surgical procedures, hold the greatest potential for savings in hospitals (DeJohn 2005). Indeed, surgeons' decisions are frequently based on factors unrelated to cost, such as their experience with a particular product, their sense of what is in the best interests of a particular patient, or their relationship with a manufacturer's representative (Montgomery and Schneller 2007). Savings on these products can be generated through supply standardization strategies or utilization management (Governance Committee 1997). However, such strategies ultimately depend on physician participation (Montgomery and Schneller 2007; Aston 2010).

The hospital, due to the particularities of its internal supply chain, consequently merits greater attention, and solutions are needed that address its unique situation. The benefits that can be generated through sound management of the hospital supply chain are equally unique. Whereas using effective logistics in the industrial and retail sectors can lead to reduced costs and increased customer service, efficient logistics in the healthcare sector can yield other substantial gains. For example, improved logistics in this sector can become a tool to enhance job satisfaction among clinical staff. To this end, given the complexity and challenges being faced by the healthcare sector, it is important to integrate the internal supply chain in order to fully benefit from its integration with the external chain (Schneller and Smeltzer 2006).

## *2.3 Organizational Structure of the Materials Management Function*

In North America, as reported by Landry and Beaulieu (2002), the emergence of the materials management or logistics department in its current form is the result of many changes that have taken place over a 100-year period within the hospital environment. Indeed, "in the early 1920s, the American College of Surgeons endorsed the concept of standardized surgical dressings and the centralized preparation and handling of all surgical supplies" (Thorsfeldt 1988, p. 64). By the 1940s, W.R. Underwood and others were paving the way for a central service organization (Thorsfeldt 1988), and in the 1970s, analysts became proponents of a centralized system to manage purchases, inventory, and distribution in hospitals (Driscoll 1981). During the same decade, this central department evolved further, integrating new functions and becoming the materials management department; one of the functions rolled into it was purchasing, although this emerging organizational structure varied from hospital to hospital (Thorsfeldt 1988).

Prior to 1950, few hospitals had a centralized purchasing department; each department managed its own purchases and inventory. However, this approach was feasible only when a limited assortment of products was involved. The technological evolution that followed the Second World War brought with it a surge in the range of products available as well as an increase in deliveries, further complicating the management of supplies. To eliminate duplication and labor costs, hospitals turned to a centralized purchasing approach (Burnette 1994).

The consolidation of a central service and purchasing unit within a new materials management department responded to a need for greater efficiency and productivity by eliminating waste and duplication in the management of material flows (Thorsfeldt 1988). It also put responsibility for purchasing supplies into the hands of professionals, which in turn led to an increased knowledge of the markets (Fearon and Ayres 1967). These centralization efforts gave birth to what we know today as healthcare supply chain management or healthcare logistics, which take concrete form through these various terms and the different roles that logistics play within a hospital. Yet, despite the clear benefits of assigning an increasing number of activities to the materials management, logistics, or supply chain department, as noted above, we continue to see many hospitals dealing with fragmented logistics activities and a host of players (Parker and DeLay 2005), which can impede the emergence of a single, credible entity to handle supply chain activities.

Moreover, in many sectors, including healthcare, the purchasing function is largely ignored by senior management (Bales and Fearon 1993; Cammish and Keough 1991). Often, responsibility for this department is assigned to a middle manager, whose authority is by definition limited (Janson 1985). In more than half of the hospitals in the USA, the materials management department reports to the financial director or CFO, which means that the manager must negotiate with decision makers in other departments as well as with his or her own internal clients (Kowalski 1993; AHRMM 2000; HFMA 2010). Chow and Heaver (1994) agree, stating that this situation initially leads to an inability to implement solutions that involve both the hospital and its suppliers and goes on to disrupt operational activities, as the department is left to find its own solutions.

Recent developments in the healthcare sector demonstrate the extent to which supply chain management is gaining the attention of leaders. In Canada and the USA, healthcare reforms are prompting a number of materials management departments to outsource a portion of what were once their traditional activities. Most departments, for example, turn to a group purchasing organization (GPO) to find suppliers and negotiate contracts; others use the stockless approach, calling on medical supply distributors to deliver products directly to nursing units (Arthur Andersen 1990; Souhrada 1998; Rivard-Royer et al. 2002). In recent years, we have also seen the emergence of consolidated service centers (AHRMM 2010), which use a shared services or 3PL approach to serve many institutions in the same province or state or indeed in several states. More and more hospitals are also creating materiel/materials management or supply chain departments and in some cases, the position of Chief Resource Officer. This shift is not unique to North America; in France, for example, the first such platforms made their debut in the late 1990s (Landry et al. 2000).

## 3 Distribution of Supplies to Nursing Units

As mentioned above, the distribution of medical supplies to nursing units represents a key component of the hospital's internal supply chain, given the many nursing units found in a hospital, the large quantity of items replenished on a daily basis, the number of clinical staff impacted, and the cost of these items. Methods of distributing supplies to nursing units can be classified according to whether decisions regarding the quantities to be replenished are centralized in the materials management/logistics department or decentralized in the nursing units themselves, and, on another level, according to whether supplies are managed in nursing units as perpetual inventory (online real-time inventory status) or as periodic inventory (where items on hand have to be counted to establish reordering quantities).

In most cases, supply purchases are charged to the nursing unit (or user department) at the time of delivery and no longer appear on the hospital's books as an asset. This "unofficial inventory" can represent up to ten times the value of the official inventory, that is, the supplies kept in central stores and managed through a perpetual inventory management system (Berling and Geppi 1989).

Based on the literature on this subject (primarily Perrin 1994) and our own field experience, we note that the methods most commonly used to distribute supplies to hospital nursing units have ranged from clinically driven requisition-based systems (decentralized-periodic inventory), exchange carts (centralized-periodic), and periodic automatic replenishment or par level system (centralized-periodic) to the more recently introduced two-bin system (centralized-periodic), RFID-enabled two-bin system (centralized-perpetual), weight control bins (centralized-perpetual), and user-driven unitary demand capture systems (centralized-perpetual). See Table 18.1 for a detailed description of these different types of inventory management systems.

The order in which the above replenishment systems are presented is not random; it follows the same sequence as their introduction into the healthcare sector. This evolution has come in three waves. In the 1970s, the exchange cart system began to overtake the requisition system in popularity (Perrin 1994). Although rarely used today, the exchange cart concept nevertheless introduced a key objective that subsequent replenishment systems only served to reinforce: transfer responsibility for replenishment from clinical personnel to a centralized administrative body [e.g., hospital central stores or, in the case of stockless materials management, distributors (Rivard-Royer et al. 2002)], which would perform this task for all of the hospital's nursing units. To a certain degree, this is a sort of "internal" vendor managed inventory (VMI) system. This division of duties allowed for greater specialization of functions, with clinical personnel now able to focus on their core mission of patient care. It also allowed staff in the centralized administrative unit to spend time establishing minimum and maximum thresholds for the various supplies kept in stock and identifying the optimal replenishment frequency—in short, managing the inventory, a task that clinical personnel were often forced to neglect (Landry and Philippe 2004).

In the 1980s, the par level system proved itself more efficient than the exchange cart system by delivering appreciable gains through reductions in stock and storage space in central stores, as it eliminated the need to manage duplicate mobile supply carts. Par level was also more flexible, in that it could be used with any and all storage equipment in the nursing unit (fixed carts, fixed shelving, cabinets, etc.), while enabling staff to manage a wider range of products than with the exchange cart system, as the process was no longer limited to using a particular type of mobile cart. Moreover, it permitted the use of portable readers to enter quantities in stock or scan label barcodes. However, these gains were achieved at the expense of additional contact between the material handler (a stores clerk who is part of the materials management department) and clinical staff (mainly nurses or nursing aides), as the handler now had to spend more time in the nursing unit to count the supplies requiring replenishment and put away delivered products. It was nevertheless possible to diminish the disruption by conducting rounds to scan barcodes and put away supplies during the evening or night shifts, when there was less activity on the unit.

The third development came at the end of the 1980s, when the two-bin or kanban system emerged from Denmark and Holland (early 1990s in France; late 1990s in North America) and delivered significant gains over its predecessors. During scanning rounds (order taking), rather than material handlers drawing on their experience only and "eyeballing" the materials as with par level (Leone and Rahn 2010), they could now simply scan labels that had been removed from empty bins and affixed to a wall-mounted board within each nursing unit storage area. It is important to note that the gains generated by the two-bin system did not come at the expense of increased inventory. In fact, compared to the par level system, the two-bin method did not double the quota of supplies, but rather divided it between each

**Table 18.1** Description of distribution methods (adapted and expanded from Landry and Beaulieu 2010)

| Method | Description |
|---|---|
| Requisition | Nursing or clinical support staff conduct regular inventory counts combined with consumption estimates (a form or fixed-interval periodic review system). Products identified as low in inventory are noted on a requisition form that is forwarded, either manually or electronically, to the materials management department. Based on this requisition, required supplies are picked or ordered from external vendors and sent to the nursing unit in question. With this mode, it is often clinical personnel who are assigned the task of putting away the delivered products in the storage units |
| Exchange carts | Medical supplies are placed on a cart positioned in a storage area on the nursing units. Products are taken from the cart and consumed, with the cart being exchanged according to a predetermined schedule by an identical, fully stocked replacement cart (fixed-interval periodic review system). During the replenishment period, the first cart is returned to central stores to be restocked. According to the set schedule, the newly replenished cart will later be exchanged for the cart on the nursing unit |
| Par level | Rounds of the nursing units to be replenished are conducted according to a predetermined schedule. During the rounds, a material handler identifies items that need replenishment on the nursing unit through a visual evaluation or a more formal inventory count. Normally, a product is identified by scanning a barcoded label on the shelf, bin, or packaging, and the quantities counted are entered into a handheld computer. The information is then downloaded to the materials management information system, which compares the quantities counted with established quotas and generates a pick list or requisition in the case of nonstock items (fixed-interval periodic review system). The picked or ordered products are then delivered to the nursing units and put away by a material handler. Some hospitals use a min/max variation of the par level system |
| Two-bin/kanban | Each quota of medical supplies is divided between two compartments. When the first of the two compartments is empty, clinical staff remove the label identifying the product from the front of the compartment and affix it to a wall-mounted kanban board (with rails). Rounds of the nursing units to be replenished are conducted according to a predetermined schedule. Thus, two conditions must be satisfied to trigger the replenishment process: the bin must be empty, and the replenishment process must be underway (hybrid inventory management system, which combines both fixed-order and fixed-interval characteristics). During the rounds, a material handler scans the labels on the board. The replenishment information is then transferred to the materials management information system, which generates either a pick list for items stored in the central warehouse or a requisition for items sourced externally (direct purchase). The medical supplies are delivered to the nursing unit and put away in the empty compartments by a material handler after having rotated the stock<br>Some hospitals have implemented this system using individual plastic bins. In this case, each supply is divided between two labeled plastic bins that are either stacked or placed end to end on a shelf. When a bin is empty, it is set aside and collected by a material handler to be brought back to central stores for replenishment (Graban 2009) |

**Table 18.1** (continued)

| Method | Description |
|---|---|
| User-driven unitary demand capture systems | With these automated systems, items are stored in the nursing units in closed cabinets or open bins. Each unit removed is recorded by the employee (through various means such as pushing buttons or scanning transponders or barcodes), thus capturing consumption. At any point in time (generally at fixed intervals), communication is established with the materials management information system to enable replenishment based on this on-hand quantity. The collected data is also transferred to the hospital's billing application to charge patients for the supplies used to treat them |
| Weight control bins | This system stores the various items in bins and maintains a perpetual inventory in the nursing unit based on the weight of these items. Replenishment is triggered when the bin reaches the preset weight for each product type. Communication is established with the materials management system to enable a request to be generated |
| RFID-enabled two-bin/kanban systems | In this version of the two-bin system, the bin's label looks like any other but is equipped with a passive (no battery) high-frequency (HF) RFID transponder. A reader is installed behind each of the replenishment boards where labels from empty bins are affixed. This board is connected to the hospital's information technology network. The moment an RFID label enters the reading range of the antenna, communication is established with the materials management system to enable a request to be generated at fixed intervals or according to preestablished replenishment rules. In addition to being used with compartments, RFID transponders can also be affixed to the individual plastic bins, thus representing another way to automate the two-bin system. RFID technology eliminates the necessity of conducting rounds to scan the labels of empty bins. The system also maintains a perpetual inventory of bins (i.e., plastic bins or compartments) |

bin. The two-bin system also brought with it greater control over the quantities to order (a fixed quantity per bin). In addition, the two-bin system forced stock rotation and in doing so reduced the risk of products expiring. It could also be combined with a high-density storage system, which enabled a greater variety of products to be stored in the same storage area, including direct purchases that could be managed with this system (Landry et al. 2004).

In the early 1990s, the USA saw the introduction of automated storage cabinets in nursing units, the first user-driven unitary demand capture systems to provide perpetual inventory management. Used primarily for medical supplies and pharmaceutical products, these systems emerged in the American healthcare sector as a result of private hospitals seeking to better reconcile the supplies consumed by patients with those invoiced to them. A few years later, less expensive point-of-use technologies using open bins with transponders were introduced. However, the challenge with these systems, both closed and open, has been compliance by users to record consumption, with one notable consequence being inventory inaccuracies. In the case of the automated cabinets, once the door of a closed cabinet is opened, clinical staff can remove items without accounting for them. Indeed, in a study that targeted the dispensing of pharmaceutical products, Klibanov and Eckel (2003) found that 19.5% of 2,895 drawers contained incorrect inventory. Moreover, in certain countries, such as the USA, the practice of invoicing patients or insurance companies for individual items used is gradually changing to a diagnostic and treatment-related system. For this reason, some are questioning the ongoing practicality of using such sophisticated permanent inventory systems in nursing units.

In the early 2000s, in an effort to reduce compliance issues, some vendors introduced RFID transponder technology, thus eliminating the requirement for nursing unit staff to record transactions (Bendavid et al. 2010). Still, the deployment of this technology has been limited, as each item must be conditioned by affixing an RFID transponder to it, an activity that can become cost-prohibitive given the relative low cost of most medical supplies (Bendavid et al. 2010). The use of this technology has therefore been restricted to a small group of products, such as implants, in specialized areas (operating rooms, cath labs, etc.).

In the mid-2000s, in a further effort to reduce compliance and demand capture issues, a weight control bin solution for general supplies was adapted from the industrial sector and introduced in the US healthcare sector. The solution automatically triggers the replenishment process using order point logic. However, the solution is challenging from a space utilization point of view; not only is it a wall-mounted system, but in many cases the walls used must be reinforced. The offering has a limited assortment of bins and can also be unreliable in a live environment (technology failure, recalibration, items returned to the wrong bin, monitoring of expiry dates, etc.).

The mid-2000s also saw the development of the RFID-enabled two-bin replenishment system. Initially developed in Canada (Beaulieu and Landry 2010; Bendavid et al. 2010), this replenishment system has since become popular in Europe, particularly in France and Spain. In 2011, a computer vision-enabled version of the two-bin replenishment system (video capture) was introduced in North America at the annual AHRMM conference (Association for Healthcare Resource & Materials Management). Using video cameras, the new application is capable of recognizing empty bins and generating replenishment orders. This further reduces human intervention while maintaining the benefits of the two-bin system. Building on the gains delivered by the two-bin system, both RFID and video capture technology allow for the elimination of data collection through preestablished rounds and provide real-time, remote visibility of inventory levels and replenishment needs. At the bin level, the periodic review model has thus evolved into a perpetual inventory model.

All of these inventory management systems use a fixed-interval reordering process (periodic review system), order point logic, or a combination of both (hybrid system). This means that the review period duration, maximum inventory level, order point, reordering quantities, and safety stock can all be calculated using various stochastic inventory models to try to find the right balance between ordering costs and inventory carrying costs. Research has shown that keeping these inventory management parameters up to date can lead to improved performance (Landry et al. 2004). Unfortunately, in many hospitals, the rule of thumb prevails; too often demand is not tracked, and parameters are not kept up to date.

Over and above what has been presented for a typical nursing unit, the operating room presents unique challenges, as a large proportion of the items it carries are nonstock (direct purchase) and consignment items. Moreover, these supplies are often very expensive, with inventory costing five to six times more than that stocked in the hospital's central stores and with an inventory turnover rate of 2.5, compared to 12 in central stores (Park and Dickerson 2009). However, the OR offers a rare opportunity in healthcare, as material usage could theoretically be planned days or sometimes weeks in advance by taking advantage of the forward visibility of the OR schedule and surgeons' preference lists (bills of materials). Material requirements planning (MRP) systems, common in the manufacturing sector for dependent demand items, could then be used (Steinberg et al. 1982; Lafond and Landry 2001). To our knowledge, however, very few examples of such utilization exist in practice in the OR. Currently, forward visibility is restricted to using OR schedules and preference lists (for predictable items) to enable the preparation of case carts before a surgical procedure. Rather than having OR staff pick supplies and instruments from OR storerooms right before the operation, the ability to prepare in advance also paves the way for automating of charge capture, with data collection greatly streamlined for patient charging where applicable.

And finally, the OR has recently seen the introduction of RFID technology to manage high value items via a number of applications, such as RFID shelves, cabinets, and receptacles (Bendavid and Boeck 2011). Although they use the same technology to collect data, RFID-enabled shelves and cabinets manage the process differently from RFID receptacles: shelves and cabinets read tags on product packages when within the field of the antenna and therefore in inventory, and removing a product from a shelf will deplete this product from inventory. In the case of the receptacle, data is read when a product is consumed, with its RFID-tagged packaging disposed of in an RFID-enabled receptacle. The logic here is that a product recorded in the conditioning process is in inventory until its packaging is disposed of in the receptacle, thus requiring less technology per item managed. As for the RFID-enabled shelves and cabinets, these are used to track the usage of unpredictable high value items—usually consignment products—such as orthopedic prostheses (Philippe and Beaulieu 2010).

## 4 Best Practices in Medical Supply Distribution Methods

In many situations, the two-bin/kanban replenishment method has proven to be a better inventory management system for medical supplies and common drugs, office supplies, etc. than clinically driven requisition-based methods, exchange carts, par level, or more expensive automated cabinets (Landry et al. 2004; Black and Miller 2008; Graban 2009; Landry and Beaulieu 2010; Leone and Rahn 2010).

As mentioned above, generally speaking, the two-bin system offers the following advantages over other periodic review systems:

- No-count replenishment system (built-in decision rule; no “eyeballing,” as is often the case with other systems)
- Reduces the average inventory level, because it increases the quality of information at the point of use
- Reduces the time taken for the ordering process (four to seven times faster than par level systems, Landry et al. 2004) and thus reduces the time spent by material handlers in the nursing unit (less chance of disruption to clinical activities)
- Reduces product handling and increases event-related sterility (infection control)
- Reduces the risk of products expiring (built-in stock rotation)
- Manages products with different replenishment cycles in the same storage units
- Leads to better ergonomics when implemented with high-density storage systems
- Integrates a number of lean healthcare features (visual management, standardized process, kanban; Landry and Beaulieu 2010)

The addition of RFID technology has further improved the two-bin system. It has eliminated the need to conduct rounds of the nursing unit to scan the labels of empty bins, thus doing away with movements with little or no added value (elimination of waste) and disruptions on the nursing unit, particularly in hard to access areas (Landry and Beaulieu 2010). Moreover, RFID technology, combined with a materials management information system, can immediately alert the materials management department via pager or other device that there is a stockout in the nursing unit (i.e., that the label from the second bin has been affixed to the board; Landry and Beaulieu 2010). The emerging application of voice technology used in conjunction with portable RFID readers also improves the put away process by locating labels on the board faster and reducing the risk of errors.

While we consider the RFID-enabled two-bin system a better way to manage supplies in nursing units or specialty areas such as the OR, one must remain open minded about using other systems or techniques in specific circumstances. For example, we have seen the exchange cart work well in dialysis, with sourcing directly from the vendor. Indeed, according to Szulanski (1996), best practices are replicated organizational routines where "practice refers to the organization's routine use of knowledge and often has a tacit component, embedded partly in individual skills and partly in collaborative social arrangements." Winter (1995) states that these organizational routines intuitively rely on behaviors that generate a predictable result. On a more conceptual level, these routines "can be conceived as a web of coordinating relationships connecting specific resources" (Winter 1995). A practice is therefore not limited to technology and work processes. Under these circumstances, a practice can qualify as "best" based on the environment where it is deployed (Moore 1999).

For example, Hôpital du Sacré-Cœur de Montréal, a hospital in Montreal, Canada, transformed the secondary storage locations in its emergency department examination rooms into more than 30 primary storage points replenished by the hospital's central stores. These storage locations had previously been replenished by clinical support staff from a central storage area in the emergency department itself. The change gave material managers a greater awareness of the replenishment process of a very busy area of the hospital (Beaulieu and Landry 2010). It also generated substantial clinical productivity gains that could then be refocused on patient care.

This transformation was supported by the implementation of an RFID-enabled two-bin application housed in a high-density storage system, a solution perfectly suited to the challenges at hand. But, in keeping with the above account of the impact of combining resources, behaviors, and a practice, it was also supported by a revamping of the working methods of materials management employees and the training of clinical personnel. In this example, not only was a technology deployed that offered intrinsic advantages, but the hospital implemented it within a perspective of optimizing the entire logistics process in terms of who does what at each step of the process. The hospital also moved away from replenishing nursing units during the busy day shift, when there is a greater likelihood of disrupting clinical flows and when access to elevators is more challenging (Beaulieu and Landry 2010). With respect to putting away supplies, Sacré-Cœur opted to have two designated teams deliver supplies to nursing units and put the supplies away in each bin while performing stock rotation. This further illustrates what we cited about best practices earlier in the section and the fact that better practices become "best" in real-life situations and in a specific context that integrates both hard and soft dimensions.

We can therefore define best practices as a set of organizational characteristics that produce superior performance. A best practice may take the form of a technology, work method, work organization, or a combination of all of these elements (Landry et al. 2000). During its 10-year initiative, Sacré-Cœur has demonstrated the importance of patience, continuous improvement, and innovation while maintaining a clear vision of what it means to have an integrated internal supply chain. Today, the logistics department is considered as playing a strategic role within the hospital

(Amaya et al. 2010). Its involvement goes beyond the simple transferring of supplies in response to the needs of clinical staff; it has become a major stakeholder in developing solutions to the various logistics problems faced by the hospital, for example, patient movements and the positioning of medical equipment. This department has demonstrated its ability to coordinate resources in order to produce new organizational routines. In short, when it comes to best practices, perhaps W. Evert Welch said it best: "There are no bad techniques [. . . ] just bad applications" (cited in Plossl 1994, p. 288).

## 5 Future Research Opportunities

Due to the real-time visibility of inventory status associated with the two-bin replenishment concept, RFID technology and more recently video capture have made it possible to proactively manage supplies by triggering replenishment rounds based on a range of criteria (number of labels on the kanban board, time elapsed since a label has been on the board, stockout situation, etc.) (Landry and Beaulieu 2010). These innovations open the door to a large number of research avenues. For example, when deciding on replenishment triggers, what criteria might help optimize the management of inventory? Also, in terms of transport, this makes possible a field of research on deterministic or stochastic inventory-routing problems in the world of medical supply distribution. Currently, the replenishment of nursing units is for the most part done according to a predetermined schedule that is rarely updated (e.g., a hospital may decide to replenish a nursing unit every day, on Mondays and Thursdays, or once a week). This schedule can remain unchanged over a number of months or even a year and beyond. Increased knowledge of the stock status in the nursing units can enable staff to plan replenishments according to a variable schedule based on one or more criteria, while optimizing the transport route in consideration of the supplies being delivered, the location of the deliveries (nursing units), the hospital configuration (corridors and elevators), and the capacity of the transport cart (capacitated vehicle routing problem).

In addition, when combined with the implementation of a warehouse management system, the RFID-enabled two-bin system offers the possibility of tracing each medical supply lot from the moment it enters the hospital, through the receiving area, until the moment it is used in a nursing unit, by linking each lot to specific storage bins. This opens yet another research avenue. Simulation may constitute a way of approaching these questions and addressing issues that include, for example, whether secondary storage locations in the nursing units should be replenished by clinical staff from primary locations or whether central stores should replenish these directly. In other words, what should the key drivers be in defining a primary vs. secondary storage location? Simulation also offers the possibility of weighing different options during a major renovation project or the construction of new nursing units or even a new hospital.

The emergence of centralized distribution platforms also generates interesting research possibilities. For example, does a better model exist between shared services and third party logistics (3PL) providers, or under what conditions should one or the other organizational mode or governance structure be selected? Does a third option exist? Given the emergence of distribution platforms, which often are region-wide, what impact might these platforms have on upstream partners in the supply chain, primarily GPOs and distributors?

Change management also offers interesting research opportunities, as the implementation of the innovative systems discussed in this chapter (e.g., two-bin system, RFID-enabled two-bin system, or the preparation of case carts using surgeons' preference lists) introduces varying levels of transformation to the organization.

## 6 Conclusion

We have demonstrated the important role of supply chain management in a hospital setting and the various ways of structuring the associated activities, and we have emphasized the main challenges of distributing medical supplies to nursing units. Operational excellence is thus achieved through the use of "best" inventory management and distribution systems, combined with continuous supply chain process improvements and better integration with the patient care process. With respect to the latter, dramatic clinical advancements have been made in recent years. However, generally speaking, the management processes supporting the delivery of care have not moved at the same pace (Spear 2009). Emerging healthcare supply chain innovations, many of which, such as RFID technology and the lean approach, originated in the industrial sector, offer an opportunity to fill the gap. The implementation of best practices in the distribution of medical supplies to nursing units is a good illustration of how improving the healthcare supply chain releases clinical staff from the frustrations associated with managing inventory at the nursing unit level and provides them with the time to focus on problems that more directly involve them and that they have been trained to resolve.

## References

AHA (2012) Fast facts on US hospitals. www.aha.org Accessed January 3, 2012.

AHRMM (2000) National performance indicators for healthcare materials management. Association for Healthcare Resource & Materials Management, Chicago

AHRMM (2010) Healthcare supply chain, resource & logistics processes. Association for Healthcare Resource & Materials Management, Chicago

Amaya CA, Beaulieu M, Landry S, Rebolledo C, Velasco N (2010) Potenciando la contribucion de la logistica hospitalaria: tres casos, tres trayectorias. Int Manag 14(4):85–98

Andersen A (1990) Stockless materials management: how it fits into the healthcare cost puzzle. HIDA Educational Foundation, Alexandria

Aston G (2010) Teaming with physicians can drive down costs. Hosp Health Netw 84(1):13

Bales WA, Fearon HE (1993) CEOs'/Presidents' perceptions and expectations of the purchasing function. CAPS, Tempe

Beaulieu M, Landry S (2010) Le déploiement d'une stratégie logistique à l'Hôpital du sacré-coeur de Montréal. Int J Case Stud Manag 8(1):35

Bendavid Y, Boeck H (2011) Using RFID to improve hospital supply chain management for high value and consignment items. Procedia Comput Sci 5:849–856

Bendavid Y, Boeck H, Philippe R (2010) Redesigning the replenishment process of medical supplies in hospitals with RFID. Bus Process Manag J 16(6):991–1013

Berling RJ Jr, Geppi JT (1989) Hospitals can cut materials costs by managing supply pipeline. Health Care Financ Manag 43(4):19–22

Black J, Miller D (2008) Toyota way healthcare excellence. ACHE Management Series, Chicago

Blackburn JD (1991) The quick response movement in the apparel industry, in Blackburn JD time-based competition—the next battle ground in American manufacturing. Business One Irwin, Homewood, pp 246–269

Bourgeon B, Constantin A, Karolszyk G, Marquot JF, Pedrini S, Landry S, Diaz A, Estampe D (2001) Évaluation des coûts logistiques hospitaliers en France et aux pays Bas. Logistique Manag 9(1):81–87

Burnette SW (1994) Efficient materiel handling and distribution: a design perspective. Hosp Mater Manag Q 16(2):24–34

California Healthcare Foundation (2012) Health care costs 101. Slow but Steady, CHCF, Oakland.

Cammish R, Keough M (1991) A strategic role for purchasing. McKinsey Q (1):22–39

Chow G, Heaver TD (1994) Logistics in the Canadian health care industry. Canadian Logist J 1(1):29–73

Consulting CSC (1996) EHCR, Efficient healthcare consumer response, improving the efficiency of the healthcare supply chain. American Hospital Association/American Society for Healthcare Materials Management, Chicago

DeJohn P (2005) The last frontier: saving on M.D. preference items. Hosp Mater Manag 30(6):1, 9–11

Driscoll RS (1981) Supply, processing, and distribution: an overview. Hosp Mater Manag Q 2(4):21–24

Fearon HE, Ayres DL (1967) Effect of centralized purchasing on hospital costs. J Purchasing 3(3):22–35

Fereng J (2010) How are your nurses spending their time? Hosp Health Netw 84(5):14

Graban M (2009) Lean hospitals. CRC, New York

HFMA (2010) HFMA's healthcare financial pulse. HFMA, Westchester

Hunter NA, Valentino P (1995) Quick response—ten years later. Int J Clothing Sci Technol 7(4):30–40

Janson RL (1985) Future trends in hospital materiel management. Hosp Mater Manag Q 7(1): 11–17

Klibanov OM, Eckel SF (2003) Effects of automated dispensing on inventory control, billing, workload, and potential for medication errors. Am J Health Syst Pharm 60:569–572

Kowalski JC (1993) Managing hospital materials management. Kowalski Dickow Associates, Washington

Kurt Salmon Associates Inc. (1993) Efficient consumer response: enhancing consumer value in grocery industry. Food Marketing Institute, Washington

Lafond N, Landry S (2001) Gérer plus efficacement les stocks du bloc opératoire à partir de la programmation des interventions chirurgicales. Gestions Hosp (405):259–263

Landry S, Beaulieu M (2002) Logistique hospitalière: un remède aux maux du secteur de la santé? Gestion 26(4):34–41

Landry S, Beaulieu M (2007) The hospital: not just another link in the healthcare supply chain. In: Starr MK (ed) Foundations of production and operations management. Thomson, New York, p 281

Landry S, Beaulieu M (2008) Recognizing the special character of the hospital link in the healthcare supply chain. In: Starr MK (ed) Production and operations management, 2nd edn. CENGAGE Learning, Stamford, pp 402–404

Landry S, Beaulieu M (2010) Achieving lean healthcare by combining the two-bin kanban replenishment system with RFID technology. Int J Health Care Manag 1(1):85–98

Landry S, Philippe R (2004) How logistics can service health care. Supply Chain Forum 5(2): 24–30

Landry S, Beaulieu M et al (2000) Étude internationale des meilleures pratiques de logistique hospitalière, vol 00–05. CHAÎNE Research Group, HEC Montréal, Montreal

Landry S, Blouin JP, Beaulieu M (2004) Réapprovisionnement des unités de soins: portrait de six hôpitaux québécois et français. Logistique Manag (Special issue):13–20

Leone G, Rahn RD (2010) Lean in the OR. Flow Publishing, Boulder

Montgomery K, Schneller ES (2007) Hospitals' strategies for orchestrating selection of physician preference items. Milbank Q 85(2):307–335

Moore R (1999) Making common sense, common practice. Cashman Dudley, Houston

Nachtmann H, Pohl EA (2009) The state of healthcare logistics: cost and quality improvement opportunities. Center for Innovation in Health care Logistics, Arkansas

Park KW, Dickerson O (2009) Can efficient supply management in the operating room save millions? Curr Opin Anesthesiol 22(2):242–248

Parker J, DeLay D (2005) The future of the health care supply chain. Health Care Financ Manag 62(4):66–69

Pedersen J (1996) Product standardization: playing to win. In Vivo 14(6):15–20

Perrin RA (1994) Exchange cart and par level supply distribution systems: form follows function. Hosp Mater Manag Q 15(3):63–76

Philippe R, Beaulieu M (2010) Supply chain processes in ORs can be improved using industrial practices. Can Health Care Technol 15(4):13–15

Plossl GW (1994) Orlicky's material requirement planning. McGraw-Hill, New York

Rivard-Royer H, Landry S, Beaulieu M (2002) Hybrid stockless—a case study: lessons for health care supply chain integration. Int J Oper Prod Manag 22(4):412–424

Schneller ES, Smeltzer LR (2006) Strategic management of the health care supply chain. Jossey-Bass, San Francisco

Souhrada L (1998) Sky's the limit. Mater Manag Health Care 7(7):24–26

Spear SJ (2009) Chasing the rabbit. McGraw-Hill Companies, New York

Steinberg E, Khumawala B, Scamell R (1982) Requirements planning systems in the health care environment. J Oper Manag 2(4):251–259

Szulanski G (1996) Exploring internal stickiness: impediments to the transfer of best practice within the firm. Strategic Manag J 17(Winter special issue):27–43

The Governance Committee (1997) Richest sources of savings, lessons from America's lowest-cost hospitals. Advisory Board Company, Washington

Thorsfeldt H (1988) Why today's central service is an integral part of materiel management. Hosp Mater Manag Q 9(3):63–70

Tucker AL, Edmondson AC (2003) Why hospitals don't learn from failures. California Manag Rev 45(2):55–72

Tudor TL, Wollridge AC, Bates MP, Philips PS, Butler S, Jones K (2008) Utilizing a "systems" approach to improve the management of waste from health care facilities: best practice case studies from England and Wales. Waste Manag Res 26(3):233–240

Tudor TL, Townend WK, Cheeseman CR, Edgar JE (2009) An overview of arising and large-scale treatment technologies for health care waste in the United Kingdom. Waste Manag Res 27(4):374–383

Tyagi RK, Vachon S, Landry S, Beaulieu M (2010) Reverse supply chain in hospitals: lessons from three case studies in Montreal. In: Ferguson ME, Souza GC (eds) Closed-loop supply chains—new developments to improve the sustainability of business practices. CRC/Taylor & Francis Group, Boca Raton, pp 181–194

Thitchie L, Burnes B, Whittle P, Hey, R (2000) Benefits of reverse logistics: the case of the Manchester Royal Infirmary Pharmacy, Supply Chain Management 5(5): 226–233.

Winter SG (1995) Four Rs of profitability: rents, resources, routines, and replication, Montgomery CA resource-based and evolutionary theories of the firm. Kluwer Academic, Boston, pp 144–178

# Chapter 19
# Overcoming the Challenges of the Last Mile: A Model of Riders for Health

**Jessica H. McCoy**

## 1 Introduction

More than eight million children under the age of 5 died in 2009, largely of preventable causes; the majority of these deaths (74%) were in developing countries. A child dies every 30 s from malaria alone, and each day, 1,500 women die from causes related to pregnancy and childbirth that could be avoided by care from trained health professionals. These deaths are preventable, but even so, millions of people die every year because they do not have access to basic healthcare (World Health Organization 2011).

People in resource-limited regions face numerous barriers to healthcare access. Countries with strong health resources in their capital cities may not have the channels necessary to distribute those resources beyond the city limits. Resources and funding for healthcare may come from all over the world, but are of little use unless they reach the intended patients. We refer to the final leg of the journey of healthcare from a point of access (such as a local clinic or traveling health worker) to patients as the "last mile" (Fig. 19.1). Whether healthcare is accessible to a community often comes down to this critical last mile of health delivery. Upon her election as director-general of the World Health Organization in 2006, Margaret Chan noted that "all the donated drugs in the world won't do any good without an infrastructure for their delivery" (United Nations 2006).

Last-mile challenges are diverse and widespread. It is difficult to build and support remote clinics, particularly in the absence of a road network. In addition, some regions may be unreachable during certain seasons (e.g., rainy season). Transient populations are difficult to reach consistently. For example, rates of childhood vaccination and assisted births in nomadic populations in Chad are

J.H. McCoy (✉)
Stanford University, Stanford, CA 94305, USA
e-mail: jhmccoy@stanfordalumni.org

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_19, © Springer Science+Business Media New York 2013

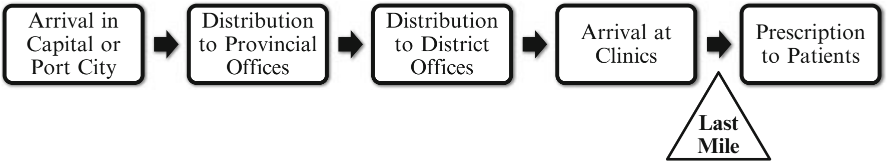

Arrival in Capital or Port City
Distribution to Provincial Offices
Distribution to District Offices
Arrival at Clinics
Prescription to Patients
Last Mile

**Fig. 19.1** Hypothetical supply chain for medical supplies such as drugs. The "last mile" is the final step of providing healthcare to the end consumer of the supply chain: the patients

unacceptably low (Hampshire 2002). To diagnose and treat patients properly, clinics must have appropriate equipment, capacity, and resources such as nurses and drugs. Health worker shortages are widespread: trained health workers are in short supply in many parts of Africa, so some rural health workers in Africa are each responsible for thousands of people (Chen et al. 2004; World Health Organization 2011).

Accessibility is affected by a number of financial, social, and physical factors. The high cost of treatment, in addition to the cost of transportation and missed work, is prohibitive for many (Weiser et al. 2003). Social stigma attached to diseases such as HIV may deter patients from seeking treatment (Mills et al. 2006; Weiser et al. 2003). Some patients are unable to seek routine care due to unstable lifestyles (Chesney 2000; Hampshire 2002). Additionally, unsafe environments caused by armed conflict can disrupt healthcare access (Reilley et al. 2002).

Transportation barriers to healthcare are significant in developing regions and have far-reaching implications regarding the health of rural populations (Porter 2002). Physical access challenges such as the long distance to a treatment center or the need to rely on public transportation to travel to that treatment center are common reasons for nonadherence to tuberculosis treatment (Shargie and Lindjørn 2007). Poor roads directly affect a population's ability to access emergency obstetric care, and people in developing countries who live more than 4 or 5 miles from a clinic receive less frequent healthcare (Howe and Richards 1984; Samai and Sengeh 1997). Improving transport could lead to better health outcomes (Ahmed and Hossain 1990; Operations Evaluation Department 1996). Despite these findings, transportation options tend to be severely limited in most developing countries. In the poorest 15 countries in sub-Saharan Africa, only 37% of people living in rural areas are within 2 km of an all-season road (Transport Sector Board 2008).

Some researchers propose to bypass physical barriers altogether by implementing alternatives to traditional health delivery. Telemedicine programs use communications technology to facilitate diagnoses and appropriate care of rural patients. For example, Project Tristan connects patients on the remote island of Tristan da Cunha to specialists at the University of Pittsburgh Medical Center via a dedicated satellite-internet connection; otherwise, healthcare for these patients is a 6- to 7-day boat ride away (Economist 2008). Real-time videoconferencing can connect cancer patients worldwide with oncologists in order to provide advanced care (Hazin and Qaddoumi 2010). Using text messages to remind patients to take their antiretroviral medication improved treatment adherence rates in a trial in Kenya (Lester et al. 2010).

Even though telemedicine programs have not yet been collectively demonstrated to be cost-effective in the developed world (Whitten et al. 2002), many researchers believe that such programs could improve health delivery in the developing world in a cost-effective manner, especially given advancements in wireless technology (Anupindi et al. 2009; Istepanian et al. 2006). Although telemedicine programs have the potential to help alleviate the logistics challenges of the last mile, they are unlikely to eliminate those challenges on their own.

In this chapter, we focus on the *physical* accessibility of healthcare; our discussion applies broadly to the developing world, but many of our examples are in the context of sub-Saharan Africa. Rural populations are often underserved because both patients and their local health workers lack transportation options. Even if roads do exist, individuals in rural communities do not own motor vehicles (Transport Sector Board 2008). In Zambia,[1] we interviewed rural health workers who resort to hiring taxis to transport critically ill patients to the closest hospital—but can do so only if the patients can afford it. Other health workers we interviewed bring fuel to the local market to trade for rides in order to conduct outreach visits to outlying communities. While some clinics may have a motorcycle or car, the vehicle is of little use if it is broken down or out of fuel. Public health budgets are stretched far, and frequently there is insufficient funding for fuel or fleet maintenance.

The remainder of this chapter is organized as follows. In Sect. 2 we provide a brief review of the literature on operations research as applied to resource-limited settings and last-mile problems. We adapt a traditional operations research modeling approach to the context of last-mile health delivery in Sect. 3 and illustrate this framework by presenting a case study and model in Sect. 4. We review model next steps and limitations in Sect. 5 and conclude with a discussion of the challenges of modeling in this context in Sect. 6.

## 2 Literature Review

Operations research is a powerful methodology for process improvement and can be applied to model the last mile of health delivery to help decision makers choose viable solutions. While the application of operations research to health delivery is well studied, modeling in the context of resource-limited regions is less common. In this section, we review work that has been done in developing regions and, more specifically, with regard to the critical last mile.

The relatively new field of humanitarian logistics contains numerous examples of operations research applied in the developing world (for more extensive reviews, see (Apte 2009) and (McCoy 2008)). For example, Beamon and Kotleba proposed

[1] Over the course of several trips to Zambia in 2010–2012, our team interviewed dozens of health workers and health officials. We mention some of their anecdotes here but suppress names to preserve anonymity.

the use of $(Q,R)$ inventory policies for the ongoing relief supply chain in southern Sudan (Beamon and Kotleba 2006). McCoy and Brandeau (2011) used dynamic programming to develop optimal shipping policies from a stockpile to a relief operation, informed by the authors' work with the United Nations High Commissioner for Refugees in sub-Saharan Africa and the Middle East. De Angelis et al. (2007) worked with the World Food Programme to build a vehicle-routing schedule for cargo planes in Angola.

Researchers have also studied the supply chains of sub-Saharan Africa to understand the challenges of operating in resource-limited regions. Yadav et al. (2011) developed a case on the differences between medical and beverage supply chains in sub-Saharan Africa. Spiliotopoulou and Yadav used a dynamic compartmental model governed by differential equations to model the pharmaceutical supply chain in Africa and understand the impact of drug assortment on disease resistance to malaria medication (Spiliotopoulou et al. 2011). Additionally, researchers have linked operations within clinics to successful treatment programs. Using an optimization model of clinic capacity management that incorporated treatment adherence, McCoy and Johnson (2012) proposed more effective treatment programs for rural clinics treating infectious diseases such as HIV. Deo et al. (2010) used a time and motion study to inform a simulation of clinic wait times.

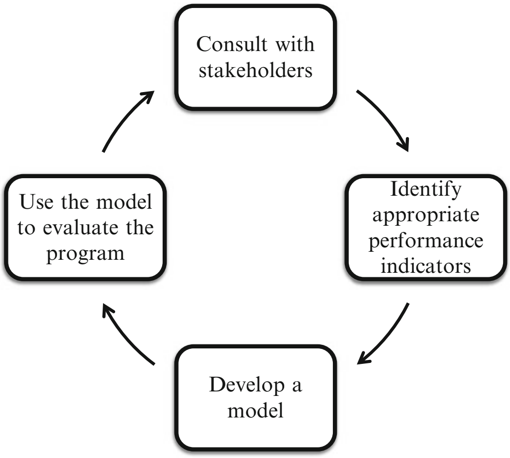
These models also shed light on the challenges of and possible solutions for the last mile of health delivery. In addition, Balcik, Beamon, and Smilowitz developed a mixed integer program to allocate resources and route vehicles between local distribution centers and final beneficiaries (Balcik and Beamon 2008). Other researchers have focused on fleet management as a method for overcoming the last mile. Pedraza Martinez and van Wassenhove (2012) modeled the International Federation of Red Cross and Red Crescent Societies' optimal vehicle replacement policy as an optimal stopping problem, and Pedraza Martinez et al. (2011) used game theory to develop contracts that alleviate incentive misalignments in vehicle replacement.

Operations research models and methods are well suited to identifying solutions to problems with the last mile in resource-limited regions. Operations in developing regions are different enough from those in developed regions that models for developed regions are not directly applicable. Models created for the developed world may have assumptions and objectives that do not apply to developing regions such as sub-Saharan Africa. We contribute to these growing streams of literature by modeling the effectiveness of fleet management in health delivery. We present our approach in the following section.

## 3 Framework for a Model-Based Approach

Not all of the problems that plague health delivery supply chains are common to both the developed and the developing world, and so models that are insightful and useful in the context of the former do not necessarily apply to the latter and vice versa;

**Fig. 19.2** Adapted operations research modeling framework


Consult with stakeholders
Identify appropriate performance indicators
Develop a model
Use the model to evaluate the program

for example, operating room scheduling might not be a priority in rural Botswana, while models on the spread of malaria do not apply in Canada. Each organization in the health delivery supply chain in a region likely operates under different objectives with different incentives, so the context of the problem is very important. In addition, translating model results and policy into sustainable programs is itself a tremendous hurdle (Madon et al. 2007; Sanders and Haines 2006). In recent years, research on implementation science has burgeoned to understand how policies can be implemented more successfully; for examples, see Chokshi and Kesselheim's analysis of why access to vaccines remains inadequate in resource-limited countries (Chokshi and Kesselheim 2008), and Zwarenstein et al.'s work on improving the quality of care to children with asthma in South Africa (Zwarenstein et al. 2007).

With these concerns in mind, we apply a traditional operations research modeling implementation approach (e.g., see Chap. 2 of Hillier and Lieberman (2005)) to our context: last-mile health delivery in resource-limited regions. In this section, we adapt the approach (Fig. 19.2) to frame our work with Riders for Health in Sect. 4.

First, in order to understand the problems of the last mile, researchers must be sure to include stakeholders in the modeling process and identify how stakeholders contribute to and are affected by each challenge in order to ensure relevant results. Health delivery supply chains in developing regions involve a number of stakeholders, not just the decision makers or funders of the supply chain. The end consumers of these supply chains are the local communities, families, and individuals that need care. The health workers who serve these communities are the next members of the supply chain, followed by their employers, who are health officials at the district, provincial, and national levels. Nonprofit organizations comprise a large fraction of the health sector in some resource-limited regions. Finally, global aid agencies and donor nations funnel donations (both monetary and in-kind) into health delivery supply chains.

Second, all of these entities are invested in the success of the health delivery supply chain but may have different approaches for improving operations as well as different metrics to gauge any improvement. Patients may value low waiting times and inexpensive treatment at their local clinic (cf. Deo et al. (2010)), while health workers might want additional coworkers to lessen their workload. Administrators may be concerned with minimizing costs, while donors might propose programs to increase intervention access or coverage (c.f. Bryce et al. (2004) and Rosero-Bixby (2004)). After identifying the challenges faced by different stakeholders, researchers should select a reasonable objective that is supported by many if not all of the stakeholders.

Having identified the problem and clarified an objective function, the next step is to apply the tools and techniques of operations research to build an insightful model. Creating models for the developing world involves some unique challenges. For example, reliable data may be especially scarce in a resource-limited environment. Lack of data presents a barrier to an organization's ability to estimate need (e.g., Samii et al. (2002)). One possible solution is to identify proxy variables for variables that are unavailable, unreliable, or costly to measure. Conversations with stakeholders can identify what data are available and how available data relate to desired but unavailable data; for examples in health modeling, see Acharya and Cleland (2000), Dovlo (2005), and Ranson et al. (2003). Modelers should take care to identify metrics that are measurable and that make sense in the context of the supply chain.

At the final stage of the modeling approach, hopefully any model results will answer the questions provided by stakeholders in earlier stages of the project. If the model relies on uncertain data, sensitivity analyses are critical for validating the results. By identifying which parameters the model is robust or sensitive to, sensitivity analyses can help decision makers determine which variables should be estimated most carefully. Evaluating a program goes hand in hand with communicating the results with stakeholders in a clear, relevant manner. If model results are not reasonable, stakeholders can inform the adjustment of assumptions or parameters to improve the model. Involving stakeholders in each step of the modeling process not only instills confidence in the results but also leaves them in a better position to implement any findings.

In the next section, we illustrate this framework by presenting a case study and model from our experience working with Riders for Health.

## 4 Case Study: Evaluating the Effectiveness of Riders for Health

Riders for Health (referred to henceforth as "Riders") is a nonprofit organization that focuses on improving last-mile health delivery by providing transportation solutions to health workers. In particular, Riders' approach relies on managing motorcycles since motorcycles are cheaper to procure and maintain than automobiles and are also more suited to the unpaved roads and footpaths common in rural areas. Rural clinics whose health workers previously walked or used bicycles to visit outlying villages can increase the number of health interventions they deliver using motorcycles managed by Riders (Lee and Tayan 2007). For example, Riders' involvement in the Gambia has coincided with an increase in infant vaccination coverage from 62 to 73% (Riders for Health 2011). In a district of Zimbabwe where all community health workers are mobile thanks to Riders, deaths from malaria have decreased by 20%.

Riders' rigorous fleet management enables it to maintain fleets with few or no breakdowns and to extend the life of each motorcycle in the fleet (Rammohan 2010). Riders achieves these results through its core competencies: driver training programs, a hub-and-spoke service system, and efficient spare parts inventory management. When operating in a region, Riders trains everyone who will be using a Riders motorcycle in safe driver practices and basic maintenance. In addition, Riders establishes a service network and hires and trains mechanics. For more details on Riders' inception and its fleet management program, see Lee and Tayan (2007) and Lee et al. (2011).

After enjoying success on a small scale, Riders is now interested in securing national-level contracts. Before pursuing such contracts, Riders must evaluate its program effectiveness. Recently, our team at Stanford University has worked to design a 2-year trial (begun in 2011) to measure Riders' effectiveness at improving health worker mobility. The insights in this chapter are based on our experiences developing and initiating this trial. First, we explain the strategic and tactical decisions that have driven the trial design.

Though it is a nonprofit organization, Riders charges its customers for the fleet management it provides. Together with potential clients (typically ministries of health), Riders develops a contract of fleet management to suit the needs of the client. The contract specifies both the level of service that Riders will provide (e.g., management of existing fleet or provision and management of a new fleet) and the cost that Riders will charge the client per kilometer driven by the fleet.

From conversations with national-level health officials in Zambia, Riders chose to base the trial in the Southern Province of Zambia. The Southern Province is home to 1.6 million people, spread out over 85,283 km$^2$ (Encyclopædia Britannica 2010b; Zambia Central Statistical Office 2010).[2] For the trial, we have randomly selected four experimental and four control districts from the 11 districts of the Southern Province. Each district has between 10 and 40 health centers, each of which is responsible for supporting several health posts and/or outreach sites (Zambia Ministry of Health 2010). Each district has 5–30 motorcycles to meet the health needs of their populations; prior to the start of the trial, more than

[2] As a comparison, note that Austria is smaller at 83,879 km$^2$ yet has more than five times the population with 8.4 million people (Encyclopædia Britannica 2010a).

half of these motorcycles were broken down. The quality of garages varies widely from well-stocked dealerships in the main cities of Lusaka and Livingstone to small operations run out of entrepreneurial mechanics' homes in more rural areas. The majority of the road network in Zambia is unpaved; only 18% of Zambia's 37,000 km of gazetted roads are paved, and only 57% of those roads are in good condition (Organisation for Economic Co-operation and Development 2006; National Road Fund Agency 2005). Compared to neighboring countries, Zambia's rural road accessibility is poor: only 17% of rural Zambians live within 2 km of an all-season road, compared to the average of 37% across sub-Saharan Africa (World Bank 2010). As of 1998, half of rural Zambians lived more than 5 km from the nearest health center, and only a third attempted to consult with a medical professional when ill (World Bank 2006).

The parameters of the trial, including the number of existing motorcycles that Riders will manage as well as the number of new motorcycles that Riders will procure, were established through negotiation between donors, Riders, and Zambian officials. After selecting a location and determining the size of the program, Riders extensively interviewed health workers throughout the experimental districts of the Southern Province to identify specifically where to station the motorcycles (i.e., the actual health facilities and health workers). In parallel, Riders worked to establish a service system to support the fleet by setting up a garage and hiring mechanics. Additionally, Riders worked to train each person that would be using one of the fleet motorcycles. During the trial, Riders will manage the fleet of the experimental districts according to the parameters outlined in the contract with the client organization.

As evaluation partners, our team at Stanford is tasked with demonstrating the link between Riders' program and changes in health outcomes in the experimental districts. We propose to do so using three levels of evaluation. First, we plan to quantify the effectiveness of Riders' management in reducing fleet downtime. Second, any changes in fleet availability should be connected to changes in health intervention delivery; for example, perhaps more vaccinations are administered, or more bed nets are distributed by a health worker with a reliable motorcycle than by a health worker traveling on foot. And finally, the accumulation of the increased number and improved quality of health interventions delivered over time may contribute to measurable improvements in health outcomes (e.g., higher vaccination coverage, lower malaria incidence). A rigorous evaluation of such hypotheses will help Riders structure future programs and will inform potential donors as to Riders' efficacy. In this chapter, we focus on the first evaluation step. A robust model of Riders' fleet uptime (equivalently, downtime) will provide a solid foundation for the remainder of our evaluation of Riders' effectiveness. During the trial, we will collect data using surveys and GPS technology to estimate the parameters of this model.

We now illustrate the modeling considerations outlined in Sect. 3 by presenting an analytical model that we developed to help Riders understand the impact of its program in Zambia on health worker mobility. In the following sections, we present our approach (following the steps of Fig. 19.2) and preliminary results.

## 4.1 Step 1: Consult with Stakeholders

During ten trips to Zambia over the past 2 years, our team interviewed nurses, midwives, and environmental health technicians at more than 100 health centers. We have established relationships with health officials at the national, provincial, and district levels. In addition, we have been in close contact with management at Riders and with the trial funder, the Gates Foundation, throughout our study design. Each of these four groups of stakeholders has different viewpoints, but all agree that insufficient reliable transportation for health workers in Zambia is a key challenge.

Without transportation, health workers must cancel outreach visits or walk to sites that are nearby. One nurse that we interviewed walks 4 h each way to an outreach site every month in order to continue providing care to that remote community. Several health workers that we spoke with felt that they were unable to meet their commitment to provide healthcare to the local communities without reliable transportation. Others expressed frustration at having to call taxis to transport patients to the hospital and for not having enough of their own money to help patients pay the cab fare.

District and provincial health officers are similarly concerned. District-wide meetings held to bring all clinics up to speed on new developments or campaigns are sometimes canceled because so many health workers cannot make it. Attempts to manage or improve the quality of the fleet have not been very successful. Health workers do not necessarily take broken motorcycles to designated garages or to any garage. Broken down motorcycles may remain unusable for weeks or even months. The Ministry of Health in Zambia recently organized the distribution of a large donation of motorcycles from China, but has few resources set aside for the eventual repair or maintenance of those motorcycles. Fuel is a constant constraint; even though approximately 15% of each district's health budget is reserved for fuel, it is frequently insufficient. For example, one health center that is responsible for communities that are 50 km away has a fuel allotment of 20 l per month. Since motorcycles' fuel efficiency is about 20 km per liter, this allotment is nowhere near adequate. Policies outlined by the Ministry of Health have little chance of succeeding when resources are so constrained. In addition, some health officials that we spoke with believed that driver negligence was at least partly responsible for the motorcycle failure rates.

With its sizable investment of time and resources, Riders is also a stakeholder in the success of the Zambian health delivery supply chain. Riders has conducted programs in several countries in sub-Saharan Africa and chose to conduct the trial in Zambia after developing a relationship with the Ministry of Health there. Riders feels that its fleet management program can provide a tremendous impact in Zambia and similar countries because so much of the population resides in rural areas away from roads and clinics. By providing rural health workers with reliable transportation, Riders believes that it can overcome the problematic last mile of health delivery in Zambia.

Finally, donors are also stakeholders due to current and potential future involvement (both monetary and political). The Gates Foundation relies on rigorous evidence-based decision-making to identify effective ideas and interventions (The Gates Foundation 2011). By funding this evaluation, the Gates Foundation demonstrates its dedication to finding solutions to last-mile challenges. If the trial results indicate that fleet management is a cost-effective way to improve health outcomes in rural regions, the quantitative analysis can be used to attract and retain future donors.

Our initial trips and interviews provided strong baseline information on which to design the trial. During the trial, we will be interviewing approximately 120 health workers and motorcycle mechanics every week. Additionally, our team has fitted 80 motorcycles with GPS trackers to validate uptime and utilization data. We have trained our in-country data collection officers to use a number of data collection and survey tools, including handheld scanners and netbooks. The data collected over the next 2 years will be instrumental in validating and improving our analytical models of motorcycle downtime.

## 4.2 Step 2: Identify Appropriate Performance Indicators

We focused on choosing a performance measure that addresses the needs of multiple stakeholders. “Performance” could be interpreted in several ways. According to its website, Riders’ vision is “of a world in which no one will die of an easily preventable or curable disease because barriers of distance, terrain or poverty prevent them from being reached” (Riders for Health 2011). This vision touches on several aspects of healthcare delivery and encompasses many possible organizational objectives. For example, Riders may be interested in monitoring the utilization of its motorcycles by health workers or may want to estimate how factors such as road density and terrain affect motorcycle utilization. To isolate the problem of insufficient transportation, and to create a measurable benchmark, we turned to Riders’ mission, which is to “manage [vehicles] on a planned, preventive basis so that the vehicles do not break down however difficult the conditions.” When Riders operates in a region, it trains drivers, maintains the fleet, and establishes a spare parts supply chain. We hypothesized that these components of the Riders program are responsible for reducing the number of breakdowns. Thus, we chose to measure performance as the expected number of motorcycles down in a given period. We reasoned that fewer down motorcycles would permit health workers to perform more outreach and follow-up visits and to transport patients to local hospitals for further treatment. Riders’ provision of driver training could help decrease wear and tear on the motorcycles and thus contribute to improvements in performance. Additionally, a functioning supply chain for service and spare parts would reduce motorcycle downtime.

Our performance metric can be trivially extended to measure the expected *percentage* of a fleet that is down. This modification would enable Riders to compare

**Fig. 19.3** Riders can reduce three out of the four simplified causes of motorcycle downtime

| | | Downtime Source: | |
|---|---|---|---|
| | | Awaiting Repair or Spare Parts | Receiving Repair or Installation |
| Type of Failure: | Repairable | ✓ | ✓ |
| | Consumable | ✓ | |

its operations between regions and over time if fleet size changes. Alternatively, our performance metric could easily be adjusted to measure "up" motorcycles instead of "down" motorcycles.

Through interviews with management at Riders, we learned about the causes of motorcycle downtime and grouped the causes into two broad categories based on the spare parts inventory literature (cf. Sherbrooke (2004)): repairable failures and consumable failures (Fig. 19.3). Repairable failures refer to largely unpredictable failures that necessitate motorcycle repair. For example, a fender bent from an accident or a leaking tire may inhibit a motorcycle from working properly, so that the motorcycle must be sent to a repair shop. From what we have learned in Zambia, motorcycles needing more involved repairs tend to wait weeks and even months before being taken to a garage. Hence, the primary factor determining the length of downtime following a repairable failure appears to be human delay.

Consumable failures involve the replacement of a spare part and in general are more predictable than repairable failures. For example, brake pads must be replaced every 12,000 km for most motorcycles, and the drive chain every 24,000 km. Without servicing at regular intervals, these parts may wear out to the point of causing a motorcycle to fail (either directly or indirectly through being unsafe to drive). Then, once the motorcycle is down and in need of a spare part, the motorcycle remains inoperable until the spare part arrives and can be installed. Because most of these installations are straightforward and take considerably less than a week (the proposed period length of our analysis) to perform, for the purposes of our model, they are instantaneous. From our observations in Zambia, motorcycles needing a quick part installation are typically repaired much more quickly than those suffering from repairable failures.

Once a motorcycle has failed, the length of the ensuing downtime depends on the availability of trained mechanics and, in the case of consumable failures, on the availability of appropriate spare parts. Riders' comprehensive fleet management program addresses both the occurrence of failures and the length of consequent downtimes and thus may be able to reduce the expected number of down motorcycles in a fleet. Several factors influence the number of motorcycles down due to repairable and consumable failures, such as the length of the repair time and the inventory management system in place for spare parts. We were interested in understanding how the different initiatives in Riders' program (namely, driver training, routine maintenance, and inventory management) ultimately impact the number of motorcycles down. In the next section, we organize these relationships into an analytical model.

## 4.3 Step 3: Develop a Model

Consider a region containing $N$ clinics, where clinic $i$ has $M_i$ motorcycles, $i = 1,2,\ldots,N$; the regional fleet has $M_T = \sum_{i=1}^{N} M_i$ motorcycles (see Table 19.1 for a definition of all model parameters). We first develop a status quo model of the expected number of motorcycles down assuming that these clinics rely on private garages to repair fleet motorcycles. Then, recognizing that Riders is able to impact the motorcycle failure rates and inventory stocking levels with its driver training, maintenance, and inventory management, we show how the expected number of motorcycles down changes when Riders is contracted to manage the fleet. In particular, we develop an optimization model to understand how Riders should invest in its driver training and maintenance programs to minimize the expected number of motorcycles that are down, subject to a budget constraint.

Suppose that in the status quo scenario, each clinic in the region uses a different garage to maintain its motorcycles. (A "garage" could be as informal as a neighbor with repair knowledge or a mechanic working out of his or her home.) Let garage $i$ be the garage that manages repairs for clinic $i$. Let $X_T \leq M_T$ be the random number of motorcycles that are down in a given period. Then, per our assumptions on the causes of downtime, the expected number of down motorcycles can be expressed as the expected number down due to a repairable failure (driver negligence/accidents or insufficient maintenance) plus the expected number down due to a consumable failure (waiting for a spare part). We consider each term separately. Consistent with the spare parts literature (cf. Sherbrooke (2004)), we assume that the failures of a single motorcycle follow a Poisson process with rate $\lambda_1 + \lambda_2$, where $\lambda_1$ is the rate of repairable failures and $\lambda_2$ is the rate of consumable failures. We assume that each motorcycle operates and fails independently of the other motorcycles in the fleet (i.e., the failure processes are independent). The failure rates of the fleet at clinic $i$ are therefore $M_i\lambda_1$ and $M_i\lambda_2$.

Repairable failures require the time and resources of a mechanic to return the motorcycle to a functioning state. We suppose that once a motorcycle is down due to a repairable failure, it takes a deterministic amount of time to repair it due to the amount of human delay mentioned earlier. Let $S$, a constant, denote the number of periods that a motorcycle is down following a repairable failure. Then, at any point in time (assuming that the system is at steady state), the expected number of motorcycles down due to repairable failures at clinic $i$ is given by Little's Law and is simply $M_i\lambda_1 S$.

**Table 19.1** Comprehensive list of parameters

| Parameter | Description |
|---|---|
| $\alpha$ | Service level for the spare part in the centralized inventory management case |
| $\beta_j$ | Effectiveness of program $j$ at reducing failure rate $\lambda_j$; $j = 1,2$ |
| $\lambda_1$, $\lambda_1(e_1)$ | Rate of repairable failures for a single motorcycle (exogenous and as a function of effort $e_1$) |
| $\lambda_2$, $\lambda_1(e_2)$ | Rate of consumable failures for a single motorcycle (exogenous and as a function of effort $e_2$) |
| $\mu_i$, $\mu_T$ | Mean of demand for spare part at garage $i$ and for the region |
| $\phi(\cdot)$, $\Phi(\cdot)$ | Standard normal density and distribution functions, respectively |
| $\sigma_i$, $\sigma_T$ | Standard deviation of demand for spare part at garage $i$ and for the region |
| $A, \tilde{A}$ | Riders' available budget and adjusted budget defined for notation convenience: $\tilde{A} \equiv A - c_3\sqrt{\beta_2 M_T}$ |
| $b_1(e_1)$, $b_2$ | Benefit–cost ratios for driver training and routine maintenance, respectively |
| $B_i$ | Number of backorders for the spare part at garage $i$ |
| $c_j$ | Cost of effort $e_j$; $j = 1,2$ |
| $e_j$ | Effort that Riders invests in its core competencies of training programs ($e_1$) and maintenance ($e_2$) |
| $f(e_2, y_i)$ | Current-period holding and penalty costs of the inventory policy with stocking level $y_i$ and effort $e_2$ |
| $h$ | Unit holding cost per period of the spare part |
| $D_i$, $D_i(e_2)$ | Random demand for spare part at clinic $i$ in one period (exogenous and as a function of effort $e_2$) |
| $L(z)$ | Standard normal loss function |
| $M_i$, $M_T$ | Number of motorcycles at clinic $i$ ($i = 1,2,\ldots,N$) and in the regional fleet |
| $N$ | Number of clinics in the region |
| $p$ | Unit penalty cost of the spare part |
| $S$ | Number of periods that a motorcycle is down following a repairable failure |
| $\text{TB}_\text{d}$, $\text{TB}_\text{r}$ | Total expected number of backorders per period for the decentralized and Riders scenarios, respectively |
| $\text{TC}_\text{d}$, $\text{TC}_\text{r}$ | Total expected cost per period for the decentralized and Riders scenarios, respectively |
| $X_i$ | Number of motorcycles down at clinic $i$ |
| $X_T, X_T(e_1, e_2, y)$ | Total number of motorcycles down across the region (exogenous and as a function of decisions $e_1$, $e_2$, and $y$) |
| $y_i$, $y_T$ | Stocking level for the spare part at garage $i$ and the regional stocking level for the spare part under Riders' centralized inventory management |
| $z$ | Score of the standard normal distribution corresponding to service level $\alpha$ |

The discussion of consumable failures is more nuanced. When a motorcycle fails in need of a critical spare part, the motorcycle will remain inoperative until that part is available. For simplicity, we consider a single critical spare part. We assume that if the part is in stock, it is available more or less immediately. On the other hand, if the spare part is not in stock when the motorcycle fails, the motorcycle will remain down until the spare part arrives and can be installed. Thus, the downtime due to consumable failures depends on the inventory policy of the local garage. In fact, the expected number of motorcycles down at a clinic that need the spare part is exactly the expected number of backorders for that spare part at the clinic's garage. We arrive at the following performance metric to evaluate fleet management:

$$\mathbb{E}[X_i] = M_i\lambda_1 S + \mathbb{E}[B_i] \; = \; M_i\lambda_1 S + \mathbb{E}\left[(D_i - y_i)^+\right] \;, \tag{19.1}$$

where $X_i$ is the number of motorcycles down at clinic $i$ and $B_i$ is the number of backorders for the spare part at corresponding garage $i$ in a period (given stocking level $y_i$ and random demand $D_i$ for the spare part). The expected total number of motorcycles down for the region is $\mathbb{E}[X_T] = \sum_{i=1}^{N} \mathbb{E}[X_i]$.

The expected number of backorders at garage $i$ depends on the garage's inventory management policy. We assume that the garage uses a base-stock policy; that is, in each period, garage $i$ replenishes its inventory of the spare part up to $y_i$ units. For simplicity, we focus on a spare part for which there is low demand and high holding cost (per the definition of consumable items in the spare parts literature (Sherbrooke 2004)), such as clutch overhaul kits. Typically, a motorcycle clutch should be overhauled every 30,000 km; by Riders' estimates, health workers' motorcycles travel 900 km per month on average. Hence, a motorcycle should only need a clutch overhaul kit about every 3 years, so demand for this part is low. Further, compared to other spare parts for the low-end motorcycles typically used in sub-Saharan Africa, clutch overhaul kits are expensive for local garages to stock. Based on our initial survey data from Zambia, clutch overhaul kits cost about \$35, a large sum in a country where the average monthly income for a family is about \$100 (Encyclopædia Britannica 2010b).

We assume that the associated costs are linear, including a unit holding cost $h$ per period and a unit penalty cost $p$.[3] By our failure rate assumptions, demand for clutch overhaul kits per motorcycle per period is Poisson with parameter $\lambda_2$. Then demand at clinic $i$, $D_i$, is Poisson with parameter $M_i\lambda_2$. Because our model context is a developing region, our assumption that the fleet failure rate grows linearly in the fleet size is reasonable. (For a resource-rich region, a larger fleet size might imply that individual motorcycles are used less frequently and so a linear relationship between fleet size and fleet failure rates would not hold.)

Now, if Riders is contracted to take over fleet management in the region, the status quo model in (19.1) will change to reflect the impact that Riders has on the failure rates. Until now, we have assumed that the failure rates $\lambda_1$ and $\lambda_2$ are exogenously given and constant. However, in reality, the failure rates depend on the effort that Riders puts into its program initiatives. Let $e_j \in [0,1]$ represent the effort that Riders invests in its core competencies: $e_1$ is the effort invested in driver training programs to reduce the repairable failure rate (now a function of $e_1$, $\lambda_1(e_1)$), and $e_2$ is the effort invested in routine maintenance intended to reduce the consumable failure rate (now a function of $e_2$, $\lambda_2(e_2)$). In addition, Riders is able to reduce the expected number of motorcycles that are down by choosing the spare part stocking levels $y_i$ (which affect the expected number of spare parts backorders). Because Riders' choices of decision variables $e_1$, $e_2$, and $y_i$ drive demand for spare parts and

[3] We omit procurement costs in our model, as the average procurement over the long term is independent of the stocking decision.

ultimately the total number of motorcycles that are down, these values are now also functions of the decision variables: $D_i(e_2)$ and $X_T(e_1,e_2,y)$, respectively, where $y=(y_1,y_2,\ldots,y_N)^T$.

If Riders had an unlimited budget, it would be able to hold large spare part inventories and to invest maximum effort in driver training and maintenance programs. Let $c_j$ be the cost of fully investing in program initiative $j$, $j=1,2$ (i.e., $c_j$ is the cost incurred when $e_j=1$); then, the cost of exerting effort $e_j$ is $c_je_j$. In the absence of an unlimited budget, Riders must trade off which core competencies it will invest in (driver training, routine maintenance, inventory management). To investigate this cost tradeoff, suppose that Riders seeks to maximize regional mobility by minimizing the expected number of motorcycles that are down, $\mathbb{E}[X_T(e_1,e_2,y)]$. Riders' challenge can be formulated as the following optimization problem:

$$\begin{aligned}
\min \quad & \mathbb{E}[X_T(e_1,e_2,y)] && (P)\\
\text{s.t.} \quad & c_1e_1+c_2e_2+\textstyle\sum_{i=1}^{N}\mathbb{E}[f(e_2,y_i)] \leq A,\\
& \qquad\qquad\qquad\qquad 1\geq e_1,e_2\geq 0,\\
& \qquad\qquad\qquad\quad y_1,y_2,\ldots,y_N\geq 0,
\end{aligned}$$

where $c_1,c_2>0$, and $A>0$ is the available budget. In addition, $\mathbb{E}[f(e_2,y_i)]$ is the expected current-period holding and penalty costs of the inventory management policy given stocking level $y_i$ and effort $e_2$:

$$\mathbb{E}[f(e_2,y_i)] = h\mathbb{E}\left[(y_i-D_i(e_2))^+\right]+p\mathbb{E}\left[(D_i(e_2)-y_i)^+\right]. \tag{19.2}$$

Modeling demand as a function of invested effort has been done in the marketing and operations management literature; for examples, see Taylor (2002) and Rao (1990).

In the remainder of this chapter we work with the probability density and distribution functions of this aggregated Poisson distribution. To simplify our analysis, we approximate the Poisson distribution (parameter $M_i\lambda_2(e_2)$) using a normal distribution with mean $M_i\lambda_2(e_2)$ and standard deviation $\sqrt{M_i\lambda_2(e_2)}$; let $\phi(\cdot)$ and $\phi(\cdot)$ be the standard normal density and distribution functions, respectively. This approximation improves as $M_i\lambda_2(e_2)$ increases, so it is more accurate for larger fleet sizes.[4] Of course, one could proceed without applying a normal approximation;

[4] The Poisson distribution with parameter $\lambda$ can be approximated using a normal distribution with mean and variance $\lambda$ by the central limit theorem. Concordant with the theorem, as $\lambda$ increases, the approximation becomes more accurate. For our base case in Sect. 4.4, $\lambda\approx 200$. When tested against the null hypothesis that 1,000 random samples drawn from a Poisson distribution with parameter 200 followed a normal distribution with mean and variance 200, a Kolmogorov–Smirnov test resulted in a $p$-value of 0.3364. In addition, the normal distribution is commonly used to model demand in operations management literature, despite its being a continuous distribution with probability placed on negative values (cf. Lee et al. (1997) and Lee and Özer (2007)). Because our scenario here is for high values of $\lambda$, the probability of negative demand is negligible, and we thus conclude that the normal distribution is a reasonable approximation.

however, the calculations would be considerably more challenging. The results that follow are an approximation that reflects the trends and provides useful insights.

Existing data to populate this model are scarce. As part of the 2-year trial, we are collecting data on holding and penalty costs ($h$ and $p$) through conversations with mechanics. Interviews with health workers will help us estimate the length of downtime following a failure ($S$) and the motorcycle failure rates ($\lambda_1$ and $\lambda_2$). Regarding the cost parameters, the cost of inventory management can be calculated once stocking level $y_i$ and demand $D_i(e_2)$ are established. Estimating $c_1$ and $c_2$ is more involved. While $e_j = 0$ corresponds to doing nothing towards program initiative $j$, Riders may calibrate the meaning of $e_j = 1$ as well as the corresponding cost of effort $c_j$ ($j = 1,2$) differently for each region it operates in. Here, we assume that the cost of $e_1 = 1$ is the cost of providing extensive and ongoing driver training for all drivers in a region and that the cost of $e_2 = 1$ is the cost of providing routine maintenance to all motorcycles in the region. Service schedules are well-known and can be used to estimate the former, whereas experience in maintaining motorcycles is necessary to estimate the latter.

In the next section, we solve a special case of $(P)$ to understand how Riders should implement its fleet management program.

## *4.4 Step 4: Use the Model to Evaluate the Program*

In the previous section, we developed a status quo metric of motorcycle downtime for a decentralized system of $N$ clinic–garage pairs, and then showed how Riders could use that metric to balance efforts invested in its three core competencies of driver training, routine maintenance, and inventory management. We now analyze this model to gain insights into Riders' performance. In the following sections, we first explore how Riders improves the status quo scenario by centralizing resources, then solve Riders' optimization problem under a special case of an exogenous service level and quadratic failure rates, and finally perform univariate sensitivity analyses to understand the robustness of the model to various parameters.

### 4.4.1 Centralized Spare Parts Inventory Management

Riders' optimization problem $(P)$ can be simplified by observing that the status quo model involves decentralized inventory management. If Riders is contracted to manage the regional fleet, its management of the inventory will be centralized, and so the expected penalty and holding costs as well as the expected number of backorders will all be lower under Riders.

Instead of considering the needs of each clinic individually, Riders can analyze the demand patterns of all clinics in a region and leverage the aggregated demand to keep less inventory. We now establish the benefits of centralization through demand pooling; our results here parallel those of Eppen (1979). First, we define our notation. Let $L(z)$ represent the standard normal loss function: $L(z) \equiv \mathbb{E}\left[(\xi - z)^+\right]$, where $\xi$ is a standard normal random variable. Using properties of the normal distribution, we can rewrite the expected current-period cost in (19.2) for garage $i$ under the decentralized system as $c_3\sigma_i$, where $c_3 = hz + (p+h)L(z)$, $y_i = \mu_i + \sigma_i z$, $\mu_i = M_i\lambda_2(e_2)$, and $\sigma_i = \sqrt{M_i\lambda_2(e_2)}$. Similarly, we can rewrite the expected number of backorders at garage $i$ during a period as $L(z)\sigma_i$. Then the **t**otal expected **c**ost and **b**ackorders per period for the **d**ecentralized system are, respectively, $\mathrm{TC_d} = c_3\sum_{i=1}^{N}\sigma_i$ and $\mathrm{TB_d} = L(z)\sum_{i=1}^{N}\sigma_i$.

We now calculate the total expected cost and backorders for the centralized system. Suppose that the fleets of all $N$ clinics are managed centrally by Riders. Then the demand faced by the system is the sum of the demands faced by each clinic, and we can approximate the aggregated Poisson demand distribution as a normal distribution with mean $\mu_T$ and standard deviation $\sigma_T$, where

$$\mu_T = \sum_{i=1}^{N}\mu_i = M_T\lambda_2(e_2) \quad \text{and} \quad \sigma_T = \sqrt{\sum_{i=1}^{N}\sigma_i^2} = \sqrt{\lambda_2(e_2)M_T}\,.$$

Now, the total per-period expected cost and backorders of Riders' centralized system are, respectively, $\mathrm{TC_r} = c_3\sigma_T$ and $\mathrm{TB_r} = L(z)\sigma_T$. Several observations are immediate. Since $\sigma_T \le \sum_{i=1}^{N}\sigma_i$, the total expected cost and backorders are lower in Riders' centralized system; that is, $\mathrm{TC_r} \le \mathrm{TC_d}$ and $\mathrm{TB_r} \le \mathrm{TB_d}$. Further, we observe that if each clinic has the same number of motorcycles (i.e., $\sigma_i = \bar{\sigma}$ for all $i$), then $\mathrm{TC_d} = c_3N\bar{\sigma}$ and $\mathrm{TC_r} = c_3\sqrt{N}\bar{\sigma}$. (Similarly, $\mathrm{TB_d} = L(z)N\bar{\sigma}$ and $\mathrm{TB_r} = L(z)\sqrt{N}\bar{\sigma}$.) Thus, economies of scale are present: when each clinic has the same number of motorcycles, the total expected cost and backorders in a centralized system increase as the square root of the number of clinics (whereas this increase is linear in a decentralized system).

We conclude that Riders' central management of a system of clinics' fleets can be expected to bring better performance (i.e., fewer expected backorders) at a lower cost than the management of those same fleets by a decentralized network of private garages.

### 4.4.2 Exogenous Service Level and Quadratic Failure Rates

Because of the analytical intractability of $(P)$, we now propose two modifications. First, we consider controlling the spare parts inventory policy not with a series of stocking levels but with a single, system-wide service level. We have so far assumed that the stocking levels $y_i$ are decision variables. In light of the demand pooling, Riders only chooses $y_T = \mu_T + \sigma_T z$, where $z$ corresponds to the (type I) service level $\alpha$: $z \equiv \phi^{-1}(\alpha)$. Incorporating $\alpha$ as a decision variable complicates the solution to $(P)$; at optimality, it is a function of the dual variable associated with the budget constraint and so offers some insights, but the problem is difficult to solve analytically. For the following analysis, we assume the service level, $\alpha$, is given exogenously. This special case illustrates Riders' fundamental tradeoff: should Riders invest more effort into reducing the repairable failure rate through driver training programs or into reducing the consumable failure rate through improved routine maintenance?

In addition, the objective of $(P)$ is neither convex nor concave in its variables for general $\lambda_1(e_1)$ and $\lambda_2(e_2)$. The failure rates should each be decreasing in effort exerted, and arguably, there are diminishing returns in failure rate reduction as more effort is exerted. Thus, it is reasonable to assume that $\lambda_j'(e_j) \leq 0$ and $\lambda_j''(e_j) \geq 0$, $j = 1,2$. To satisfy these conditions, we propose that $\lambda_j(e_j) = \beta_j(1-e_j)^2$ for $j = 1,2$.[5] We can interpret $\beta_j > 0$ as the effectiveness of program $j$ at reducing failure rate $\lambda_j$, $j = 1,2$. The revised problem is a quadratic program:

$$\begin{aligned} \min \quad & M_T\beta_1(1-e_1)^2\, S + \mathrm{TB_r} \\ \text{s.t.} \quad & c_1e_1 + c_2e_2 + c_3\sqrt{\beta_2 M_T}(1-e_2) \leq A, \\ & \qquad\qquad\qquad\qquad 0 \leq e_1, e_2 \leq 1, \end{aligned}$$

where $c_1, c_2 > 0$, and $c_3$ are as defined above. From our earlier derivation, we substitute $\mathrm{TB_r} = L(z)\sigma_T$ and $\sigma_T = \sqrt{\lambda_2(e_2)M_T}$. Additionally, we define $\tilde{A} \equiv A - c_3\sqrt{\beta_2 M_T}$. Then the above becomes

$$\begin{aligned} \min \quad & M_T\beta_1(1-e_1)^2 S + L(z)\sqrt{\beta_2 M_T}(1-e_2) \qquad (P'), \\ \text{s.t.} \quad & c_1e_1 + c_2e_2 - c_3\sqrt{\beta_2 M_T}\; e_2 \leq \tilde{A} \\ & \qquad\qquad\qquad 0 \leq e_1, e_2 \leq 1. \end{aligned}$$

Now, the marginal benefit (i.e., decrement to the expected number of motorcycles down) of investing more in driver training is $2M_T\beta_1 S(1-e_1)$, and the marginal benefit of investing more in maintenance is $L(z)\sqrt{\beta_2 M_T}$. Similarly, the marginal costs of investing more in driver training and maintenance are $c_1$ and $c_2 - c_3\sqrt{\beta_2 M_T}$, respectively. We define the benefit–cost ratios for driver training and maintenance, respectively, as

$$b_1(e_1) \equiv \frac{2M_T\beta_1 S(1-e_1)}{c_1} \quad \text{and} \quad b_2 \equiv \frac{L(z)\sqrt{\beta_2 M_T}}{c_2 - c_3\sqrt{\beta_2 M_T}}.$$

The solution to $(P')$ can be interpreted using these marginal benefit–cost ratios. Since $(P')$ is a convex program, its solution can be found using the Karush–Kuhn–Tucker conditions:

**Solution to** $(P')$ . *If* $\tilde{A} \geq c_1 + c_2$, *then* $e^* = (1,1)$. *Otherwise*:

[5] Riders' goal is to maintain a fleet with zero breakdowns, so here we allow the failure rates to decrease to zero given maximum effort. If failures still occur given $e_j = 1$, a minimum failure rate $\lambda_0$ can easily be incorporated into this formulation so that $\lambda_j(1) > 0$, $j = 1,2$.

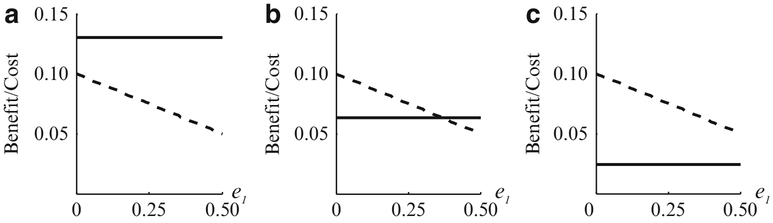


**Fig. 19.4** Hypothetical scenario illustrating the three possible solutions; the *solid line* is $b_2$ and the *dashed line* is $b_1(e_1)$. In all graphs, $\beta_1 = 0.1$, $\beta_2 = 10$, $S = 1$, $c_1 = c_2 = 200$, $\tilde{A} = 100$, $p = 0.02$, and $h = 0.01$. We vary the service level $\alpha$ to distinguish between the three solutions: $\alpha = 0.25$ (Fig. 19.4a), $\alpha = 0.50$ (Fig. 19.4b), and $\alpha = 0.75$ (Fig. 19.4c)

1. *If $b_1(\tilde{A}/c_1) > b_2$, then it must be that $\tilde{A} \le c_1$, and driver training should be prioritized: $e^* = (\tilde{A}/c_1, 0)$.*
2. *If $b_1(0) < b_2$, then maintenance should be prioritized: if $\tilde{A} \le c_2$, then $e^* = (0, \tilde{A}/(c_2 - c_3\sqrt{\beta_2 M_T}))$, and if $c_2 < \tilde{A} < c_1 + c_2$, then $e^* = (\frac{\tilde{A}-(c_2-c_3\sqrt{\beta_2 M_T})}{c_1}, 1)$.*
3. *Otherwise, there exists $\hat{e} \in [0, \tilde{A}/c_1]$ such that $b_1(\hat{e}) = b_2$. The solution is to balance the available budget between the two initiatives: $e^* = (\hat{e}, \frac{\tilde{A}-c_1\hat{e}}{c_2-c_3\sqrt{\beta_2 M_T}})$.*

The solution relies on the comparison between $b_1(e_1)$, which is a decreasing linear function of $e_1$, and $b_2$, which is constant in $e_1$; Fig. 19.4 illustrates the three possibilities. For restrictive budgets (i.e., $\tilde{A} < c_1 + c_2$), Riders must balance its two program initiatives. When the marginal benefit–cost ratio for driver training is greater than that of maintenance for all feasible $e_1$ (sufficient to check for $e_1 = \tilde{A}/c_1$), it is optimal to use the entire budget to fund driver training (solution 1, Fig. 19.4c). This comparison is equivalent to $\frac{2M_T\beta_1 S(c_1-\tilde{A})}{c_1^2} > \frac{L(z)\sqrt{\beta_2 M_T}}{c_2-c_3\sqrt{\beta_2 M_T}}$. Hence, driver training becomes a more effective investment as the fleet size ($M_T$) increases, as its effectiveness coefficient ($\beta_1$) increases, and as the length of downtime following a repairable failure ($S$) increases. If $c_1 > 2\tilde{A}$, $b_1(e_1)$ is decreasing in the marginal cost of driver training; otherwise, $b_1(e_1)$ is increasing in $c_1$.

On the other hand, when the marginal benefit–cost ratio for driver training is less than that of maintenance for all feasible $e_1$ (sufficient to check for $e_1 = 0$), it is optimal to prioritize the maintenance program (solution 2, Fig. 19.4a). Only if there are funds remaining after investing the maximum in the maintenance program should funds be allocated to the driver training program. The marginal benefit–cost comparison for this scenario is equivalent to $\frac{2M_T\beta_1 S}{c_1} > \frac{L(z)\sqrt{\beta_2 M_T}}{c_2-c_3\sqrt{\beta_2 M_T}}$. The maintenance program becomes a more attractive investment as its effectiveness coefficient ($\beta_2$) increases, as the fleet size increases, and as the cost $c_3$ increases (hence, $b_2$ is also increasing in the unit penalty and holding costs, $p$ and $h$, respectively). Additionally, $b_2$ is decreasing in the service level ($\alpha$) and the marginal cost of the maintenance program ($c_2$).

**Table 19.2** Parameter settings for sensitivity analysis. The remaining parameters were held constant at $B = 100$, $N = 100$, and $M_i = 1$ for all $i$; $c_2 = 200$, $p = 0.02$, $h = 0.01$, $\beta_1 = 0.1$, and $S = 2$

| Parameter | Base value | Range |
|---|---|---|
| $\alpha$ | 0.75 | $[0.50, 1.00]$ |
| $\beta_2$ | 0.5 | $[0.1, 2.0]$ |
| $c_1$ | 150 | $[50, 400]$ |

Finally, if there exists an $\hat{e} \in [0,\tilde{A}/c_1]$ such that the marginal benefit–cost ratios for driver training and for maintenance are the same (solution 3, Fig. 19.4b), then Riders should split its budget between the two program initiatives. In particular, $\hat{e} = 1 - \frac{c_1 L(z)\sqrt{\beta_2 M_T}}{2M_T \beta_1 S(c_2 - c_3\sqrt{\beta_2 M_T})}$. Since $e_1^* = \hat{e}$, we see that Riders should shift more of its budget to the driver training program as the costs $c_1$ and $c_3$ decrease, or as the cost $c_2$ increases. That is, Riders should shift its budget toward driver training as that option becomes a more effective investment in terms of reducing the number of motorcycles that are down on average.

### 4.4.3 Sensitivity Analysis

In addition to identifying the analytical solution, we have also performed univariate sensitivity analyses on some model parameters to learn which parameters the model is most sensitive to. The base case and ranges are given in Table 19.2. For simplicity, we assumed that each of $N = 100$ clinics has one motorcycle each: $M_i = 1$ for $i = 1, 2, \ldots, N$. Further, we held some cost parameters constant: $B = 100$, $c_2 = 200$, $p = 0.02$, and $h = 0.01$. We observed that changes in $c_2$, $p$, and $h$ did not impact the model. In addition, changes in the effectiveness coefficient and the repairable failure downtime had little impact on the expected number of motorcycles down; for example, changing $S$ from 0.5 to 4 periods only elicited a change from 2.06 to 2.10 motorcycles down on average, and changing $\beta_1$ from 0.01 to 4 resulted in a change from 1.97 to 2.11 motorcycles down on average. We held these parameters constant at $\beta_1 = 0.1$ and $S = 2$ for the results presented here.

For the base case of parameters outlined in Table 19.2, the total cost of maximum effort is 300—far exceeding the actual budget $\tilde{A}$. If Riders chooses to do nothing, then for the base case, on average 22.1 motorcycles are down out of the total fleet of 100. The optimal solution for the base case setting is 2.1 motorcycles; that is, with a budget of 100, Riders can decrease the expected number of motorcycles down from 22.1 to 2.1. We verify that the expected number of motorcycles down decreases as the service level increases (Fig. 19.5a), since a higher service level leads to fewer backorders on average. Similarly, we find that as the effectiveness of maintenance programs ($\beta_2$) increases, the expected number of motorcycles down increases (Fig. 19.5b). This result is not universally true; by taking a first-order condition, we find that the objective evaluated at $e^*$ is not necessarily increasing or

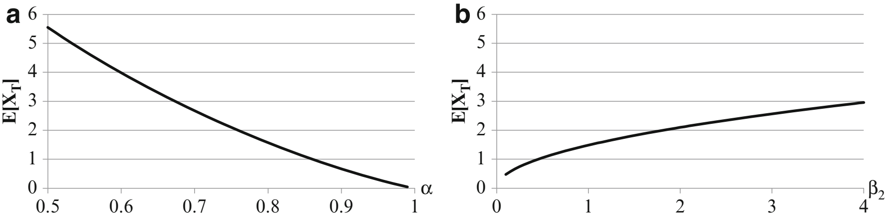
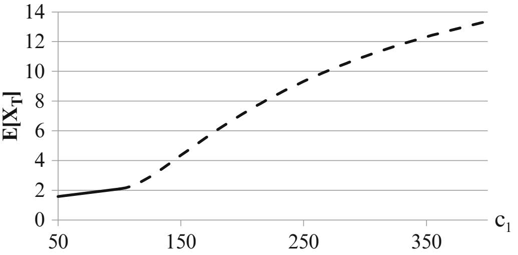


**Fig. 19.5** $\mathbb{E}[X_T]$ as a function of the type-I service level $\alpha$ (Fig. 19.5a) and of the maintenance effectiveness coefficient $\beta_2$ (Fig. 19.5b). Splitting the budget between driver training and maintenance was optimal in all of these scenarios

**Fig. 19.6** $\mathbb{E}[X_T]$ as a function of the effort costs $c_1$. The *solid line* indicates that splitting the budget was optimal; the *dashed line* indicates that allocating the entire budget to the driver training programs was optimal



decreasing in $\beta_2$. For the base case parameter setting, however, it is the case that the expected number of motorcycles down is increasing in $\beta_2$. As the effectiveness of the maintenance program increases, the marginal benefit of investing in maintenance increases, and the marginal cost decreases, so the solution shifts to increase $e_2^*$ at the expense of $e_1^*$. However, due to the relative benefits and costs in the base parameter setting, the expected number of motorcycles down actually increases as $\beta_2$ increases.

The model is more sensitive to changes in the cost of driver training programs, $c_1$. In Fig. 19.6, we see that increasing $c_1$ changes the optimal solution from splitting the budget between the initiatives to allocating the entire budget to driver training. Recall that the benefit–cost ratios intersect when the optimal solution is to split the budget. When $c_1$ increases, the intercept of $b_1(e_1)$ decreases and the slope is less and less negative, effectively pulling $b_1(e_1)$ above $b_2$ for all feasible $e_1$. In addition, the expected number of motorcycles that are down increases in this cost.

These solutions articulate how Riders should allocate its budget in order to maximize regional mobility. In the next section, we provide an analysis of this model, its limitations, and possible extensions.

## 5 Model Results: Next Steps

The model that we have presented here represents a first pass at understanding the effectiveness of Riders' fleet management program. We found that by increasing investment in driver training programs and routine maintenance, Riders can decrease the motorcycle failure rates and hence the expected number of motorcycles down in a regional fleet. The attendant increases in fleet availability can improve health worker mobility and help health workers accomplish their outreach objectives.

Specifically, our model connects reductions in motorcycle failure rates and improvements in spare parts inventory management with decreases in motorcycle downtime. The cost and effectiveness parameters of the model determine how Riders should split its investment in driver training and maintenance. To estimate the various parameters of the model, we have developed logistics surveys to ask health workers and motorcycle mechanics each week during the trial. Questions about the uptime and utilization of the health workers' motorcycles, and about causes of downtime, can help us estimate the failure rates and lengths of ensuing downtimes. In the garage survey, we will track both a high-demand, low-value critical spare part and a low-demand, high-value critical spare part to see how often mechanics order and use the parts in repairs. Other questions about these parts are designed to estimate the procurement, holding, and penalty costs associated with each part.

The model provides useful insights, but has some limitations that could be addressed in future research. First, the failure rates realistically depend on the number of kilometers that each motorcycle is driven per period since a motorcycle that is not driven at all cannot fail. The model could be modified so that the failure rates are also functions of this distance, and the objective could be changed to maximize the average number of kilometers driven per period. Second, the service level could also be treated as a decision variable: how responsive should Riders' inventory management program be? And finally, Riders may be able to proactively impact the length of downtime following a repairable failure; that is, $S$ may be a function of $e_2$.

Our estimation of status quo operations (i.e., the decentralized network of private garages) is very conservative and likely overestimates status quo performance. For example, one garage owner that we interviewed kept no inventory at all. Instead, when a customer needed a part, he scheduled a 2-day trip to the capital city to buy the part there (the relative importance of the customer influenced how quickly the trip was undertaken). We also observed a "culture of waiting" in Zambia and believe that in general, garage owners and customers alike may not emphasize high service levels or fast turnaround times. The differences between the status quo and operations under Riders are probably greater than indicated by our model because of our conservative assumptions.

Our team has learned a great deal about conducting research in a resource-limited environment. Because health delivery and fleet management in Zambia are so different from those in the developed world, early in the trial design phases, we sketched out a list of primary data that we would need to collect in order to evaluate the effectiveness of Riders compared to status quo operations. To collect primary data in the Southern Province, we had to plan and build a massive infrastructure of data collection officers, weekly surveys, and GPS trackers. The surveys alone took a year of testing and retesting to finalize, and installing the GPS trackers on a randomized selection of motorcycles itself took months due to the challenges of moving from clinic to clinic over poor roads. A lack of cell phone signal in many areas has hampered our ability to regularly make appointments or conduct surveys. The data that we will collect in the trial will be unique in its breadth and depth, and we hope that future research will be able to build on what we have learned.

From a modeling standpoint, we have refined our assumptions and understanding continuously as we learn more about the operations on the ground in Zambia. As we have emphasized, the modeling process is iterative. The results that we have identified here will be reviewed and either validated or rejected by our stakeholders. Based on their feedback, we plan to modify the model and its assumptions appropriately.

## 6 Conclusions

Health delivery in developing countries is typically underfunded and understaffed. Developing countries bear much more than their fair share of the global disease burden while accounting for only a small fraction of global healthcare spending (Schieber and Maeda 1999). Exacerbating the lack of funding in many nations are accessibility issues. Barriers in the last mile of health delivery prevent millions of people from accessing even basic healthcare. These last-mile issues could be resolved through an influx of resources and support: new clinics, more health workers, more drugs, and better road networks. But policy makers in developing regions are constrained by very limited resources. Identifying the best solution in light of constraints is a strength of operations research. Thus, researchers can use operations research modeling tools to help improve the last mile of the health delivery supply chain.

In this chapter, we have adapted traditional problem-solving guidelines to frame our model of a problem in the context of sub-Saharan Africa. In particular, we identified challenges including a lack of data and a need to tailor the problem to operations on the ground. To illustrate one approach to addressing these challenges, we have provided a case study on our work with Riders for Health, an organization devoted to overcoming the challenges of health delivery in rural Africa by increasing the accessibility of healthcare.

For our project in Zambia, we have worked to address the lack of available data by establishing an infrastructure to collect primary data. We have also performed sensitivity analyses on model parameters to understand which parameters should be most carefully estimated by decision makers. In addition, we tailored our model to Riders' objective and chose to minimize the expected number of motorcycles down subject to a budget constraint rather than formulate the optimization as a cost minimization problem. Many of our modeling assumptions were informed by interviews conducted in Zambia and conversations with Riders management over the past 2 years.

Using our model, we are able to quantify the reduction in costs and backorders from Riders' centralized spare parts inventory management. In addition, we have modeled the connection between Riders' investment in its driver training and routine management programs and the reduction in repairable and consumable failure rates, respectively. We are collecting data in our trial in the Southern Province of Zambia in an ongoing effort to improve our estimates of the parameter values that we have used here; the data will serve to validate our model or inform modifications. This work is an initial step in our evaluation of Riders' performance. Future steps include linking motorcycle availability to increased numbers of health interventions delivered and finally to regional health outcomes.

Research on health delivery supply chains for resource-limited regions is still nascent. Our hope is that operations researchers can use what we have learned in the design and implementation of this trial to model other last-mile problems. Particularly in resource-limited regions, these problems are complex, and researchers must engage stakeholders at every step of the process to ensure that the resulting models are relevant and applicable. Improvements in health delivery could have a significant impact on health in developing nations, particularly when implemented using rigorous approaches from implementation research. Operations research modeling will be instrumental to improving the healthcare delivery supply chain in the developing world, and evaluations such as ours are necessary to identify effective programs and support evidence-based decision-making.

**Acknowledgements** This material is based upon work supported under the National Defense Science and Engineering Graduate Fellowship, the Stanford Graduate Fellowship, and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1147470. The author would like to thank the editor and the anonymous referee for their insightful suggestions which greatly strengthened the content of this chapter. Additionally, Margaret Brandeau, Hau Lee, and Hugo Mora provided excellent feedback on the model and analysis.

## References

Acharya LB, Cleland J (2000) Maternal and child health services in rural Nepal: Does access or quality matter more? Health Pol Plann 15(2):223–229

Ahmed R, Hossain M (1990) Developmental impact of rural infrastructure in Bangladesh. International Food Policy Research Institute and the Bangladesh Institute of Policy Studies, Washington

Anupindi R, Aundhe MD, Sarkar M (2009) Healthcare delivery models and the role of telemedicine. In: Swaminathan JM (ed) Indian economic superpower: Fiction or future? World Scientific, New Jersey

Apte A (2009) Humanitarian logistics: A new field of research and action. Found Trends Technol Inf Oper Manag 3(1):1–100

Balcik B, Beamon BM (2008) Facility location in humanitarian relief. Int J Logist: Res Appl 11(3):101–121

Beamon BM, Kotleba SA (2006) Inventory modelling for complex emergencies in humanitarian relief operations. Int J Logist: Res App 9(1):1–18

Bryce J, Victora CG, Habicht JP et al (2004) The multi-country evaluation of the integrated management of childhood illness strategy: Lessons for the evaluation of public health interventions. Am J Publ Health 94(3):406–415

Chen L, Evans T, Anand S et al (2004) Human resources for health: Overcoming the crisis. Lancet 364(9449):1984–1990

Chesney M (2000) Factors affecting adherence to antiretroviral therapy. Clin Infect Dis 20(S2):S171–S176

Chokshi DA, Kesselheim AS (2008) Rethinking global access to vaccines. BMJ 336(7647): 750–753

De Angelis V, Mecoli M, Nikoi C et al (2007) Multiperiod integrated routing and scheduling of World Food Programme cargo planes in Angola. Comput Oper Res 34(6):1601–1615

Deo S, Topp S, Garcia A et al (2010) Two queue or not two queue: The impact of integrating HIV and outpatient health services in an urban health clinic in Zambia. Working paper

Dovlo D (2005) Wastage in the health workforce: Some perspectives from African countries. Hum Resource Health. doi:10.1186/1478-4491-3-6

Encyclopædia Britannica (2010) Austria. Available via Encyclopædia Britannica Online. http://www.britannica.com/EBchecked/topic/44183/Austria. Cited 3 Mar 2011

Encyclopædia Britannica (2010) World data: Zambia. Available via Encyclopædia Britannica Online. http://media-2.web.britannica.com/eb-media/39/77939-004-0C4ABAD0.pdf. Cited 3 Mar 2011

Economist (2008) Telemedicine comes home. Available via The Economist Online. http://www.economist.com/node/11482580. Cited 10 Mar 2011

Eppen GD (1979) Effects of centralization on expected costs in a multi-location newsboy problem. Manag Sci 25(5):498–501

The Gates Foundation (2011) www.gatesfoundation.org. Cited 20 Jul 2011

Hampshire K (2002) Networks of nomads: Negotiating access to health resources among pastoralist women in Chad. Soc Sci Med 54(7):1025–1037

Hazin R, Qaddoumi I (2010) Teleoncology: Current and future applications for improving cancer care globally. Lancet Oncol 11(2):204–210

Hillier FS, Lieberman, GJ (2005) Introduction to operations research, 8th edn. McGraw-Hill, New York

Howe J, Richards P (1984) Rural roads and poverty alleviation. Intermediate Technology Publications, London

Istepanian RSH, Laxminarayan S, Pattichis CS (2006) M-health: Emerging mobile health systems. Springer, New York

Lee HL, Özer Ö (2007) Unlocking the value of RFID. Prod Oper Manag 16(1):40–64

Lee HL, Tayan B (2007) Riders for health: Healthcare distribution solutions in sub-Saharan Africa. Stanf GSB Case GS-58

Lee HL, Padmanabhan V, Whang S (1997) Information distortion in a supply chain: The bullwhip effect. Manag Sci 43(4):546–558

Lee HL, Rammohan SV, Sept L (2011) Innovative logistics in extreme conditions: The case of healthcare delivery in Gambia. In: Bookbinder JH (ed) Global logistics. Springer, New York

Lester RT, Ritvo P, Mills EJ et al (2010) Effects of a mobile phone short message service on antiretroviral treatment adherence in Kenya (WelTel Kenya1): A randomised trial. Lancet 376(9755):1838–1845

Madon T, Hofman KJ, Kupfer L et al (2007) Implementation science. Science 318(5857): 1728–1729

McCoy JH (2008) Humanitarian response: Improving logistics to save lives. Am J Disaster Med 3(5):283–293

McCoy JH, Brandeau ML (2011) Efficient stockpiling and shipping policies for humanitarian relief: UNHCR's inventory challenge. OR Spectrum 33(3):673–698

McCoy JH, Johnson ME (2012) Clinic capacity management: Planning treatment programs that incorporate adherence. Prod Oper Manag (to appear)

Mills EJ, Nachega JB, Bangsberg DR et al (2006) Adherence to HAART: A systematic review of developed and developing nation patient-reported barriers and facilitators. PLoS Med 3(11):2039–2064

National Road Fund Agency (2005) Road maintenance. http://www.nrfa.org.zm/road_maintenance.htm. Cited 12 Jan 2011

Operations Evaluation Department (1996) Kingdom of Morocco impact evaluation report: Socioeconomic influence of rural roads. World Bank, Washington, DC

Organisation for Economic Co-operation and Development (2006) African economic outlook: Zambia. Available via African Economic Outlook. http://www.africaneconomicoutlook.org/. Cited 20 Jan 2011

Pedraza Martinez A, van Wassenhove LN (2012) Vehicle replacement in the International Committee of the Red Cross. Prod Oper Manag, DOI: 10.1111/j.1937-5956.2011.01316.x

Pedraza Martinez A, Hasija S, van Wassenhove LN (2011) An operational mechanism design for fleet management coordination in humanitarian operations. INSEAD Working Paper 2010/87/TOM/ISIC

Porter G (2002) Living in a walking world: Rural mobility and social equity issues in sub-Saharan Africa. World Dev 30(2):285–300

Rammohan SV (2010) Fueling growth. Stanf Soc Innov Rev (Summer 2010):68–71

Ranson MK, Hanson K, Oliveira-Cruz V et al (2003) Constraints to expanding access to health interventions: An empirical analysis and country typology. J Int Dev 15(1):15–39

Rao RC (1990) Compensating heterogeneous salesforces: Some explicit solutions. Mark Sci 9(4):319–341

Reilley B, Abeyasinghe R, Pakianathar MV (2002) Barriers to prompt and effective treatment of malaria in northern Sri Lanka. Trop Med Int Health 7(9):744–749

Riders for Health (2011) www.riders.org. Cited 20 Jan 2011

Rosero-Bixby L (2004) Spatial access to healthcare in Costa Rica and its equity: A GIS-based study. Soc Sci Med 58(7):1271–1284

Samai O, Sengeh P (1997) Facilitating emergency obstetric care through transportation and communication, Bo, Sierra Leone. Int J Gynecol Obstet 59(S2):S157–S164

Samii R, van Wassenhove LN, Kumar K et al (2002) Choreographer of disaster management: The Gujarat earthquake. INSEAD Case Study

Sanders D, Haines A (2006) Implementation research is needed to achieve international health goals. PloS Med 3(6):0719–0722

Schieber G, Maeda A (1999) Healthcare financing and delivery in developing countries. Health Aff 18(3):193–205

Shargie EB, Lindjørn B (2007) Determinants of treatment adherence among smear-positive pulmonary tuberculosis patients in southern Ethiopia. PLoS Med 4(2):280–287

Sherbrooke C (2004) Optimal inventory modeling of systems: Multi-echelon techniques, 2nd edn. Springer, New York

Spiliotopoulou E, Boni M, Yadav P (2011) Impact of treatment heterogeneity on drug resistance and supply chain costs. Working paper

Taylor T (2002) Supply chain coordination under channel rebates with sales effort effects. Manag Sci 48(8):992–1007

Transport Sector Board (2008) Safe, clean, and affordable: Transport for development. World Bank, Washington, DC

United Nations (2006) New UN health chief pledges to focus on Africans and women worldwide. Available via UN News Centre. http://www.un.org/apps/news/story.asp?NewsID=20556&Cr=WHO&Cr1=. Cited 26 Jan 2011

Weiser S, Wolfe W, Bangsberg DR et al (2003) Barriers to antiretroviral adherence for patients living with HIV infection and AIDS in Botswana. J AIDS 34(3):281–288

Whitten PS, Mair FS, Haycox A et al (2002) Systematic review of cost effectiveness studies of telemedicine interventions. BMJ 324(7351):1434–1437

World Bank (2006) Africa development indicators. World Bank, Washington, DC. http://siteresources.worldbank.org/INTSTATINAFR/Resources/ADI_2006_text.pdf. Cited 15 Jan 2011

World Bank (2010) Zambia's infrastructure: A continental perspective. Africa Infrastructure Country Report

World Health Organization (2011) http://www.who.int. Cited 4 Mar 2011

Yadav P, Stapleton O, van Wassenhove LN (2011) Always cola, rarely essential medicines: Comparing medicine and consumer product supply chains in the developing world. INSEAD working paper

Zambia Central Statistical Office (2010) Census of population and housing: Preliminary report. http://www.zamstats.gov.zm/. Cited 27 Feb 2011

Zambia Ministry of Health (2010) Numerous interviews of health officials and health workers

Zwarenstein M, Bheekie A, Lombard C et al (2007) Educational outreach to general practitioners reduces children's asthma symptoms: A cluster randomised controlled trial. Implement Sci 2:30

# Chapter 20
# Allocating Scarce Healthcare Resources in Developing Countries: A Case for Malaria Prevention

Jacqueline Griffin, Pinar Keskinocak, and Julie Swann

## 1 Introduction

One of the primary challenges in healthcare operations, whether for disease prevention or treatment, is the imbalance between the availability of resources and public health needs. As a result, allocation of scarce resources is among the most critical decisions within healthcare operations management. While financial constraints are prevalent in almost all healthcare settings, other limited resources include hospital beds, operating room time, nurses, and vaccines.

Decisions regarding the best use of scarce resources can become increasingly complex in the setting of a developing country due to greater disease incidence, poorer healthcare system infrastructure, and other societal factors. As a result, significant tradeoffs associated with resource allocation in developing countries provide opportunities for the use of analytical techniques. We demonstrate tradeoffs in the operation of a malaria prevention campaign with an optimization model and decision support tool. Both the model and tool include features that specifically address challenges of the developing world and the variety of resources that must be allocated in interrelated stages of decision making. The performance of different resource allocation heuristics is explored.

The remainder of this chapter is organized as follows. In Sect. 2 we present a summary of recent research concerning resource allocation decision making in developing countries. In Sect. 3, we present a case study of the development of a

J. Griffin (✉)
Northeastern University Boston, MA, USA
e-mail: ja.griffin@neu.edu

P. Keskinocak • J. Swann
Georgia Institute of Technology, Atlanta, GA, USA
e-mail: pinar@isye.gatech.edu; jswann@isye.gatech.edu

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184,
DOI 10.1007/978-1-4614-5885-2_20, © Springer Science+Business Media New York 2013

mathematical model and decision support tool for the allocation of scarce resources in a malaria prevention campaign. This includes a description of the system including the constraints balancing supply (Sect. 3.1.2) and demand (Sect. 3.1.1) and the resource allocation decisions (Sect. 3.1.3). The mathematical model is defined in Sect. 3.2. The use of the decision support tool is demonstrated through a numerical example in Sect. 3.3.

## 2 Literature Review

Resource allocation models, regardless of location or industry, examine decisions for optimizing a stated objective function provided a limited budget or other set of resources. With regard to improving health in developing countries, both the types of decisions to be made and the measurement of the objectives can vary. While most objectives relate to improving overall health, metrics can include disability-adjusted life years (DALYs) (Lasry et al. 2008), quality-adjusted life years (QALYs) (Zaric and Brandeau 2001), and the number of infections (Brandeau et al. 2003). Additionally, models specific to developing countries must consider unique characteristics such as road infrastructure (Balcik et al. 2008; Rahman and Smith 2000), spatial accessibility (Carr and Jallah 2008; Wilson and Blower 2005), nomadic populations (Ndiaye and Alfares 2008), and rainy seasons (Oppong 1996).

In addition to the variety of performance metrics utilized, resource allocation models can consider a variety of types of decisions including the funding of different health programs (Epstein et al. 2006). Flessa (2000) develops a linear program to study allocation of curative and preventative treatments for a variety of diseases in developing countries (Flessa 2000). Often, the funding of different programs is determined by calculating the cost-effectiveness of each program to identify appropriate prioritization (Hansen and Chapman 2008). While cost-effectiveness measurements are useful, when prioritizing investment in programs which address the same health condition, the interactions between different programs should be incorporated into resource allocation models, often requiring nonlinear components (Alistar and Brandeau 2012).

Rather than addressing prioritization among health programs, some models address decisions for the prioritization of patient populations (Brandeau et al. 2003) which can differ by a variety of factors including health condition (Lee et al. 2010) and location (Balcik et al. 2008). Wilson and Kahn (2006) examine the epidemiological impact of drug allocation strategies on urban and rural populations through the development of a spatially explicit model.

Budget allocation decisions are further complicated by the sequential distribution of funds among different stages (e.g., regional and local levels) within a healthcare delivery system. For example, Lasry et al. (2007) examine the distribution of funds for HIV prevention at multiple levels, by comparing the use of heuristic and optimal policies, and the interactions between decisions at the different stages (Lasry et al. 2007). Similarly, Zaric and Brandeau (2007) use an optimization model to examine budget allocation at two levels to address the difference between proportional and efficient distribution policies in HIV prevention.

When modeling resource allocation for the treatment or prevention of infectious diseases, often the epidemiology, or interaction of the disease and the population, is explicitly modeled. One approach is the use of compartmental models (Zaric and Brandeau 2001). The use of epidemic models addresses the importance of timing in resource allocation and the dynamic nature of disease prevalence. Through use of optimization and epidemic models, Brandeau et al. (2003) examine prioritization of independent populations for resource allocation within a limited time horizon. Zaric and Brandeau (2002) find that heuristic methods for allocation perform well and that reallocation of resources throughout a limited time horizon produces superior results to making a single allocation in the first period.

While there is interest in developing analytical models to aid in resource allocation, often simple heuristics and gut instincts are employed in practice. Researchers have examined the loss of efficiency or effectiveness with the use of these heuristics. For example, Deo and Sohoni (2011) compare the use of simple heuristics and optimal resource allocation for locating diagnostic testing equipment in developing countries.

While many complex models have been developed for addressing public health concerns in developing countries, few are developed specifically for implementation. Lasry et al. (2008) develop a resource allocation model for HIV prevention. The model is implemented in a spreadsheet-based decision support tool, S4HARA. The tool deconstructs decision making into four interrelated stages including prioritization of HIV prevention methods and allocation of funds (Lasry et al. 2008).

In Sect. 3, we present a case study of a decision support tool for allocating resources in a malaria prevention program across time and locations. Similar to previous models, this model considers the disease dynamics and decisions for population prioritization. Additionally, the model includes characteristics that are unique to developing countries including a focus on road infrastructure quality. Similar to Lasry et al. (2007) and Zaric and Brandeau (2007), the model can be deconstructed into multiple levels of decision making and the relationship between decisions at different stages is examined. Similar to Lasry et al. (2008), a spreadsheet-based decision support tool incorporating these stages is developed using the model.

## 3 Case Study: Indoor Residual Spraying Operations Management

Malaria is a vector-borne illness transmitted by infected *Anopheles* mosquitoes. Malaria poses a serious global public health threat, but is especially prevalent and has its most significant impact in the developing world. Approximately 500 million clinical cases of malaria occur each year, resulting in approximately one million deaths worldwide (Thomson et al. 2006). While almost half of the world's population lives in an area at risk of malaria transmission, 89% of deaths occur in Africa, resulting in the second largest number of deaths from an infectious disease after HIV/AIDS in Africa (Centers for Disease Control and Prevention Division of Parasitic Disease 2010).

Despite being a preventable and curable disease, there is a continued prevalence of malaria in Africa due in part to increasing drug resistance towards inexpensive and readily available medications (Cox et al. 1999) and severe underfunding by the international community for achieving malaria control goals (Kiszewski et al. 2007; Snow et al. 2008). Also, many pharmaceuticals for malaria treatment are prohibitively expensive for widespread use in developing countries. Due to the state of healthcare infrastructure, the logistics of diagnosing and treating infected individuals can be challenging. Hence, increased focus has been given to prevention methods.

The most prevalent malaria prevention methods, long lasting insecticide-treated nets (LLINs) and indoor residual spraying (IRS), both work by reducing the rate of mosquito bites, which is the method of transmission for malaria. LLIN programs provide bed nets for individuals to drape over their beds while sleeping. In addition to providing a physical barrier, the insecticide repels and kills mosquitoes that come in contact with the net. With the IRS method, trained spray teams apply insecticide to the walls and ceilings of residences. Both LLIN and IRS require retreatment at regular intervals. While both methods have been shown to be effective, Ministries of Health choose a prevention method based on the individual circumstances of their country (Over et al. 2004). In addition to government agencies, some non-governmental organizations (NGOs) participate in the treatment and distribution of bed nets. For example, the United Nations Foundation has created a *Nothing But Nets* campaign to distribute bed nets across Africa (United Nations Foundation 2012).

Regardless of the choice of prevention program, operational decisions for a malaria prevention campaign include (1) which areas to target for prevention, (2) when to schedule spraying or net distribution given the seasonal nature of the disease, and (3) how to allocate resources such as insecticide, labor, and trucks. Often, these decisions are made by members of the Ministry of Health or NGOs based on personal experience and gut instincts, rather than analytical methods that capture the true complexities of the system. Because of the complexities, these practices can result in poor allocation of resources with impaired effectiveness due to undercoverage, overcoverage, or poor timing.

In Sect. 3.1, we describe the development of an operations research model for an IRS program and present its formulation in Sect. 3.2. In Sect. 3.3, we discuss how this model is adapted into a decision support tool. With a small numerical example, we demonstrate the value of the tool in evaluating complexities and tradeoffs that exist in the allocation of limited resources for an IRS program.

## *3.1 IRS Problem Definition*

One goal for an IRS campaign is to minimize the number of contracted cases of malaria. The number of cases of malaria prevented is a consequence of both the deployment of limited resources for prevention and the population at risk of contracting malaria, or the demand. As a result, decisions in the design of a malaria prevention program must address both supply and demand. The model presented below addresses strategic, tactical, and operational decisions in the design and operation of an IRS program.

### 3.1.1 Demand

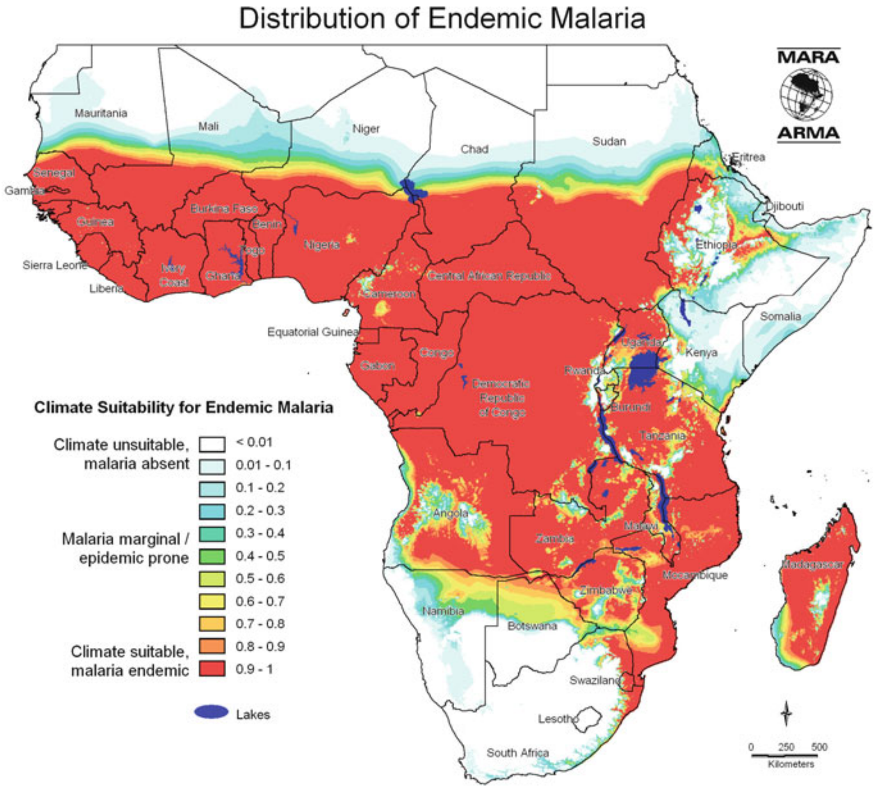
Demand for malaria prevention is a function of several factors including population size, location, and timing. To represent this demand, the total area being considered for treatment is divided into zones. For each zone, the total population and malaria risk factors are considered.

The rate of transmission of malaria exhibits a seasonal pattern, consistent with mosquito population trends. In addition to time of the year, mosquito prevalence is also affected by environmental factors (i.e., annual rainfall and temperature) (Craig et al. 1999). Due to these characteristics, a model for malaria prevention should account for dynamic decision making, incorporating changes in risk by time and location.

To model the disease dynamics, forecasts for malaria risk factors can be estimated with data compiled by the Mapping Malaria Risk in Africa (MARA) initiative a group of scientists that have processed and summarized data pertaining to geographic representations of the malaria burden in Africa (www.mara.org.za) (Mapping Malaria Risk in Africa 2004). MARA considers environmental factors such as rainfall and temperature in forecasting annual malaria risk on a 0–1 scale, with 1 as the highest risk. This risk factor is calculated for 25 km$^2$ zones in Africa. Additionally, data pertaining to the first and last month of the high transmission season further classifies the time-dependent nature of malaria risk. This data on the overall risk factor and the timing of the high transmission season is combined to calculate a risk factor for each time period and zone considered by the model. A map developed by MARA depicting the endemic risk of malaria for the African continent is shown in Fig. 20.1.

### 3.1.2 Supply

The supply of resources in an IRS campaign includes distribution centers, teams of trained sprayers, spray equipment, insecticide, and transportation vehicles. Each team of trained sprayers is employed at a distribution center (DC). During spraying operations, a team utilizes spray equipment as well as a truck for transport to and from the assigned zone. We assume that each team only visits one zone per time



This map is a product of the MARA/ARMA collaboration (http://www.mara.org.za). July 2002, Medical Research Council, PO Box 70380, Overport, 4067, Durban, South Africa
CORE FUNDERS of MARA/ARMA: International Development Research Centre, Canada (IDRC); The Wellcome Trust UK; South African Medical Research Council (MRC);
Swiss Tropical Institute, Multilateral Initiative on Malaria (MIM) / Special Programme for Research & Training in Tropical Diseases (TDR), Roll Back Malaria (RBM).
Malaria distribution model: Craig, M.H. et al. 1999. Parasitology Today 15: 105-111.
Topographical data: African Data Sampler, WRI, http://www.igc.org/wri/sdis/maps/ads/ads_idx.htm.

**Fig. 20.1** A map of endemic risk of malaria in Africa based on climatic factors including rainfall and temperature as calculated by Mapping Malaria Risk in Africa's (MARA) (www.mara.org.za)

period and returns to the distribution center at the end of the period. While trucks and spray equipment can be used in subsequent time periods, the inventory of insecticide is depleted as spraying occurs.

Quantities of these resources are limited by the supply of funds or the total budget. In the model, we assume there is a fixed cost for opening a distribution center and a variable cost for each time period in which the distribution center operates. There is a unit cost for training sprayers and renting equipment and trucks. Additionally, there is a unit cost for the insecticide needed to treat each home. Lastly, there is a transportation cost between distribution centers and zones for which the budget is allocated.

Unlike developed countries, road infrastructure is less reliable in developing countries, and as a result, the time to travel a similar distance may take much longer on roads in undeveloped areas. Additionally, it is possible that a road may be impossible to cross during certain parts of the year (i.e., the rainy season). As a result, it is unrealistic to treat all roads as equivalent or to assume a constant multiplicative factor for cost per mile in the model. Instead, a road quality-adjusted transportation cost is calculated accounting for the classification of the roads in the associated zone. These costs can be estimated using data from the World Health Organization's HealthMapper, a public health mapping application (World Health Organization 2012). The estimated transportation cost between a distribution center and a zone incorporates the quality of the roads. By considering the road quality, it is possible for the best location of distribution centers or assignment of zones to differ considerably as compared to a model that only considers road distance.

### 3.1.3 Model Decisions

Decisions for matching supply and demand to maximize the number of prevented cases of malaria must address the quantity and timing of resource allocation and system constraints. For instance, insecticide treatment continues to be active for a limited number of time periods following application. Capacity restrictions for equipment and insecticide at the distribution centers limit the number of teams deployed. There is also a limit on the number of homes that can be treated by one spray team in one time period. To prevent duplicate spraying, each zone is assigned to exactly one open distribution center. Since it is assumed that a team can only visit one zone per period, routing is not considered.

With the objective of maximizing the number of cases of malaria prevented, the following decisions are made for IRS program operations:

- How many distribution centers (DC) should be opened?
- Where should the distribution centers be located?
- When should the distribution centers be operational?
- How should zones be allocated to the (open) distribution centers?
- How many teams (i.e., employees, spray equipment, vehicles) should be located at each distribution center?
- When should teams be deployed to each zone?
- How many homes should be treated from each zone in each time period?
- How much insecticide (DDT) should be purchased at each distribution center?

While all of these decisions are interrelated, the decisions can be classified into three stages of decision making:

1. Location of distribution centers [strategic]
2. Assignment of zones to distribution centers [tactical]
3. Scheduling team deployment [operational]

**Table 20.1** Decision variables for team deployment MIP

| Decision variable | Definition | Type |
|---|---|---|
| $Open_j$ | 1 if DC $j$ is open, 0 otherwise | Binary |
| $Ldc_j$ | Number of spray teams to employ at DC $j$ | Integer |
| $P_{j,t}$ | 1 if DC $j$ is open at time $t$, 0 otherwise | Binary |
| $zoneServed_{i,j}$ | 1 if Zone $i$ is assigned to DC $j$, 0 otherwise | Binary |
| $L_{i,t,j}$ | Number of teams of workers allocated to Zone $i$ at time $t$ from DC $j$ | Integer |
| $Cnew_{i,t,j}$ | Number of newly protected people in Zone $i$ at time $t$ by teams deployed from DC $j$ | Integer |
| $Enew$ | Quantity of new equipment (spray tanks) to purchase | Integer |

## *3.2 Team Deployment Model*

Creating one mixed integer programming (MIP) model to incorporate all three stages of decisions would require approximately 700,000 binary variables for an instance representing Swaziland, one of the smallest countries in Africa, with approximately 800 million variables in total. By deconstructing the problem into three stages of decision making, the size of the MIP for the scheduling team deployment stage becomes more manageable, but possibly still difficult to solve to optimality for large regions.

The formulation of the MIP for scheduling team deployment including the model decision variables (Table 20.1), parameters (Table 20.2), constraints ((20.2)–(20.19)), and objective function (20.1) is provided below. This MIP can be extended to incorporate decisions about the location of distribution centers, including the number to open, and assignments of zones. These decisions are incorporated into the decision support tool discussed in Sect. 3.3.

The objective of this MIP (20.1) is to maximize the number of prevented malaria cases or the effective coverage. The total effective coverage is calculated as the sum of the product of the number of people treated and the risk in all periods in which the insecticide is active or the subsequent *activeT* periods after spraying occurs. Constraints (20.2) and (20.3) allow for each zone to be served by exactly one open distribution center. Constraints (20.4), (20.5), and (20.17) limit the number of people covered by spraying to be less than the total population and the amount defined by the assignment of teams. Constraints (20.6)–(20.10) restrict assignment of spray teams according to capacity and supplies allocated at each distribution center. Constraints (20.11)–(20.16) limit deployment to the periods of time in which a distribution center is open, which is all periods between the first and last deployment from that distribution center. Constraint (20.18) restricts the total cost of the program to be less than the available budget.

**Table 20.2** Parameters for team deployment MIP

| | Parameter | Definition | Units |
|---|---|---|---|
| Sets | $J$ | Set of DC locations | |
| | $I$ | Set of zones | |
| | $T$ | Set of time periods $\{1,\ldots,NT\}$ | |
| Zones | $population_i$ | Population in zone $i$ | People |
| | $risk_{i,t}$ | Percent of population at risk in zone $i$ at time $t$ | % risk |
| | $ph$ | Estimated number of people living in each home | People/home |
| | $dist_{i,j}$ | Road quality-adjusted straight line distance from zone $i$ to DC $j$ | km |
| | $cPerKm$ | Estimated transportation cost | \$/km |
| DCs | $capDCeq_j$ | Capacity of DC $j$ for equipment | Spray tanks |
| | $capDCtr_j$ | Capacity of DC $j$ for trucks | Trucks |
| | $capDCddt_j$ | Capacity of DC $j$ for DDT | g DDT |
| | $cOpenDC_j$ | Fixed cost to open DC $j$ | \$ |
| | $cRentDC_j$ | Cost to keep DC $j$ open per unit time | \$/time |
| | $tpdc$ | Time to prepare and close a DC | Time |
| Spray teams | M | Upper bound on the number of spray teams, $budget/cDDT$ | Teams |
| | $tt$ | Time required to train spray personnel | Time |
| | $kt$ | Workers per team (based on the capacity of a truck to hold the workers and their equipment) | Sprayers/truck |
| | $rs$ | Work rate of the spray personnel | Homes/time |
| | $ws$ | Wage of spray personnel | \$/time |
| | $cMiscEq$ | Cost of safety gear and equipment per worker | \$/sprayers |
| | $ctr$ | Cost of renting trucks per unit time | \$/time/truck |
| | $Einit$ | Initial amount of equipment available | Spray tanks |
| | $pCover$ | People that can be covered by one team, $kt \cdot rs \cdot ph$ | People |
| | $cEquip$ | Cost of tanks, safety gear, and equipment per team, $kt \cdot (chTank + cMiscEq)$ | \$/team |
| | $cHire$ | Cost to hire one team, $kt \cdot tt \cdot ws$ | \$/team |
| | $cWage$ | Wage per team per time period, $kt \cdot ws$ | \$/time/team |
| Insecticide (DDT) | $activeT$ | Time periods in which DDT is active after spraying | Time |
| | $sh$ | Amount of DDT needed per home | g/home |
| | $cgDDT$ | Total cost per gram of DDT | \$/g DDT |
| | $cue$ | Unit cost of new spray tanks | \$/spray tank |
| | $chTank$ | Total cost per tank for repair and replacement | \$/spray tank |
| | $ddtperP$ | Quantity of DDT needed to protect one person, $sh/ph$ | g DDT |
| | $ddtCover$ | Quantity of DDT needed for one team working at full capacity, $kt \cdot rs \cdot sh$ | g DDT |
| | $cDDT$ | Cost of DDT needed for one team working at full capacity, $cgDDT \cdot kt \cdot rs \cdot sh$ | \$/team |
| Financial Resources | $budget$ | Maximum monetary funds available for deployment | \$ |

$$max \qquad \sum_{i\in I,t\in T,j\in J}\ \sum_{s\in\{0,\dots,activeT:t+s\le T\}} Cnew_{i,t,j}\ risk_{i,t+s} \tag{20.1}$$

$$s.t.$$

$$\sum_{i\in I,j\in J} zoneServed_{i,j} = 1 \qquad \forall i \in I \tag{20.2}$$

$$zoneServed_{i,j} \le Open_j \qquad \forall i \in I, j \in J \tag{20.3}$$

$$\sum_{j\in J,t\in T} Cnew_{i,t,j} \le population_i \qquad \forall i \in I \tag{20.4}$$

$$\sum_{j\in J} Cnew_{i,t,j} \le pCover \sum_{j\in J} L_{i,t,j} \qquad \forall i \in I, t \in T \tag{20.5}$$

$$\sum_{i\in I} L_{i,t,j} \le Ldc_j \qquad \forall j \in J, t \in T \tag{20.6}$$

$$kt\ Ldc_j \le capDCeq_j \qquad \forall j \in J \tag{20.7}$$

$$Ldc_j \le capDCtr_j \qquad \forall j \in J \tag{20.8}$$

$$\sum_{j\in J} kt\ Ldc_j \le Einit + Enew \tag{20.9}$$

$$\sum_{i\in I,t\in T} ddtperP\ Cnew_{i,t,j} \le capDCddt_j \qquad \forall j \in J \tag{20.10}$$

$$Ldc_j \le M\ Open_j \qquad \forall j \in J \tag{20.11}$$

$$L_{i,t,j} \le M\ zoneServed_{i,j} \qquad \forall i \in I, t \in T, j \in J \tag{20.12}$$

$$P_{j,t} \le \sum_{i\in I} L_{i,t,j} \qquad \forall j \in J, t \in T \tag{20.13}$$

$$\sum_{i\in I} L_{i,t,j} \le M\ P_{j,t} \qquad \forall j \in J, t \in T \tag{20.14}$$

$$\sum_{l\in\{1,\dots,t-activeT\}} P_{j,l} \le M\ (P_{j,t-1} - P_{j,t} + 1) \qquad \forall j \in J, t \in \{activeT, \dots, NT\} \tag{20.15}$$

$$\sum_{l\in\{t+1,\dots,NT\}} P_{j,l} \le M\ (P_{j,t} - P_{j,t-1} + 1) \qquad \forall j \in J, t \in \{activeT, \dots, NT\} \tag{20.16}$$

$$\sum_{t\in T,j\in J} L_{i,t,j} \le \left\lceil \frac{population_i}{pCover} \right\rceil \qquad \forall i \in I \tag{20.17}$$

$$\begin{aligned} Enew\,cue &+ \sum_{j\in J}(cEquip + cHire)\,Ldc_j \\ &+ \sum_{i\in I,t\in T,j\in J}(2\,cPerKm\,dist_{i,j})\,L_{i,t,j} \\ &+ \sum_{i\in I,t\in T,j\in J}(ctr + cDDT + cWage)\,L_{i,t,j} \\ &+ \sum_{j\in J,t\in T} cRentDC_j\,P_{j,t} \\ &+ \sum_{j\in J}(tpdc + tt)\,Open_j + \sum_{j\in J} cOpenDC_j\,Open_j \le budget \end{aligned} \tag{20.18}$$

$$Cnew_{i,t,j} \ge 0 \qquad \forall i \in I, j \in J, t \in T \tag{20.19}$$

$$L_{i,t,j}, Ldc_j, Enew \in \{0,1,2,\ldots\} \qquad \forall i \in I, j \in J, t \in T \tag{20.20}$$

$$zoneServed_{i,j}, Open_j, P_{j,t} \in \{0,1\} \qquad \forall i \in I, j \in J, t \in T \tag{20.21}$$

## 3.3 Decision Support Tool

We created a spreadsheet-based decision support tool that incorporates the framework and structure of the MIP. Through use of the tool, decision makers can evaluate the impact of proposed strategies (i.e., conduct what-if analyses) and identify the nature of tradeoffs in resource allocation in the process of arriving at a final decision. Note that a decision maker in this case can include an actual user with an NGO or health agency, or a student in a classroom exploring alternatives and gaining insights about resource allocation.

We demonstrate the use of the decision tool with a small numerical example, defined in Sect. 3.3.1. In Sect. 3.3.2, for this instance, we present an example of the process a decision maker may go through, testing and evaluating instinct-based heuristics with use of this analytical tool, when the assignment of zones to distribution centers is fixed. We demonstrate how the use of the tool can highlight key tradeoffs and improve understanding of the deployment problem. In Sect. 3.3.3, we further demonstrate the impact of additional flexibility of decision making, including zone assignment decisions.

### 3.3.1 A Numerical Example

The decision support tool is implemented in Excel on a sample instance. This instance has five zones, four distribution centers, and three time periods. A similar decision support tool could be applied to a larger instance. A map of the location of the zones and distribution centers is shown in Fig. 20.2. Additionally, the population

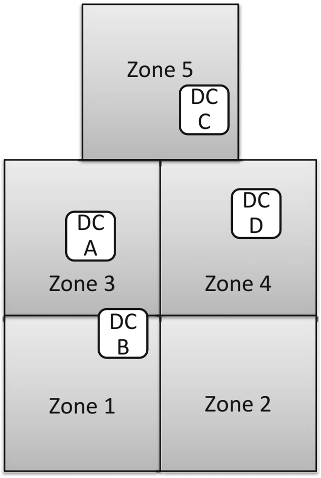


**Fig. 20.2** Map of the location of four distribution centers (A, B, C, D) and five zones for the numerical example

**Table 20.3** Population and time-dependent risk factor by zone

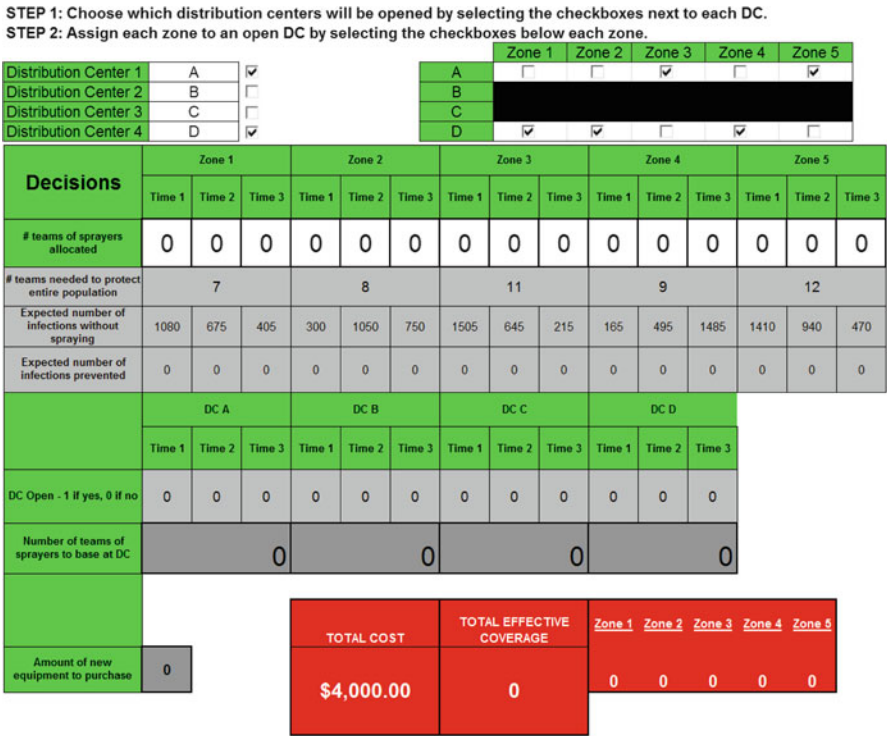
| Zone | Population | Risk period 1 | Risk period 2 | Risk period 3 |
|---|---|---|---|---|
| ($i$) | ($population_i$) | ($risk_{i,1}$) | ($risk_{i,2}$) | ($risk_{i,3}$) |
| 1 | 1,350 | 0.8 | 0.6 | 0.3 |
| 2 | 1,500 | 0.2 | 0.7 | 0.5 |
| 3 | 2,150 | 0.7 | 0.3 | 0.1 |
| 4 | 1,650 | 0.1 | 0.3 | 0.9 |
| 5 | 2,350 | 0.6 | 0.4 | 0.2 |

**Table 20.4** Quality-adjusted transportation cost between distribution centers and zones

| DC | Zone 1 | Zone 2 | Zone 3 | Zone 4 | Zone 5 |
|---|---|---|---|---|---|
| ($j$) | ($dist_{1,j}$) | ($dist_{2,j}$) | ($dist_{3,j}$) | ($dist_{4,j}$) | ($dist_{5,j}$) |
| A | 23.14 | 25.12 | 6 | 20.9 | 23.25 |
| B | 8.94 | 15.65 | 14.23 | 25 | 35.37 |
| C | 42.25 | 39.37 | 24.67 | 20.95 | 6.4 |
| D | 40.7 | 25.16 | 33.11 | 6.71 | 28.5 |

and risk factors for each time period are presented in Table 20.3. These risk factors account for both the overall risk as well as the disease dynamics. The quality-adjusted transportation costs between distribution centers and zones are shown in Table 20.4. These data sets have been defined to provide an instructive example.

In this numerical example the equipment, truck, and insecticide capacities are identical for all distribution centers, as well as the cost of opening and the variable operational costs per time period. The model assumes that spraying is effective in the time period in which the spraying occurs and all subsequent periods of the model. Reapplication is not needed within the time horizon considered here ($activeT = 2$), e.g., a 1 year period.



**STEP 1:** Choose which distribution centers will be opened by selecting the checkboxes next to each DC.
**STEP 2:** Assign each zone to an open DC by selecting the checkboxes below each zone.

| | | |
|---|---|---|
| Distribution Center 1 | A | ☑ |
| Distribution Center 2 | B | □ |
| Distribution Center 3 | C | □ |
| Distribution Center 4 | D | ☑ |

| | Zone 1 | Zone 2 | Zone 3 | Zone 4 | Zone 5 |
|---|---|---|---|---|---|
| A | □ | □ | ☑ | □ | ☑ |
| B | | | | | |
| C | | | | | |
| D | ☑ | ☑ | □ | ☑ | □ |

| Decisions | Zone 1 | | | Zone 2 | | | Zone 3 | | | Zone 4 | | | Zone 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Time 1 | Time 2 | Time 3 | Time 1 | Time 2 | Time 3 | Time 1 | Time 2 | Time 3 | Time 1 | Time 2 | Time 3 | Time 1 | Time 2 | Time 3 |
| # teams of sprayers allocated | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| # teams needed to protect entire population | 7 | | | 8 | | | 11 | | | 9 | | | 12 | | |
| Expected number of infections without spraying | 1080 | 675 | 405 | 300 | 1050 | 750 | 1505 | 645 | 215 | 165 | 495 | 1485 | 1410 | 940 | 470 |
| Expected number of infections prevented | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| | DC A | | | DC B | | | DC C | | | DC D | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Time 1 | Time 2 | Time 3 | Time 1 | Time 2 | Time 3 | Time 1 | Time 2 | Time 3 | Time 1 | Time 2 | Time 3 |
| DC Open - 1 if yes, 0 if no | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Number of teams of sprayers to base at DC | 0 | | | 0 | | | 0 | | | 0 | | |

| | |
|---|---|
| Amount of new equipment to purchase | 0 |

| TOTAL COST | TOTAL EFFECTIVE COVERAGE | Zone 1 | Zone 2 | Zone 3 | Zone 4 | Zone 5 |
|---|---|---|---|---|---|---|
| $4,000.00 | 0 | 0 | 0 | 0 | 0 | 0 |

Fig. 20.3 A view of the IRS decision support tool which calculates the total cost and total effective coverage based on the team deployment decisions entered into the spreadsheet

The spreadsheet-based decision support tool reflects the three stages of decision making defined above. First, the spreadsheet user inputs a decision concerning which distribution centers will be opened and where they will be located. Second, based on which distribution centers are open, the user makes a decision regarding the allocation of zones to distribution centers. Finally, the tool requires the user to schedule the number of teams allocated to each zone in each time period. Using the MIP model framework, the spreadsheet automatically updates defining which time periods distribution centers are open, the equipment purchased and teams trained at each distribution center, and the number of malaria cases prevented. With this design of the decision support tool, the total cost and number of prevented cases is quickly calculated based on user input, and a feasibility check is completed for all constraints. The decision support tool is constructed to use basic Excel functionality, with no macros or VBA components. A view of the decision support tool interface is shown in Fig. 20.3.

### 3.3.2 Evaluating Resource Deployment Strategies

In this section, we go into more detail on one stage of decision making demonstrating how the tool can be used to better understand the problem and to evaluate the performance of heuristic strategies. We examine the use of the tool in the team scheduling stage, in which the location of the distribution centers and the assignment of zones are fixed. In this instance, we assume a fixed budget of $50,000 and the assignment of Zones 1, 2, and 4 to DC D and Zones 3 and 5 to DC A.

In practice, strategies based on simple heuristics may be considered for scheduling team deployment. Considerations in development of strategies may include risk, transportation costs, population, and balanced allocation of resources.

Utilizing the tool to evaluate the performance of schedules based on instinctual strategies, we demonstrate how a decision maker can gain an understanding of key system behaviors and utilize this understanding to develop better strategies. With a risk-based approach, one goal may be to spray in zones with the highest overall malaria risk in the earliest time periods. This would cause the prioritization of spraying in Zone, 1 (total risk = 1.7), then Zone 2 (total risk = 1.4), and finally in Zone 4 (total risk = 1.3). Due to the fixed assignment of zones, all three of these zones are served by DC D. Possible trials of team deployment that may be considered are shown in Table 20.5. Because the number of teams deployed must be integer, the total cost for each trial is slightly less than the total budget.

In the first trial, Zone 1 is prioritized and receives enough teams to protect the total population, and additional teams are allocated to Zone 2 using the remaining budget. If instead, Zone 2 is prioritized over Zones 1 and 4, with a similar allocation heuristic, more malaria cases are prevented, with an effective coverage of 3,671 (Trial 2). This improvement is a result of the less costly transportation costs to Zones 2 and 4 rather than Zone 1. Therefore, although Zone 1 has a higher total risk, prioritizing Zone 2 leads to greater prevention overall. Further considering quality-adjusted transportation costs in addition to risk, 5,558 cases can be prevented by prioritizing Zone 2 over Zone 4 and not spraying in Zone 1.

While spraying only in the first time period may ensure the highest effectiveness per spray team deployed, it may not be the most cost-effective strategy. For example, due to the preventative nature of IRS, ideally all homes would be sprayed in the first period, to prevent infection in subsequent periods. This requires many teams to be hired in the first period, resulting in high costs for purchasing equipment and training. Additionally, due to disease dynamics, earlier spraying is more valuable in some zones compared to others. The decision support tool can be used to examine the impact of deploying teams over multiple periods.

Table 20.6 displays allocation strategies that improve on the trials from Table 20.5 with similar prioritization of risk and transportation costs, but instead allocate teams over multiple periods. Comparing Trials 1 and 4, by delaying spraying in Zone 2 until the second period, more homes receive treatment and more cases are prevented. Additionally, by utilizing teams across two periods, the total number of teams employed decreases from 10 to 7, thereby allocating less of the budget to hiring and training teams and more to spraying in Zone 4.

**Table 20.5** Trial schedules using risk-based strategies

| | Zone 1 ($\sum_j L_{1,t,j}$) | | | Zone 2 ($\sum_j L_{2,t,j}$) | | | Zone 3 ($\sum_j L_{3,t,j}$) | | | Zone 4 ($\sum_j L_{4,t,j}$) | | | Zone 5 ($\sum_j L_{5,t,j}$) | | | Trucks @ | Trucks @ | Cases | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Period ($t$) | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | A ($Ldc_A$) | D ($Ldc_D$) | prevented (obj.) | cost($) |
| Trial 1 | 7 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 3,152 | 49,596 |
| Trial 2 | 3 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 13 | 3,671 | 49,955 |
| Trial 3 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 4,245 | 48,116 |

**Table 20.6** Trial schedules using multi-period strategies

| | Zone 1 ($\sum_j L_{1,t,j}$) | | | Zone 2 ($\sum_j L_{2,t,j}$) | | | Zone 3 ($\sum_j L_{3,t,j}$) | | | Zone 4 ($\sum_j L_{4,t,j}$) | | | Zone 5 ($\sum_j L_{5,t,j}$) | | | Trucks @ | Trucks @ | Cases | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Period ($t$) | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | A ($Ldc_A$) | D ($Ldc_D$) | prevented (obj.) | cost($) |
| Trial 4 | 7 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 7 | 3,519 | 48,528 |
| Trial 5 | 4 | 0 | 0 | 4 | 3 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 3,795 | 49,775 |
| Trial 6 | 1 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 6 | 3 | 0 | 0 | 0 | 0 | 0 | 15 | 4,549 | 49,309 |
| Trial 7 | 2 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 1 | 7 | 0 | 0 | 0 | 0 | 0 | 11 | 4,773 | 49,731 |

Similarly, improvements over Trial 2 can be achieved through scheduling team deployment to Zone 2 among multiple time periods (Trial 5). Improvements over Trial 3, through rebalancing of team deployments to minimize the number of teams hired, are demonstrated in Trials 6 and 7.

While the zones with the highest overall risk are served by teams from DC D, the zones with the highest population, Zones 3 and 5, are assigned to DC A. The decision support tool can be used to examine the impact of deploying teams from both distribution centers and accounting for population by zone and transportation costs. Table 20.7 includes strategies that utilize both distribution centers. In Trial 8, deployment is prioritized with respect to the expected number of malaria cases in each zone. By reducing the number of teams deployed from DC D, 5,712 malaria cases are prevented in Trial 9. This is an 81% increase in total effective coverage over the starting strategy in Trial 1, demonstrating the valuable information a decision maker may gain from use of the decision support tool.

### 3.3.3 Evaluating Distribution Center and Zone Assignment Strategies

As demonstrated in Sect. 3.3.2, the decision support tool can be used to highlight key features in the development of an effective resource deployment strategy including transportation costs, malaria risk levels, population, and disease dynamics. In this section, we demonstrate the value of the tool in the tactical decision making stage in which decisions for the best allocation of zones to DCs A and D are made.

The initial assignment of zones (A: 3,5; D: 1,2,4) balances the total population covered by each distribution center. Under this assignment, the maximum number of malaria cases prevented, with a budget of $50,000, is 5,938, only a 4% improvement over Trial 9. In the optimal strategy with fixed locations and assignments (Table 20.8), all resources are deployed to Zones 2, 3, and 4, the zones with the smallest transportation costs. Due to the high population and risk in Zone 3, teams are deployed in the first time period. Deployment from DC D balances the number of teams used in the first and second periods. This demonstrates the usefulness of the tool in identifying the key system tradeoffs.

Heuristics for assigning zones to distribution centers may be similar to those considered when making deployment decisions. One strategy may be to balance the total risk among zones assigned to each distribution center. With this goal, Zones 2 and 4 would be served by one distribution center, and Zones 1, 3, and 5 by the other distribution center. For the two possible assignments of zones to DC A and DC D using these groupings, the optimal schedules and the number of prevented cases are shown in Table 20.9.

Another strategy a decision maker might select is to balance the total expected number of infections in zones for each distribution center. Under this strategy, Zones 1, 2, and 3 would be served by one distribution center, and Zones 4 and 5 served by the other. For the two possible assignments to DC A and DC D using these groupings, the optimal schedules and the number of prevented cases are shown

**Table 20.7** Trial schedules with deployment from both DCs

| | Zone 1 | | | Zone 2 | | | Zone 3 | | | Zone 4 | | | Zone 5 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $(\sum_j L_{1,t,j})$ | | | $(\sum_j L_{2,t,j})$ | | | $(\sum_j L_{3,t,j})$ | | | $(\sum_j L_{4,t,j})$ | | | $(\sum_j L_{5,t,j})$ | | | Trucks @ | Trucks @ | Cases | Total |
| Period ($t$) | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | A ($Ldc_A$) | D ($Ldc_D$) | prevented (obj.) | cost($) |
| Trial 8 | 0 | 0 | 0 | 0 | 4 | 0 | 11 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 11 | 9 | 5,489 | 49,744 |
| Trial 9 | 0 | 0 | 0 | 0 | 5 | 0 | 11 | 0 | 0 | 7 | 2 | 0 | 0 | 0 | 0 | 11 | 7 | 5,712 | 49,823 |

**Table 20.8** Optimal schedule with DC A serving Zones 3 and 5 and DC D serving Zones 1, 2, and 4

| | Zone 1 | | | Zone 2 | | | Zone 3 | | | Zone 4 | | | Zone 5 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $(\sum_j L_{1,t,j})$ | | | $(\sum_j L_{2,t,j})$ | | | $(\sum_j L_{3,t,j})$ | | | $(\sum_j L_{4,t,j})$ | | | $(\sum_j L_{5,t,j})$ | | | | | | Total |
| | | | | | | | | | | | | | | | | Trucks @ | Trucks @ | Cases | cost |
| Period ($t$) | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | A ($Ldc_A$) | D ($Ldc_D$) | prevented (obj.) | ($) |
| Optimal | 0 | 0 | 0 | 6 | 0 | 0 | 10 | 0 | 0 | 1 | 7 | 0 | 0 | 0 | 0 | 10 | 7 | 5,938 | 49,607 |

**Table 20.9** Balanced total risk covered by DC A and DC D with associated optimal schedules

| | | | | Zone 1 | | | Zone 2 | | | Zone 3 | | | Zone 4 | | | Zone 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Cases prevented | $(\sum_j L_{1,t,j})$ | | | $(\sum_j L_{2,t,j})$ | | | $(\sum_j L_{3,t,j})$ | | | $(\sum_j L_{4,t,j})$ | | | $(\sum_j L_{5,t,j})$ | | |
| Period ($t$) | DC A | DC D | (optimal) | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| Assignment 1 | 2, 4 | 1, 3, 5 | 4,040 | 1 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 1 | 7 | 0 | 0 | 0 | 0 |
| Assignment 2 | 1, 3, 5 | 2, 4 | 6,063 | 5 | 0 | 0 | 0 | 0 | 0 | 10 | 1 | 0 | 4 | 4 | 0 | 0 | 0 | 0 |

in Table 20.10. As shown here, even when a reasonable heuristic is developed for grouping zones, the decision support tool can aid in making the appropriate assignments (3,959 vs. 6,101 prevented cases). Assuming that DCs A and D are open, Assignment 4 is an optimal strategy for both allocation and assignment, with 6,101 malaria cases prevented.

The same effective coverage, 6,101, is achieved in the optimal solution when assignments of zones are made to the closest distribution center with respect to quality-adjusted transportation costs (Table 20.11). This is the optimal effective coverage that can be obtained while opening DCs A and D, as determined by solving a revised MIP to optimality.

As demonstrated above, the use of the decision support tool can be valuable in gaining an understanding about the best allocation and assignment strategies and for learning about the tradeoffs within the system. Additionally, the model can be used to examine sensitivity to inputs or to examine objectives not in the model, such as geographical equity. By utilizing a decision support tool, rather than a strict optimization model, it is possible for decision makers to identify a strategy that accounts for the many tradeoffs of the model in addition to social and political concerns.

## 4 Conclusion

The model for IRS malaria prevention presented here provides an example of the types of healthcare resource allocation models that can be created for the developing world. First, the model addresses the role of poor infrastructure, a significant challenge in developing countries, in allocation decisions. Additionally, resource allocation must incorporate the interdependencies of different levels of decision making. Here, the decision support tool addresses the tradeoffs in location of distribution centers, assignment of distribution centers to zones, allocation of teams to distribution centers, and scheduling of teams.

To have practical value in affecting health in developing countries, models incorporating the many levels of decisions are needed. While this model can be extended to address alternate malaria prevention methods, such as LLIN, or prevention of other vector-borne diseases, research is needed to develop similar models and applications for other diseases and illnesses. In such cases, resource allocation models should also address specific disease characteristics, such as seasonality, risk factors, or the heterogeneous nature of disease incidence by geographic location.

While there is potential for the development and application of new tools and analytical models for use in developing countries, a variety of challenges exist. For example, limitations in our model include the use of a finite time horizon and difficulties in estimating model parameters, such as quality-adjusted transportation costs and expected number of infections. Another key roadblock in model development is the availability of data about infrastructure and population

**Table 20.10** Balanced potential malaria cases to DC A and DC D with associated optimal schedules

| | | | | Zone 1 | | | Zone 2 | | | Zone 3 | | | Zone 4 | | | Zone 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Cases prevented | $(\sum_j L_{1,t,j})$ | | | $(\sum_j L_{2,t,j})$ | | | $(\sum_j L_{3,t,j})$ | | | $(\sum_j L_{4,t,j})$ | | | $(\sum_j L_{5,t,j})$ | | |
| Period ($t$) | DC A | DC D | (optimal) | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| Assignment 3 | 4, 5 | 1, 2, 3 | 3,959 | 0 | 0 | 0 | 5 | 2 | 0 | 0 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 0 |
| Assignment 4 | 1, 2, 3 | 4, 5 | 6,101 | 1 | 0 | 0 | 0 | 6 | 0 | 10 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 0 |

**Table 20.11** Smallest distribution cost assignments to DC A and DC D with associated optimal schedules

| | | | | Zone 1 | | | Zone 2 | | | Zone 3 | | | Zone 4 | | | Zone 5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Cases prevented | $(\sum_j L_{1,t,j})$ | | | $(\sum_j L_{2,t,j})$ | | | $(\sum_j L_{3,t,j})$ | | | $(\sum_j L_{4,t,j})$ | | | $(\sum_j L_{5,t,j})$ | | |
| Period ($t$) | DC A | DC D | (optimal) | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| Assignment 5 | 1, 2, 3, 5 | 4 | 6,101 | 1 | 0 | 0 | 0 | 6 | 0 | 10 | 0 | 0 | 4 | 4 | 0 | 0 | 0 | 0 |

density. Therefore, general assumptions may need to be made pertaining to the transportation costs or other model input. Due to the scarcity of reliable data, developers should consider model robustness, e.g., uncertainties or inaccuracies in the data.

In addition to model development, a key feature in the design of a decision support tool for use in developing countries is the software platform that is used. For example, spreadsheet-based tools may be easier to use and more affordable than sophisticated optimization software. An application that requires an expensive software package or even extensive computing power may not be appropriate for application in the field. Therefore, a key consideration must be the situational factors of those using the application. One possible consideration for applications is to utilize cell phone technology, such as text messaging, for the collection or dissemination of information. Additionally, web-based tools can prevent the need for the installation of sophisticated software technology. Although, in many ways, technology considerations are limiting, they encourage the development of creative approaches that result in high value products which are easy to implement.

With extremely limited funds, the difference in impact from making efficient or inefficient allocation decisions can be significant, as demonstrated by the numerical example in Sect. 3. Evaluating decisions can be challenging when disease prevalence changes with time and location, as is the case with malaria. For this reason, utilizing an optimization model for the development of a decision support tool can have significant benefit in evaluation of allocation decisions and teaching the value of analytic methods.

Decision makers participated in ongoing discussions about the development of the model presented here, which influenced the tool structure, input parameters, and the overall approach. Several other decision makers have expressed interest in the tool for educational or planning purposes. In addition to providing a learning experience for decision makers, decision support tools can be useful for education of a broader audience about the complexities of resource allocation decisions.

The development of decision support tools based on analytical modeling has the ability to bring awareness to the value and importance of integrating resources from a variety of research areas. Research that continues to merge different areas such as epidemiology, operations research, and economics is needed to improve decision making for resource allocation in developing countries.

**Acknowledgements** We would like to acknowledge the work of Georgia Tech students Michele Cataldi, Christina Cho, Cesar Gutierrez, Jeff Hull, Phillip Kim, and Andrew Park, who developed the initial model and study from which the current decision support tool and model were created. Additionally, we would like to acknowledge Hannah Smalley and Mallory Soldner for their work in the development of the decision support tool and a related case study. This research is supported in part by Andrea L. Laliberte, the Harold R. and Mary Anne Nash Endowment, the Claudia L. and J. Paul Raines Endowment, and Richard and Charlene Zalesky.

## References

Alistar S, Brandeau ML (2012) Decision making for HIV prevention and treatment scale up bridging the gap between theory and practice. Med Decis Making 32(1):105–117

Balcik B, Beamon B, Smilowitz K (2008) Last mile distribution in Humanitarian relief. J Intell Transport Syst 12(2):51–63

Brandeau ML, Zaric GS, Richter A (2003) Resource allocation for control of infectious diseases in multiple independent populations: Beyond cost-effectiveness analysis. J Health Econ 22(4):575–598

Carr CC, Jallah JD (2008) Improving spatial accessibility to antiretroviral treatments for HIV/AIDS. Ph.D. thesis, University of Zaragoza

Centers for Disease Control and Prevention Division of Parasitic Disease: Malaria Worldwide - Impact of Malaria (2010) Accessed 1 May, 2012. http://www.cdc.gov/malaria/malaria_worldwide/impact.html

Cox J, Craig M, Le Sueur D, Sharp B (1999) Mapping malaria risk in the highlands of Africa. Tech. Rep., December 1999

Craig M, Snow R, Le Sueur D (1999) A climate-based distribution model of malaria transmission in sub-Saharan Africa. Parasitol Today 15(3):105–110

Deo S, Sohoni M (2011) Decentralization of resource-constrained health care networks: Access vs. accuracy tradeoff and network externality. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1754642

Epstein D, Chalabi Z, Claxton K, Sculpher M (2006) Mathematical programming for the optimal allocation of health care resources. https://www.york.ac.uk/media/che/documents/papers/mathprog.pdf

Flessa S (2000) Where efficiency saves lives: A linear programme for the optimal allocation of health care resources in developing countries. Health Care Manag Sci 3:249–267

Hansen K, Chapman G (2008) Setting priorities for the health care sector in Zimbabwe using cost-effectiveness analysis and estimates of the burden of disease. Cost Effect Resource Allocation 6(1):14

Kiszewski A, Johns B, Schapira A, Delacollette C, Crowell V, Tan-Torres T, Ameneshewa B, Teklehaimanot A, Nafo-Traoré F (2007) Estimated global resources needed to attain international malaria control goals. Bull World Health Organ 85(8):623–630

Lasry A, Zaric GS, Carter MW (2007) Multi-level resource allocation for HIV prevention: A model for developing countries. Eur J Oper Res 180(2):786–799

Lasry A, Carter MW, Zaric GS (2008) S4HARA: System for HIV/AIDS resource allocation. Cost Effect Resource Allocation 6:7

Lee B, Brown S, Korch G, Cooley P, Zimmerman R, Wheaton W, Zimmer S, Grefenstette J, Bailey R, Assi T, Burke DS (2010) A computer simulation of vaccine prioritization, allocation, and rationing during the 2009 H1N1 influenza pandemic. Vaccine 28(31):4875–4879

Mapping Malaria Risk in Africa: Mapping Malaria Risk in Africa (2004) http://www.mara.org.za

Ndiaye M, Alfares H (2008) Modeling health care facility location for moving population groups. Comp Oper Res 35(7):2154–2161

Oppong J (1996) Accommodating the rainy season in third World location-allocation applications. Soc-Econ Plann Sci 30(2):121–137

Over M, Bakote'e B, Velayudhan R, Wilikai P, Graves PM (2004) Impregnated nets or DDT residual spraying? Field effectiveness of malaria prevention techniques in Solomon Islands, 1993–1999. Am J Trop Med Hyg 71(Suppl 2):214–223

Rahman S, Smith DK (2000) Use of location-allocation models in health service development planning in developing nations. Eur J Oper Res 123(3):437–452. doi:10.1016/S0377-2217(99)00289-1

Snow RW, Guerra CA, Mutheu JJ, Hay SI (2008) International funding for malaria control in relation to populations at risk of stable Plasmodium falciparum transmission. PLoS Med 5(7):e142

Thomson MC, Doblas-Reyes FJ, Mason SJ, Hagedorn R, Connor SJ, Phindela T, Morse AP, Palmer TN (2006) Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature 439:576–579

United Nations Foundation: Nothing but nets: About the campaign (2012) http://www.nothingbutnets.net/about-us/

Wilson D, Kahn J (2006) Predicting the epidemiological impact of antiretroviral allocation strategies in KwaZulu-Natal: The effect of the urban–rural divide. Proc Natl Acad Sci USA 103(38):14228–14233

Wilson DP, Blower SM (2005) Designing equitable antiretroviral allocation strategies in resource-constrained countries. PLoS Med 2(2):e50

World Health Organization: Public Health Mapping and GIS: The HealthMapper (2012) http://www.who.int/health_mapping/tools/healthmapper/en/

Zaric GS, Brandeau M (2002) Dynamic resource allocation for epidemic control in multiple populations. IMA J Math Appl Med Biol 19:235–255

Zaric GS, Brandeau ML (2001) Resource allocation for epidemic control over short time horizons. Math Biosci 171(1):33–58

Zaric GS, Brandeau ML (2007) A little planning goes a long way: Multilevel allocation of HIV prevention resources. Med Decis Making: Int J Soc Med Decis Making 27(1):71–81

# Index

**A**

Access, 1, 32, 50, 75, 180, 186, 207, 232, 253, 321, 366, 389, 405, 423, 469, 485, 514

Adaptive behavior, 1–16, 54, 56

Ambulance relocation, 317

Ambulance re-positioning, 292, 307–312

Ambulance simulation, 291

Ambulatory care, 193–194, 212, 254

Ambulatory clinic, 253–285

Appointment scheduling, 15, 75–88, 112, 122, 210, 249, 253, 254, 259, 283, 407–411, 420

Approximate dynamic programming, 76, 99, 100, 311, 314, 408, 418

**B**

Base location, 292, 297, 298, 301–308

**C**

Capacity allocation, 100, 110, 194, 207–228

Capacity management, 168, 170, 172, 410, 488

Capacity planning, 75, 76, 95–104, 112, 159, 161, 164, 167, 172, 173, 179, 181, 192, 194, 196, 198, 199, 208, 209, 228, 336, 361, 408, 412, 416, 419

Case-mix, 96, 144, 199, 227

Case study, 133, 156, 172, 174–180, 192, 196, 368, 390, 399–403, 416–419, 429, 460–462, 487, 490–505, 507, 513, 515–530

Chempack, 399–403

Clinic, 3, 19, 50, 76, 156, 186, 210, 231, 253, 321, 352, 390, 407, 425, 449, 467, 485, 515

Clinic operations, 193, 226, 233–235, 238, 239, 241, 248, 249, 272–274, 283, 284, 321–347

Clinic Operations Management System (COMS), 238–244, 249–251

Combining queues, 35–37

COMS. *See* Clinic Operations Management System (COMS)

Continuity of care, 215–217, 271, 413

Coordination, 192, 194, 202, 239, 248, 249, 423–425, 427–431, 435, 437–444, 446

**D**

Decision support tool, 132, 173, 203, 513–516, 520, 523–530, 532

Discharge, 3, 5–7, 12–15, 24, 53, 56, 59, 107, 155, 156, 158, 160, 162, 164, 167, 171, 174, 177–179, 181, 186, 198, 203, 364, 369, 436, 439, 441, 442, 452, 453

Discrete convex analysis, 76

Discrete event simulation, 22, 45, 47, 139, 161, 162, 165, 166, 175, 176, 180, 181, 193–198, 210, 291, 363, 368, 410, 416, 425, 428, 442

Disease screening, 109, 112

Distribution, 14, 50, 77, 124, 162, 194, 211, 243, 274, 297, 326, 362, 402, 406, 433, 450, 467, 488, 514

Dynamic decision making, 3, 517

B.T. Denton (ed.), *Handbook of Healthcare Operations Management: Methods and Applications*, International Series in Operations Research & Management Science 184, DOI 10.1007/978-1-4614-5885-2, © Springer Science+Business Media New York 2013

E

Electronic medical record, 5, 192, 272, 284, 425

Emergency, 1, 49, 81, 125, 155, 186, 208, 232, 254, 291, 322, 351, 389, 410, 424, 457, 469, 486

Emergency Department, 1, 22, 49, 106, 155, 186, 232, 351, 400, 415, 424, 480

Emergency medical services, 198, 329, 392

Emergency preparedness, 323–325, 341, 389

Exponential service times, 172, 303

F

Facility Location, 109, 110, 112, 392–396, 399

FIFO. *See* First-in-first-out (FIFO)

Finite waiting room, 21, 25–31

First-in-first-out (FIFO), 20, 25, 53, 411

Flexibility, 42, 47, 49, 181, 190–191, 193, 207–229, 331, 343, 413, 444, 445, 523

G

General service times, 24

H

Healthcare coordination, 429, 443

Health operations management, 322

Health systems, 38, 188, 190, 191, 198–199, 201–203, 207, 321, 323, 328, 359, 390, 391, 405, 406, 415, 420, 423–430, 434, 435, 443, 467–469

Heuristic, 11–13, 15, 88, 89, 98, 99, 106, 109, 110, 127, 139, 140, 142, 145, 181, 198, 211, 225, 227–229, 240, 242, 248, 249, 303, 311, 360, 365, 377, 378, 381, 392, 398–403, 409–411, 414–416, 457, 513–515, 523, 526, 528, 530

Hosmer-lemeshow test, 277

Hospital, 1, 19, 49, 87, 122, 186, 208, 231, 254, 294, 327, 352, 390, 407, 425, 449, 467, 487, 513

Hospitalist service, 438–442

I

ICU. *See* Intensive care unit (ICU)

Immunization, 69, 325, 329

Inpatient, 2, 3, 8, 12, 22, 24, 31, 36, 40, 42, 51, 52, 96, 97, 99, 125–127, 145, 156, 157, 161, 164, 170, 174–178, 188, 190, 191, 196–198, 201–203, 231, 232, 354, 358, 362, 364, 379, 410, 428, 431, 434–436, 444, 469

Integrated operating room, 127, 136, 144

Intensive care unit (ICU), 2, 3, 5–7, 12, 13, 15, 59, 65, 130–132, 136, 137, 139, 140, 142, 144, 155–182, 188, 191, 196, 429, 438, 439, 460

J

Jackson networks, 26, 38

K

Kanban, 474–476, 479, 481

L

Lean, 203, 264, 267, 268, 284, 355, 369, 380, 479, 482

Length-of-stay, 2, 5, 13, 139, 155–161, 165, 167, 168, 180, 361, 372, 375, 378, 380, 407, 429, 459

Literature review, 122, 123, 125, 139, 194, 210–211, 234, 254–272, 292, 352, 357, 360–368, 415, 426–429, 514–515

Local health departments, 321–325, 327–330, 334, 336, 337, 339–341, 343–347

Logistic regression, 274, 276–278, 280, 282, 285

Logistics, 58, 291, 303, 308, 322, 331, 335, 337, 339–340, 467, 468, 470–473, 480–482, 487, 506, 516

M

Malaria prevention, 513–532

Markovian arrivals, 25, 26, 35

Mass prophylaxis, 321–347

Mathematical programming, 83, 139–142, 145, 194, 198, 361, 398

Medical home, 425, 443, 446

Medical supplies, 58, 326, 330, 339, 450, 453, 455, 463, 467, 469, 470, 473, 475, 477, 479, 482, 486

Mixed integer programming (MIP), 87, 89, 105, 111, 127, 140, 142, 245, 248, 249, 390–396, 488, 520, 521, 525, 530

Models, 1, 19, 45, 76, 139, 156, 195, 212, 238, 255, 291, 325, 352, 389, 405, 425, 453, 478, 488, 514

Multiple servers, 22, 24, 34–37, 198, 378

**N**
No-show, 4, 8–12, 15, 77, 78, 80, 81, 97, 99, 193, 210, 211, 220, 222, 223, 227, 236, 237, 248, 253–285, 407–409, 412–414, 420, 444, 445
No-show and health outcomes, 260, 271–272
No-show interventions, 268
No-show model validation, 277
No-show predictors, 258, 269–271
No-show rate, 15, 210, 220, 222, 223, 254–256, 260, 263, 265, 267–270, 272, 278, 280, 281, 412–414
Nursing units, 467–482

**O**
Open access, 11, 12, 98, 209–211, 216, 217, 219–226, 228, 245, 259, 264, 267, 268, 284, 413, 414, 445
Operating room, 2, 51, 75, 76, 87–95, 112, 113, 121–145, 163, 174, 176, 188, 191, 197, 362, 407, 429, 469, 477, 478, 489, 513
Operating room planning and scheduling, 121–145
Operating room scheduling, 75, 76, 87–95, 112, 429, 489
Operational metrics, 234
Optima, 291, 292, 296, 300, 301, 308, 312–315
Optimization, 11, 13, 75–115, 127, 132, 133, 139, 142, 168, 201, 210, 220, 226, 228, 239, 240, 244–246, 248, 335, 336, 360, 400, 416, 451, 452, 455–457, 459, 488, 496, 499, 500, 507, 513–515, 530, 532
Organ allocation, 109, 112, 113

**P**
Pandemic, 56, 58, 69, 322–329, 333, 334, 339–341, 343, 346
Panel, 8–10, 32, 33, 42, 68, 73, 209–221, 223, 226–228, 283, 412
PASTA. *See* Poisson arrivals see time averages (PASTA)
Patient blocking, 157–158, 160, 174–180
Patient characteristics, 4, 125–128, 366, 367, 429, 457
Patient flow, 22, 23, 42, 49, 51–52, 59, 156, 161, 163, 170–173, 175, 178, 186–201, 203, 227, 228, 235, 236, 240, 249, 358, 359, 371, 372, 378–380, 382, 408, 416, 420, 429, 431, 438, 439, 442
*p*-center, 395–397, 399, 401, 402
*p*-dispersion, 398
Performance measures, 22, 76, 77, 122, 125, 128–133, 136, 145, 158–163, 167–170, 178, 180, 210, 237, 263, 301, 304, 316, 401–403, 434
Planning and scheduling, 121–145, 194, 234–243, 248, 415
*p*-median, 393–397, 399, 401, 402
Point of dispensing (POD), 322–326, 330–346
Poisson arrivals *see* time averages (PASTA), 26, 40
Pooling servers, 35
Primary care, 1, 8, 15, 32, 97, 98, 122, 207–229, 234, 255, 256, 258, 260, 262, 263, 269–272, 284, 330, 391, 424, 427, 428, 431, 437, 444
Priority queues, 22, 31, 39–42
Public health, 58, 321–324, 327–331, 333–337, 341, 343, 345, 359, 389, 392, 487, 513, 515, 519
  public health emergency, 322, 324, 327, 331, 335, 337, 341, 345

**Q**
Quality, 2, 15, 53, 68, 70, 105–107, 110, 111, 127–129, 159, 162–164, 166, 169, 172, 181, 185–187, 189, 190, 196, 197, 201, 204, 243, 300, 302, 309, 321, 335, 344, 389, 391, 392, 405, 407, 420, 423, 426, 444, 445, 451, 458, 467, 468, 474, 479, 489, 492, 493, 514, 515, 519, 521, 524, 530
Queuing, 47, 50, 51, 128, 132, 137, 140, 194–197, 199, 291, 336, 351, 361–363, 368, 406, 408, 412, 420, 428, 429

**R**
Radio-frequency identification (RFID), 51, 55, 56, 243, 251, 473, 476–482
Readmission, 6, 13, 15, 159–161
Real-time optimization, 312
Resource Allocation, 163, 175, 181, 199, 202, 513–515, 519, 523, 530, 532
RFID. *See* Radio-frequency identification (RFID)
ROC curve, 277, 280, 282

**S**
Self-reported reasons for no-show, 255, 259–264, 267, 284

Service, 2, 19, 50, 75, 121, 166, 185, 207, 231, 253, 291, 321, 352, 389, 406, 423, 451, 468, 491
Service protocols, 20–22, 25, 39–41
Simulation-optimization, 228, 459
Squared coefficient of variation, 27
Steady state, 9, 21–24, 33, 47, 166, 169, 170, 172, 376, 377, 496
Stochastic, 2, 6, 21, 38, 51, 56–59, 76, 77, 80, 82–87, 89, 90, 95–98, 109, 111, 114, 125, 127, 132, 133, 137–139, 141, 142, 161, 163, 166, 169, 170, 178, 194, 196, 210, 218–220, 222, 226–228, 297, 310, 314, 392, 406–410, 412, 414, 420, 453, 459, 460, 478, 481
Stochastic models, 6, 196, 392, 412
Stochastic programming, 76, 83–87, 89
Supply chain design, 480, 493
Supply chain management, 467–482
System dynamics, 45, 47, 49, 70, 71, 425, 427, 430–437, 457

**T**
Transplantation, 109, 110, 112, 113
Two-bin system, 473, 474, 476–479, 481, 482

**U**
Uncertainty, 2, 75–78, 84, 85, 88–90, 95, 97, 106, 112, 125, 127, 137–139, 144, 145, 185, 189–191, 194, 234–237, 250, 268, 335, 352, 382, 451, 453

**V**
Vaccine design, 109, 112
Variability, 2, 3, 7, 13, 27, 29, 88, 99, 122, 127, 139, 164, 170, 185, 196, 201, 210, 215–218, 224, 227, 236, 250, 260, 297, 301, 316, 354, 355, 406, 409, 411, 414, 429, 430, 434

**W**
Work conserving service protocols, 39, 40
Workforce scheduling, 75, 76, 105, 109, 112