International Series in
Operations Research & Management Science

Randolph Hall *Editor*

Patient Flow

Reducing Delay in Healthcare Delivery

Second Edition

Springer

International Series in Operations Research & Management Science

Volume 206

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

Randolph Hall
Editor

Patient Flow

Reducing Delay in Healthcare Delivery

Second Edition

Springer

Editor
Randolph Hall
Epstein Department of Industrial and Systems Engineering
University of Southern California
Los Angeles, California, USA

ISSN 0884-8289 ISSN 2214-7934 (electronic)
ISBN 978-1-4614-9511-6 ISBN 978-1-4614-9512-3 (eBook)
DOI 10.1007/978-1-4614-9512-3
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013955436

Printed on acid-free paper

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

Preface to the Second Edition

Over the last 100 years, quality of life and human longevity have improved in most of the industrialized world as a result of advances in human health. We have benefited from reduced exposure to disease (through such measures as vaccinations and improved water quality) and developed treatments that reduce the consequences of disease once exposed. Nevertheless, humans continue to suffer because they do not have access to appropriate healthcare, or because healthcare is delivered in a manner that is confusing or inefficient. The gap between the science and the practice of healthcare is large.

This book is dedicated to improving healthcare through reducing the delays experienced by patients. One aspect of this goal is to improve the flow of patients, so that they do not experience unnecessary waits as they flow through a healthcare system. Another aspect is ensuring that services are closely synchronized with patterns of patient demand. Still another aspect is ensuring that ancillary services, such as housekeeping and transportation, are fully coordinated with direct patient care. Past experience shows that effective management of healthcare delays can produce dramatic improvements in medical outcomes, patient satisfaction, and access to service, while also reducing the cost of healthcare.

Within the 21 chapters of this book—the *Second Edition of Patient Flow: Reducing Delay in Healthcare Delivery*—readers will be exposed to a set of techniques and strategies that can be used by clinicians and administrators to substantially reduce delays in healthcare delivery. The second edition expands on the first by providing more information on the consequences of delay, prioritizing patients, modeling integrated systems, and implementing change, all in an effort to improve healthcare in hospitals, clinics, and healthcare offices. Reflecting the highly interdisciplinary nature of this book, the chapters have been written by doctors, nurses, industrial engineers, system engineers, and geographers. Reflecting the global challenges of patient flow, authors reside in eight countries and four continents. These perspectives provide the comprehensive view needed to address the problem of patient delay.

In the first part, the book begins by examining healthcare as an integrated system. Chapter 1 provides a hierarchical model of healthcare, rising from departments, to centers, regions, and the “macro system.” The chapter also demonstrates system modeling for a large urban hospital. This is followed by a new chapter that demonstrates the use of simulation to assess the interaction of system components while seeking to achieve performance goals. The part concludes with Chap. 3, providing hands-on methods for developing process models, using these models to identify and remove bottlenecks, and developing facility plans.

The next part addresses crowding and the consequences of delay. Two new chapters (Chaps. 4 and 5) focus on delays in emergency departments, which are particularly prone to delays. The impact of delays is further explored in Chap. 6, which examines medical outcomes that result from waits for surgeries.

The third part concentrates on the management of demand, including appointments, prioritization, and triage. Chapter 7 presents a set of breakthrough strategies that use real-time monitoring systems for continuous improvement. Chapter 8 focuses on the patient appointment system, particularly through the approach of advanced access, which makes appointments more immediately available to patients. Chapter 9 concentrates on management of waiting lists for surgeries and the allocation of available capacity to meet patient demands. The part concludes with Chap. 10, an examination of triage outside of emergency departments, with a focus on allied health programs.

Part IV offers analytical tools and models to support the analysis of patient flows. Chapter 11 offers techniques for scheduling staff to match patterns in patient demand, and thus reducing predictable delays. The literature on simulation modeling, which is widely used for both healthcare design and process improvement, is surveyed in Chap. 12. The next chapter, Chap. 13, is new to the second edition and demonstrates the use of process mapping to represent a complex regional trauma system. Chapter 14 provides methods for forecasting demand for healthcare on a region-wide basis. Then Chap. 15 presents queueing theory as a general method for modeling waits in healthcare. Last in the group, Chap. 16 focuses on the rapid delivery of medication in the event of a catastrophic event, such as a pandemic or terrorist attack.

The last part of the book concentrates on achieving change. Chapter 17 provides a diagnostic for assessing the state of a hospital and using the state assessment to select improvement strategies. Chapter 18 demonstrates the importance of optimizing care as patients transition from one care setting to the next with an emphasis on clinical outcomes and the business case. Chapter 19 is new to the second edition and shows how to implement programs that improve patient satisfaction while also improving flow. Chapter 20 illustrates how to evaluate the overall portfolio of patient diagnostic groups to guide system changes. Lastly, Chap. 21 provides project management tools to guide the execution of patient flow projects.

Since the first edition was completed, considerable change has occurred in American healthcare policy, through the passage of the Affordable Care Act. This legislation aims to make healthcare insurance more available and affordable to consumers. But to achieve its larger aims of reducing the cost of healthcare, change will be needed to improve healthcare efficiencies and effectiveness, like those provided in this book.

This book is intended to motivate and guide change so that healthcare systems around the world give more priority to reducing patient delay and implement changes that dramatically improve healthcare. The chapters of this book illustrate that radical changes in the management of patient flow and patient delay are not only possible but also essential to ensuring that advances in medical practice keep pace with advances in medical science.

Los Angeles, CA Randolph Hall

Contents

Part I Integrated Healthcare Systems

Part II Crowding and the Consequences of Delay

Part III Demand, Prioritization and Appointments

Part IV Modeling and Analysis Tools

Part V Implementing Change

Contributors

Douglas L. Andrusiek Paramedical Sciences, Edith Cowan University, Perth, WA, Australia

David Belson Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA, USA

Emilio Cerdá University Complutense of Madrid, Madrid, Spain

Peter Congdon Department of Geography, Queen Mary and Westfield College, London, UK

Laura de Pablos University Complutense of Madrid, Madrid, Spain

Maged Dessouky Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA, USA

David C. Evans Department of Surgery, University of British Columbia, Vancouver, BC, Canada

Linda Green Graduate School of Business, Columbia University, New York, NY, USA

Randolph Hall Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA, USA

Shane N. Hall Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Katherine Harding Allied Health Clinical Research Office, Eastern Health, Melbourne, VIC, Australia

Sheldon H. Jacobson Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Kirk Jensen BestPractices, Fairfax, VA, USA

Hongzhong Jia Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA, USA

Alexander Kolker API Healthcare, Hartford, WI, USA

Linda Kosnik Overlook Hospital, Summit, NJ, USA

Lisa Kuramoto Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal Health Research Institute, Vancouver, BC, Canada

Adrian Levy Department of Community Health and Epidemiology, Dalhousie University, Halifax, Canada

Mark Lindsay Mayo Health System, Eau Claire, WI, USA

Megan McHugh Northwestern University, Chicago, IL, USA

Pavan Murali Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA, USA

Fernando Ordóñez Industrial Engineering Department, Universidad de Chile, Santiago, Chile

Roger Resar Institute for Healthcare Improvement, Boston, MA, USA

Maria V. Rodríguez Uría University of Oviedo, Oviedo, Asturias, Spain

Sergei Savin Graduate School of Business, Columbia University, New York, NY, USA

Zhihong Shen Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA, USA

Boris Sobolev School of Population and Public Health, The University of British Columbia, Vancouver, BC, Canada

James R. Swisher Mary Washington Hospital, Fredericksburg, VA, USA

Nicholas Taylor Allied Health Department, Eastern Health and La Trobe University, Melbourne, VIC, Australia

Jan Vissers Erasmus University Medical Centre, Institute for Health Policy and Management, Rotterdam, Netherlands

Michael Warner AtStaff, Inc., Durham, NC, USA

Michael Williams The Abaris Group, Walnut Creek, CA, USA

Part I
Integrated Healthcare Systems

Chapter 1
Modeling Patient Flows Through the Health care System

Randolph Hall, David Belson, Pavan Murali, and Maged Dessouky

Abstract Health care systems can be evaluated from four perspectives: macro, regional, center, and department. In each case, reduction of patient delay depends on improving interfaces as patients are transferred from activity to activity or department to department. This chapter presents basic tools for resolving delays at interfaces, through mapping the processes by which patients are served, and by developing and implementing measures of system performance. These tools are demonstrated through a case study of the Los Angeles County/University of Southern California Hospital.

Keywords Process charts • Performance measurement • Health care systems

1 Introduction

Health care systems have been challenged in recent years to deliver high quality care with limited resources. In the USA, large segments of the population have inadequate health insurance coverage, forcing them to rely on an underfunded public health system. At the national level, total expenditures in the year 2010 amounted to $2.6 trillion, or 17.9 % of the GDP, representing a doubling in cost in a decade. Costs are projected to rise to 19.6 % of GDP by 2021, according to the Center for Medicare and Medicaid Services.

Given the pressures to contain costs, it is critical for hospitals and health care systems to develop systems that ensure the best possible patient care within limited resources. An important aspect of this objective is to develop procedures to improve patient flow, to provide timely treatment and maximum utilization of available

R. Hall (✉) • D. Belson • P. Murali • M. Dessouky
Epstein Department of Industrial and Systems Engineering, 200 GER, University of Southern California, Los Angeles, CA 90089-0193, USA
e-mail: rwhall@usc.edu

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_1, © Springer Science+Business Media New York 2013

resources. Patient flow analysis represents the study of how patients move through the health care system.

1.1 Emergency Departments: Example of System Delays

Emergency Departments (ED) are perhaps the most challenged components of the health care system with respect to patient delay. Patients arrive at the emergency department through multiple channels, including walk-in (or drive-in) and ambulance. Depending on the nature of the emergency, the patient may be served through an ambulatory or a nonambulatory section of the emergency department. The patient (or a friend) meets with a receptionist to collect background information and meets a nurse for triage. Patients are served by physicians and nurses in treatment rooms, which may be specialized to particular injuries (e.g., orthopedics) or specialized by level of urgency. Before treating the patient, tests (X-ray, CT Scan, MRI, etc.) may be needed through a radiology department. In some cases patients must be moved to an operating room for surgery. Once emergency treatment is completed, it may be necessary to admit the patient to the hospital, in which case the patient is exposed to additional processes and delays. Eventually, the patient undergoes a discharge process, and his or her bed must be prepared for the next patient.

As stated in a recent study by the American College of Emergency Physicians (ACEP 2002), "A multitude of factors are responsible for crowding, including higher patient acuity, prolonged ED evaluations, inadequate inpatient bed capacity, a severe nursing shortage, problems with access to on-call specialists and the use of the ED by those with no other alternative to medical care, such as the uninsured." In 2010, 130 million ED visits occurred in the USA, representing 42.8 visits per 100 people. Growth in ED visits has resulted from the combined effect of a reduction in the number of people with insurance coverage (40 % of visits were paid through private or commercial insurance according to McCaig and Burt in their 2001 paper) along with the mandate under the Emergency Medical Treatment and Labor Act (EMTLA) that EDs not refuse service to any patient on the basis of ability to pay.

Emergency departments have been especially reliant on public sources, as people with no health insurance sometimes have no alternative for receiving medical care. Seventy-eight percent of hospital administrators stated that their hospitals are inadequately reimbursed for emergency care, with 80 % citing a "poor payer mix," in a 1995 survey (Greene 1995). The mismatch between available funding and potential demand has made emergency departments particularly susceptible to patient delays, with their attendant consequences on quality of care (also see Derlet and Richards (2000), Schneider et al. (2001), and Schull et al. (2001)). For instance, Bindman et al. (1991) found, in their study of San Francisco General Hospital, that 15 % had left the hospital before being called for their examination and that "almost twice as many patients who left without being seen reported at follow-up that their pain or the seriousness of their problem was worse." (Dershewitz and Paichel (1986), Buesching et al. (1985), Derlet and Nishio (1990), and Shaw et al. (1990) provide related studies.)

In this description, it should be apparent that medical care is delivered through a network of service stations, and that there is potential for delay in multiple locations. It should also be apparent that emergency departments, as a system, closely interact with other systems. Emergency departments are part of the "emergency medical system" (EMS), which includes the management of responders (fire, paramedic, ambulance), and the distribution of service among hospitals (e.g., the routing of an ambulance to a particular hospital). Emergency departments (EDs) also interact with general hospital care, as a frequent source of queueing is the inability to place a patient in a hospital bed once treatment is completed in the ED. Less obviously, emergency departments interact with clinical care, as ED demand is a by-product of the patient's ability to receive treatment through a primary care provider, access preventive care, and adopt a healthy lifestyle.

1.2 *Goal of Book*

This book presents strategies, concepts and methods that can radically improve the delivery of health care by reducing delays. Our supposition is that much of the delay accepted by the public is both unnecessary and costly. Patients are harmed in the process of delay, not only through wasted time, but through unnecessary suffering, and through adverse medical outcomes. Health care providers are harmed through the added cost and reduced efficiency resulting from the complications of handling delayed patients. For these reasons, it is imperative for all providers to seek out and implement solutions that reduce delay.

The study of health care delay is an application of the discipline of queueing theory (i.e., the study of lines and waits, Hall 1991). Health care is similar to other forms of queueing in these respects:

- Variations in the demand for service are in part predictable (e.g., result of time-of-day patterns) and in part random.
- Health services require coordination of multiple resources, such as physicians, medications, and diagnostic equipment.
- Services are provided in multiple steps, through a network of services, with the attendant issues of "grid lock" and "bottlenecks."
- Delays can be reduced through careful forecasting, scheduling, process improvement, and information management.

In these respects, reducing health care delays is similar to the efficient coordination of work in a factory. But health care has unique features, which demand specialized methods and research, as presented in this book.

- Services can usually only be provided when the patient is physically present (unlike a piece of work that can be dropped off and picked up later), which makes rapid service particularly important.
- A patient awaiting treatment may require significant continuing care (thus, waits translate into demand for more work).
- The outcome of the service—including survival, recovery time, and suffering—is adversely affected by waits.
- Schedules and plans are likely to be disrupted due to the arrival of critical patients, who can both require an exceptional amount of attention, and be exceptionally urgent.
- A patient's condition may independently change while waiting and require more or different care.

Foremost, however, health care is a system that can be improved through a better understanding of the system components and their relationships to each other, as will be discussed in the following section.

1.3 Modeling Health care as a System

EDs are but one component of the entire health care system, which might be better characterized as a system of systems, as described below.

1.3.1 Macro System

The macro system, depicted in Fig. 1.1, encompasses the set of activities that affect a person's well-being, from birth to death. From the macro perspective an individual only leaves the health care system at the end of life because he or she is constantly making decisions and engaging in activities that affect his or her health, whether or not under the direct care of a health care provider.

Figure 1.1 portrays six states of being, reflecting one's state of health, and reflecting whether (and for what purpose) one is present at a health care facility. The state of recuperation, for instance, is a period of recovery spent outside of health facilities (typically at home), with diminished health and likely under the supervision of a physician (necessitating occasional post-care visits).[1]

Broadly speaking, the goals of the macro system are simple:

- Maximize the years from birth to death (length of life).
- Maximize the proportion of one's life spent in the "well" state.
- Maximize the quality of life when not in the well state.

[1] The model is a simplification as it does not explicitly account for comorbidities. For certain chronic conditions, the patient may never be cured, but achieve an improved state of wellness, and patients may be cured of some conditions but not others.

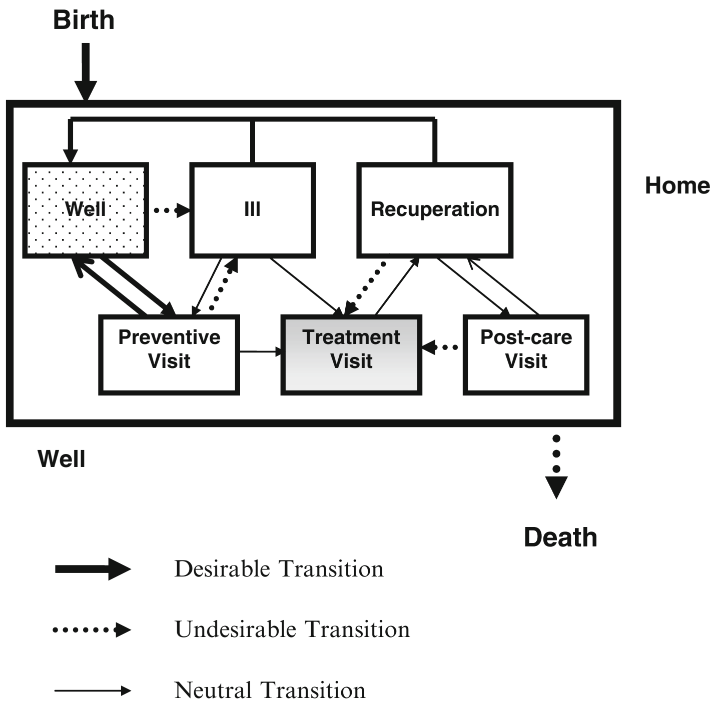


Fig. 1.1 The macro health system

Reducing patient delay and improving patient flow accomplish these macro goals by: (1) improving access to care, we increase expected length of life; (2) minimizing the length of time spent in undesirable states (e.g., at a health care facility awaiting treatment, undergoing treatment or recuperating from treatment); and (3) reducing waiting time until treatment, we improve quality of life when not well.

By improving access to care, we also hope to minimize the frequency by which we enter into the undesirable states (e.g., minimize dotted line transitions shown in Fig. 1.1, such as becoming ill or being readmitted for new treatment after beginning recuperation) and maximize the likelihood of transitioning to a desirable state (e.g., bold line transitions, such as becoming well after being ill). More broadly, personal health is optimized through an inter-related set of actions over the course of one's life, some of which are the consequence of health care delay, and some of which lead to changes in health care delay by changing societal demand for health services (Table 1.1).

Table 1.1 Goals and actions to improve macro health

State	Goal	Actions	Examples
Well	Maximize time in state; minimize likelihood of presenting well	Minimize exposure, personal lifestyle, reduce societal risks, self awareness of symptoms	Clean water and safe housing; stop smoking; no exposure to violence; patient education
Ill	Minimize time until well or treatment	Self awareness of symptoms, self care when appropriate	Patient education; non-prescription medication
Recuperation	Minimize time until well; minimize likelihood of relapse	Home care, increase compliance; medication	Home nursing services; improved discharge instructions and assessment prior to discharge
Preventive visit	Timely access to primary care; maximize likelihood of detection; apply appropriate care	*Improve patient flow*, apply appropriate diagnostics and preventive measures	Optimized appointments; apply appropriate tests based on risk factors; vaccinate on schedule
Treatment visit	Timely access and minimum length of stay; successful treatment or medications	*Improve patient flow*, correct diagnosis, apply appropriate medical care	Optimized scheduling; apply appropriate care
Post-treatment visit	Timely access to care; apply appropriate care	*Improve patient flow*, apply appropriate medical care	Optimized appointments; apply appropriate care

1.3.2 The Regional Health System

While the macro system represents the state of individual health, the regional system portrays the organizational and functional relationships among health care processes. This is invariably a hierarchy, beginning with primary care providers, through private practices and local clinics; moving to secondary care providers, through community hospitals; and then moving to tertiary providers, through regional medical centers (some highly specialized quaternary care is only available at a few national centers). As the geographical scale becomes larger, increasingly specialized care becomes available, a consequence of scale economies and a consequence of aggregated patient demand. On the other end of the spectrum, more routine care is conveniently obtained from primary care providers. The primary, secondary and tertiary providers are augmented by ancillary services, such as MRI centers, laboratories or dialysis facilities, which may support multiple providers (again influenced by scale economies); continuing care facilities, such as nursing homes; or, on a more basic scale, pharmacies.

Many variations exist within this general framework, which has evolved over time as more specialized treatments have become available, health care plans have changed, costs have changed and people have become more mobile. On the one hand, by creating multiple layers of care, some delays are inevitably created due to increasing difficulty of access. On the other hand, without multiple layers, some types of specialty care would not be available at all due to insufficiently trained caregivers or simply cost. Yet it is clear that the regional/national system should be designed with three (somewhat conflicting) goals in mind:

- Minimizing the cost of providing desired services.
- Maximizing convenience and access to services that individuals need.
- Maximizing the likelihood of a positive health outcome from service.

Reducing health care delay contributes to all three goals by: (1) removing inefficiencies in the provision of services, thus reducing cost, (2) providing timely access to the services people need, and (3) reducing waits for needed services.

It should be recognized that the regional health system is in part the result of deliberate planning (e.g., design of the emergency medical system and designation of trauma centers), in part due to happenstance (e.g., where hospitals happened to develop many years ago), in part due to market pressures (e.g., competition for patients among facilities and health plans), and in part due to factors that originate from outside the region (e.g., regulations, medical discoveries, and medical training). Thus, it would be impossible to fully optimize such a system, though it might be steered in a desirable direction.

1.3.3 Health care Center

The center is a grouping of geographically proximate facilities under the management of one organization. At a minimum, this entails two or more interacting departments, each with a distinct function (e.g., a laboratory and an outpatient clinic). At a maximum, this could encompass a larger tertiary or quaternary care medical center with dozens of departments.

A health care center operates as a system of interacting departments, which must be coordinated through the flow of patients, specimens, employees, information, materials, and pharmaceuticals. The center leadership, facility design, organizational design, employee training, and recruitment are all important factors. Centers can sometimes represent enormous multibillion dollar investments, and are frequently judged based on standards of financial return, quality of care, and medical outcomes. Patient flow is particularly important for centers, as flow from department to department needs coordination; otherwise delays at the interfaces can be significant. During a typical center visit, the patient may separately encounter waits for these services:

- Information collection as part of admission.
- Diagnostics and examinations.
- Procedures, surgeries, and therapies.
- Education.
- Rehabilitation and recuperation.
- Transportation between departments.
- Discharge processes.

In the background, patients may be delayed waiting for ancillary services, many of which are invisible to patients:

- Transfer of medical records.
- Transfer and analysis of laboratory specimens.
- Filling prescriptions.
- Housekeeping to prepare rooms for new patients.
- Communication among departments, scheduling and decision-making in preparation for patient arrivals.
- Movement and availability of wheelchairs, gurneys, and other portable equipment.
- Completion of required paperwork for internal or governmental use.

Thus, patient delays depend in part on how he or she physically flows through the center, and in part on how information, equipment, and other objects flow through the center.

In summary, the system for managing patient flows in a center should be designed and operated to achieve these goals:

- Minimizing waits as patients transition from department to department.
- Achieving a high level of synchronization among patients, employees and resources, so that services begin promptly on patient arrival and are provided with high efficiency.
- Identifying and resolving system level bottlenecks that impede the flow of patients.

These goals can only be achieved through effective coordination and communication, combined with constant attention to patient service.

1.3.4 Department

The department is the most microscopic of the systems we consider. It represents a unit within a larger center oriented toward performing a single function, or a group of closely related functions. Examples include the ED, surgery, radiology, or an inpatient ward. A department could also be ancillary, such as housekeeping, medical records, or transportation. For the patient, departments typically have clearly defined points of entry and points of exit, which may be time-stamped and correspond to responsibilities for care. Like whole centers, departments are often judged based on financial return. Medical outcomes and quality of care, however, are often more difficult to assess at the department level, as these depend on the totality of service provided by the center over the entire duration of stay or over longer periods of time.

With respect to patient flow, departments must both support the mission of the center as a whole through effective coordination, and be effective in their own right. Neither should a department create unnecessary delay within, nor should they impose delay elsewhere (e.g., delays in accepting patients, or by failing to prepare a patient for transfer when he or she is needed elsewhere). The employees must be trained and rewarded for the priority of minimizing delays through prompt service; they should employ realistic appointment systems; they should ensure proper staffing, and advance planning prior to the arrival of patients.

Later in this chapter we will use a case study to explore, in depth, patient flow issues at the department and center levels.

1.4 Highly Congested Systems

A common feature of health care systems is extreme congestion, meaning that patients have a high likelihood of encountering delay. In part, this is due to inattention to patient flow issues. However, it is also partly due to the financial structure of health care.

Again, we turn to EDs. For major public hospitals in particular, demand for service can be so large that the system rarely empties of waiting patients, either because the ED itself has insufficient capacity, or because the hospital's wards are not absorbing the inflow of patients, thus causing spillback into the ED. For this reason, contrary to conventional queueing models, the system arrival rate exceeds the system service capacity over extended periods—perhaps perpetually. As a result, the system resides in a state of dysfunctional equilibrium, where the only thing that keeps queues from growing without bound is the propensity of some patients to leave without being seen when waits are intolerable (Fig. 1.2), some of whom may return later, possibly in a worsened state, and some of whom will never receive needed treatment. At times when waits become particularly long, more patients will opt to leave without being seen, either immediately at the time of arrival, or subsequently after becoming frustrated with the wait, bringing waits back into balance. When waits become shorter, fewer patients leave, causing waits to gradually build again.

Highly congested queues exist in other contexts, such as Immigration/Naturalization Service Offices, Motor Vehicle Departments and public housing. A common characteristic is that the service provider has limited economic incentive to add capacity (because it is operated at a loss), or re-price service (because EMTALA prohibits financial screening). In fact, in some circumstances, the attraction of "latent demand" (customers who would otherwise leave) may mean that an increase in capacity causes cost to increase, but only slightly reduces average waits. To draw an analogy, expansion of highway capacity may have only small effects on congestion as overall traffic volumes increase.

Another important consideration in highly congested queueing systems is degradation in the ability to deliver quality service. Crowding in waiting rooms and treatment areas, loss in privacy, delays in accessing needed equipment, and delays in providing medication can all add to patient suffering. Continual queueing de-motivates employees, as no matter how hard they work patients will still be queued. And service will be unproductive, as patients must be shuttled in and out of

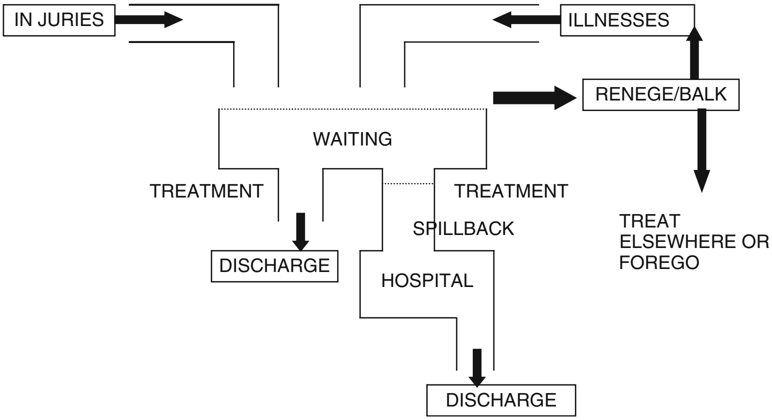


Fig. 1.2 Highly congested system creates spillback and patients who leave without being seen

treatment rooms as they wait for test results or resources. Crowding can also lead to diversion of ambulances to more distant hospitals, slowing the time until patients can be treated (ACEP 1999; Litvak et al. 2001). For all of these reasons, it is impossible to fully address problems in patient flows without considering remedies for health care finance and pricing, topics that go beyond the scope of this book.

1.5 Methods

Health care delays can be reduced through awareness of best practices, application of quantitative methods, and a commitment to change. Organizations such as the Institute for Healthcare Improvement (IHI 1996) and the American College of Emergency Physicians (ACEP 2002) have assembled numerous ideas for improvement. These and other new concepts are captured in this book. Most generally, the solutions to delay problems come in three forms (Hall 1991):

Alter the service process. Through scheduling, coordination, process changes, communication, automation, etc., increase the capacity for serving customers, and increase the synchronization between capacity and customer arrival patterns.

Alter the arrival process. Through appointments, pricing, information, education programs, etc., influence the patterns by which patients present for service, improving the alignment between capacity and demand.

Alter the queueing process. Through triage, moving waiting from the health care facility to the home, redesign of waiting areas, changes in prioritization, etc., ensure that the adverse consequences of waiting are minimized.

These three steps form a hierarchy, as the first priority should be optimizing service to meet the needs of patients; when this is infeasible or uneconomical, explore changes to patient patterns; and, if all else fails, focus on managing queues to maximum efficiency.

Within this chapter, we provide basic methods of industrial engineering that can be used to develop all three types of patient flow solutions. Our principal focus here is process planning (representing the steps needed to deliver service and the interactions between process steps), and performance measurement. These techniques are widely used to improve the performance of manufacturing, distribution and transportation systems, and are emerging as promising approaches to improve health care.

Process planning is an approach for documenting the steps entailed in delivering patient care (or an ancillary process), and redesigning the process for improved efficiency. We will show how to plan a series of processes for patient service, and we will show how to plan for the physical movement of patients. Process planning can be an effective first step toward change as it can reveal the weaknesses and strengths of the current system, and identify methods of improvement.

Performance measurement entails identifying the system goals and the measures by which attainment of the goals are judged. To be effective, performance measurement should be embedded in the continuing operation of the system, so that key decisions are influenced and evaluated, according to established objectives (JCAHO Standard LD.3.11 recommends that management identify critical patient flow processes as well as monitor relevant measures.) Performance measurement should also be transparent to all involved, so that they can witness how their actions affect the overall performance of the system, so that they can be alerted to problems when they occur, and so that they are recognized for their accomplishments. Example measures include waiting time (by step or location), number of patients waiting, number of patients served, patient satisfaction, utilization of resources, and costs. Performance measurement is ultimately useful as an approach for obtaining an accurate and meaningful picture of patient flow and helping determine where improvements can be made. Unfortunately, many hospitals have considerable difficulty making such measurements due to inadequate computer information systems or due to not having the financial resources to create and operate the necessary information system.

2 Case Study: Los Angeles County Hospital

The Los Angeles County/University of Southern California (LAC/USC) Hospital is used as a case study to demonstrate how the techniques of patient flow analysis can be used to create a system model of a center and its departments, and used to improve services (Belson et al. 2004). LAC/USC is a large urban health center serving a largely poor population. It is also the trauma center for central Los Angeles, with the busiest ED in the country, measured in admissions. Approximately 85 % of the patients admitted to beds in the hospital enter through the ED, as of the time of this study. The study was performed at the old hospital, prior to its replacement in 2008.

LAC/USC, including the General Hospital (GH) and its allied Outpatient Department (OPD) and Community Health Centers (CHC), was studied as an integrated system, to identify bottlenecks and recommend improvements. This goal has been accomplished through a series of interviews with administration in 35 hospital departments, as well as focus groups with nurses, doctors, and analysts. Through these meetings we have documented the processes for resource scheduling, patient triage, and patient routing; and we have documented caregiver perceptions of critical issues and problems in serving patients.

Separately, we have obtained, reviewed and analyzed data sources to determine their usefulness for monitoring and evaluating patient flows. We have also performed sample analyses to demonstrate patterns of patient arrivals and interdepartmental flows.

These data sources were used to create process charts that show flows through the hospital as a whole, as well as flows within individual departments. Through our analysis of these charts, as well as our own observations, we developed a series of recommendations for improving patient flows within short-term and long-term time frames, with a focus on improving the coordination among hospital departments.

2.1 Process Map for Center and Aggregate Flows

Patient flow within the center can be viewed at several levels of detail. At the highest level, the center consists of four primary areas: Emergency, Inpatient, Outpatient and the Community Health Centers. Patients frequently move between them and some may visit all four in the course of a year. Patient flow and related costs are summarized in Table 1.2.

Thus, the inpatient area served relatively few patients but represented considerable cost. The emergency area served five times as many patients, at 12 % of the total inpatient cost. Some ED visits are brief and ambulatory while others entail complex trauma care prior to patients moving to an inpatient bed. Outpatient represents an even larger number of visits (about half of the total), with per patient costs similar to ED visits.

The ED was composed of three areas: 1050, 1060, and 1350. The 1350 was for the most critical patients and 1050 for the least critical or ambulatory patients. Inpatient areas in the GH were divided between surgery wards and other medical wards. The Outpatient Department (OPD) was composed of many clinics, each with a separate medical specialty.

The patient flow between areas is summarized in Fig. 1.3, with more detail shown in Fig. 1.4. In still more detail, important flows are shown for the General Hospital in Fig. 1.5. In the following sections we will elaborate on patient flows, first illustrating processes that span departments, and then illustrating processes within departments. These descriptions are extracted from a much longer report (Belson et al. 2004) that provides detailed process maps for all of the major processes and departments in the center.

Table 1.2 Summary of patient workload (2003–2004) and costs

Area	Patients	Total costs
Inpatient	40,000	$475,000,000
Emergency	205,000	$56,000,000
Outpatient	522,000	$160,000,000
Total	749,000	$691,000,000

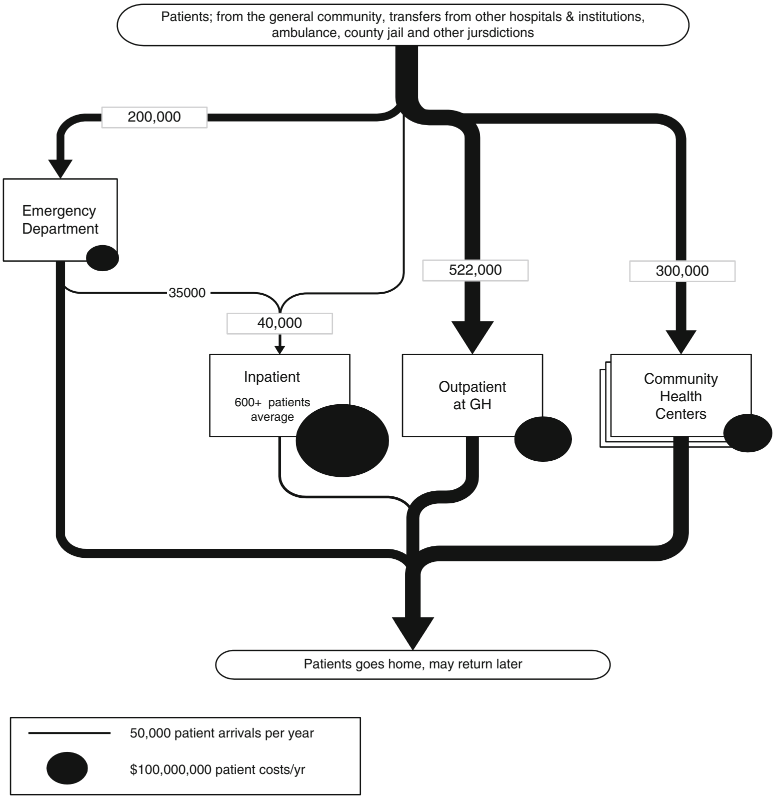

Patients; from the general community, transfers from other hospitals & institutions, ambulance, county jail and other jursdictions
200,000
Emergency Department
35000
40,000
522,000
300,000
Inpatient
600+ patients average
Outpatient at GH
Community Health Centers
Patients goes home, may return later
50,000 patient arrivals per year
$100,000,000 patient costs/yr

Fig. 1.3 Overall patient flow and costs

2.1.1 Service Process 1: Scheduling and Appointments

Patients begin their visit either as a scheduled or an unscheduled patient. Unscheduled patient visits include:

- Walk into the emergency department (these ambulatory patients represent the most common path into the hospital).
- Ambulance delivers patient at the emergency department.
- Walk into certain open clinics in the outpatient department that do not require an advance appointment.

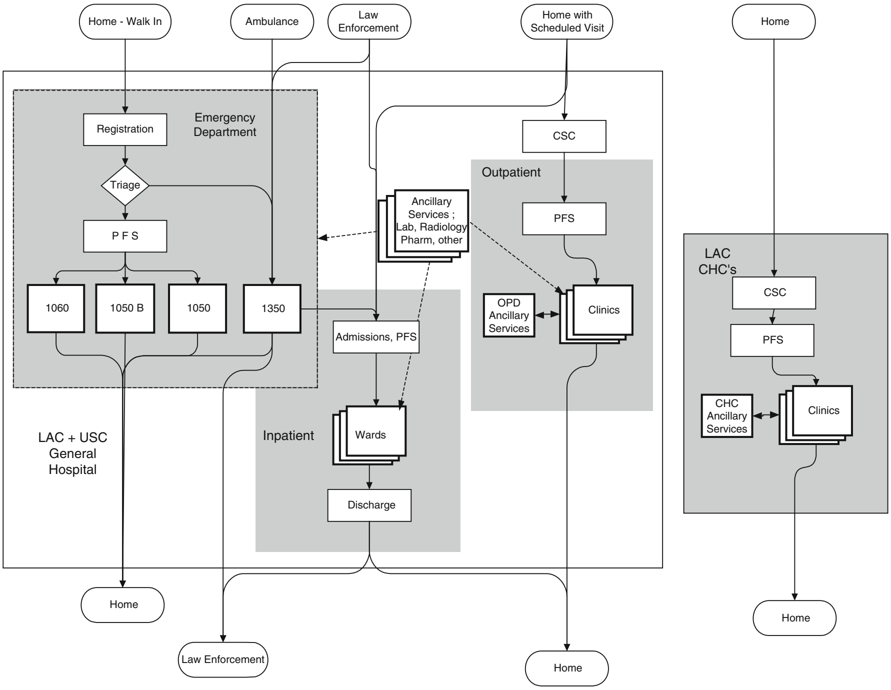


Fig. 1.4 Flow between LAC + USC units

Scheduled patient visits include:

- Scheduled clinic visits that are arranged through the hospital's Customer Service Center (CSC) or the clinics themselves
- Appointments to an inpatient ward, such as scheduled day surgeries
- Scheduled returns to certain GH areas for visits of less than a day

The most common path into an inpatient bed is from the emergency area. This movement is recorded in the hospital's computer information system that is used throughout the hospital, which also tracks bed assignments and appointments.

2.1.2 Bed Management

Movement of patients from the emergency areas to inpatient beds is the responsibility of the Bed Control Unit (BCU). This is coordinated with the staff in the emergency room as well as the hospital wards. Moreover, the Nursing Department has assigned individuals to monitor bed availability; physicians are also consulted on the appropriateness of each movement. Therefore, the bed management process involves several jurisdictions and individuals. The bed control unit locates beds for ED patients based on diagnosis, and contacts the appropriate people in the ED and

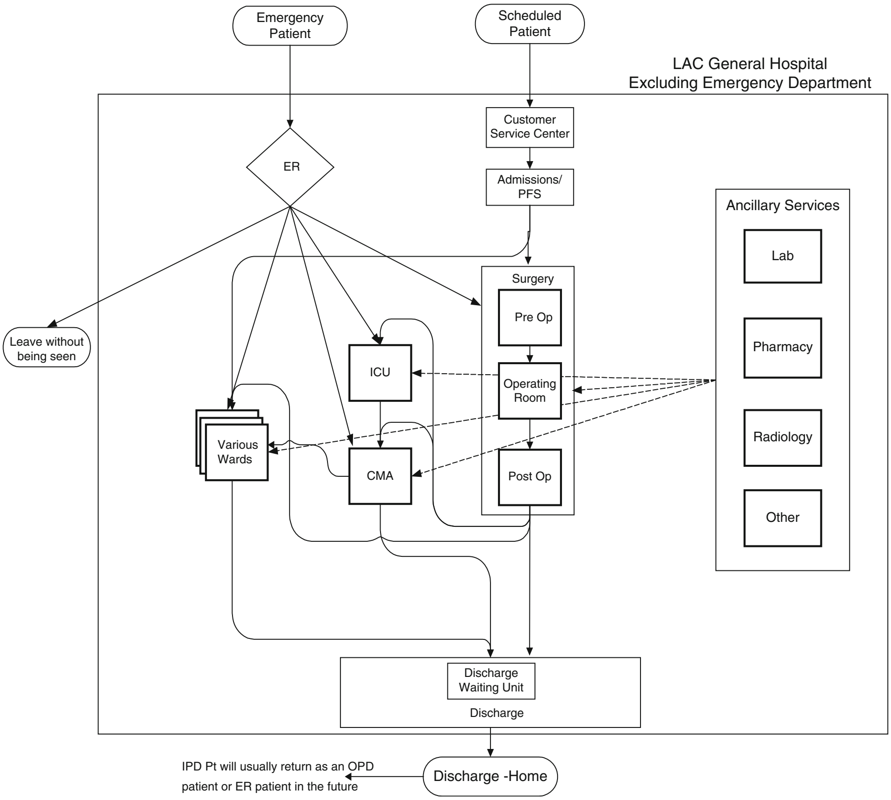


Fig. 1.5 Intra hospital patient flow

the wards as soon as a bed is made available. Information on bed availability and forthcoming discharges comes from the information system and informal communications among staff. The BCU staff has difficulty getting an overall picture of bed availability because they are not always told as soon as a bed is freed up, and because delays in housekeeping can hold up bed assignments.

An experienced patient flow manager, who is an Assistant Nursing Director, and other nurses who serve as census coordinators, walk the floors to assess the availability of beds resulting from discharging patients. The patient flow manager keeps a record of which patients are moved where and when. The manager walks through each ward and notes empty beds and potential discharges. After her rounds, she notifies the BCU of the results. She also calls up the Discharge Waiting Unit (DWU) to tell them whether any inpatients will be coming into their unit. This way she has discretion to send inpatients to the DWU and, in turn, make more beds available.

2.1.3 Discharge/Bed Preparation

Given the high percentage of beds occupied at all times, it is important to discharge patients as soon as possible. The GH instituted several efforts to improve this aspect of patient flow, including creating committees charged with removing bottlenecks, implementing buffers between processes to reduce queuing, creating a housekeeping group specialized for making the rooms ready for new patients, and creating a discharge waiting room. The nursing staff records the availability of a bed within the information system, which provides automatic notification to housekeeping that a room needs to be cleaned, which then assigns cleaning staff. However, delays sometimes force nurses to clean rooms, and because there is some ambiguity as to when the cleaning is completed, there can be confusion as to when a new patient can be assigned to a bed.

2.1.4 Staffing

Nursing administration uses the number of patients and their levels of acuity in the various wards as a basis for its staffing decisions. Other than nursing, hospital staffing levels do not fluctuate greatly from day-to-day and are based on budgetary decisions concerning each functional area. A common issue is unfilled positions, as well as absenteeism.

The nursing organization is large; there were about 1,875 RNs and LVNs and a total of about 2,700 nurses at the time of this study. The director of nursing has a staffing office where they accumulate the staffing information. Staffing, vacations, and rotation are determined at the Nurse Manager level. The nursing department uses a computer system, which reports and records the daily staffing, but actual work schedules are largely planned manually. Staffing levels are based on an acuity system. A computer system processes acuity information on the mix of patients in each area and determines the desired level of staffing. Schedules for nurses are generally posted 6 weeks in advance but are adjusted more frequently. The staffing levels are fairly stable, but the number of nurses available changes often—and is a major challenge. To manage the staff to patient ratio, which is monitored closely, there are several available options: (1) Overtime for nurses currently working, (2) Registry (the use of outside contractors), (3) Pool nurses, shared among departments, which is limited, (4) Close beds to reduce the requirement for staff.

2.1.5 Admission/Registration/PFS

A patient's entry to the hospital's data is recorded at a number of points for inpatients and outpatients. Generally this represents the entry of the patient into the information system to record a visit. Each patient is assigned a unique ID number (an "MRUN" number) on the initial visit that is used for all subsequent visits. Each visit of the patient is recorded as an “Account” in the information system.

Patient Financial Services (PFS) personnel are responsible for gathering information about the patient for reimbursement. They enter demographic data about a patient or “register” them in the system. If they detect an existing MRUN number, they locate it. If not, they generate a new one for the patient. PFS also explains financial obligations to patients.

PFS responsibilities are time consuming. They must get the general consent form completed for each patient, distribute brochures, work with triage nurses to complete patient registration, initiate the patient chart, complete a limited financial screening, and provide insurance information. For outpatient areas, certain patients meet with PFS. They can determine PFS need by looking at the backside of the patient’s ID card that they have received from a previous visit. Scheduled admissions and day surgery patients are financially screened by PFS prior to coming for surgery.

2.1.6 Transportation

LA County supplies patient transportation between the hospital and the outside when the patients cannot provide it themselves. An internal transport unit sometimes supplies patient transportation by wheelchair or gurney within the hospital, but often nursing moves the patient when the transport group is delayed. Transportation delays are common due to staffing and the difficulties in navigating through a crowded hospital. Elevator waits can also be long. These delays create a cascading effect, and have a significant impact on surgery, radiology, and bed utilization, as resources can be left idle while waiting for patients, or because appointment systems cannot be followed due to delays.

2.1.7 Medical Records

Hospital Information Management (HIM) is responsible for storage and access to patient medical records (charts). They stored a large volume of paper records in the hospital basement and used long-term storage off site.

HIM has several functions: maintenance of medical records for each patient, assuring their completeness, copying the chart for several types of requests, and alerting clinicians for missing chart elements. They also make sure that the records follow professional and legal guidelines. They stored about 1,000,000 charts at the hospital (basement) and 2,500,000 at an offsite warehouse. More information is stored in the paper chart than in the information system. The chart has the patient’s laboratory results, radiology results, X-ray results, etc. These are all bound in the patient’s chart folder.

The Central Discharge Unit (CDU) clerk checks through the information system three times a day for discharges. After the discharges have been identified, a clerk from the CDU visits the wards to pick up the patients' charts. This is done between 3:30 p.m. and 4:00 p.m. and between 7:30 p.m. and 8:00 p.m. The CDU keeps track of the records of the inpatients, scheduled admissions, and day surgeries.

All charts go through a "coding" process where data are added to the record for statistical and research purposes. The next stage is "abstracting" information, such as length of state, the Rx, tests ordered, physician attending, etc. This is done within 24 h of the time the chart is brought to the CDU. If a chart is deficient, then it is stacked on the shelves in the CDU for the physician to complete. Also, the HIM staff visits the wards three times a week to identify data deficiencies before a patient is discharged, reducing the time wasted in waiting for a physician to complete the chart.

Patients who have transferred from other hospitals have a copy of their chart brought to LAC GH. This copy is included in the chart at the GH. When they return to their initial clinic or hospital, GH sends a copy of their new GH records along with the patient. They also received about 3,000–3,500 requests a month from patients requesting a copy of their records.

2.2 Flows for Key Departments

At this point we turn to patient flows within individual departments. Several key departments are used for illustration: (1) Emergency, (2) Radiology, (3) Pharmacy, (4) Laboratory, and (5) Surgery.

2.2.1 Emergency Department

The ED was organized into three areas: Major medical/trauma (Room 1350), Minor Medical/Trauma (Room 1060), and ED Walk-in (Room 1050). The room numbers were used as the departmental identifier. Each had a separate physical area (Figs. 1.6 and 1.7). Each also served a different set of patients.

When a patient enters the ED, he or she first sees a triage nurse, who determines the severity of the patient's condition. Immediately afterwards, the PFS assigns an MRUN number if the patient is new to the hospital. With the help of the patient's name, date of birth and mother's maiden name, PFS checks in the information system for previous medical records. In some cases a duplicate ID is created, but this is rare. Patients generally have previously received an ID card, which shows their MRUN number and their financial situation regarding hospital reimbursement.

The ED is extremely busy and crowded, and suffers long waits. This is due in part to waits for admission to inpatient beds. When inpatient beds are unavailable the ED patients often must wait in ED beds until they can be moved ("borders"). Also, processes are slowed in the ED due to limited accessibility of certain ancillary services. Radiology, for example, is on a different floor and ED patients must

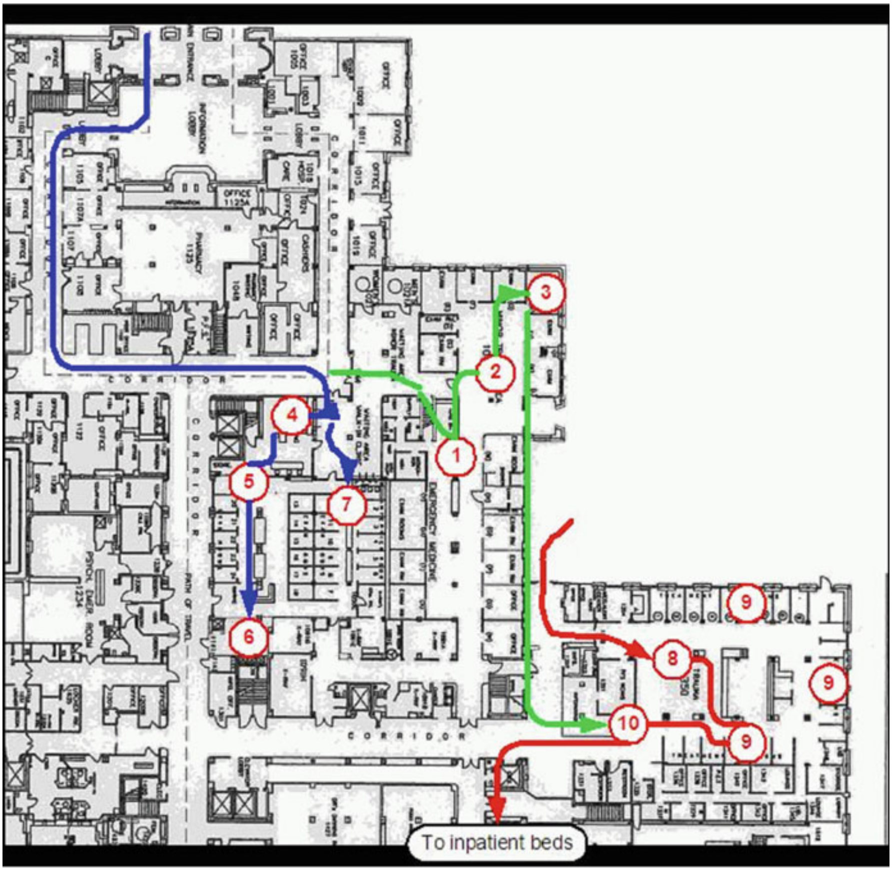


Fig. 1.6 Department of emergency medicine patient areas

sometimes be moved up to that floor for diagnostic services and then moved back down to the ED.

After the patient's condition has been stabilized, each patient is assigned a PFS worker who asks questions, enters data into the information system, takes printouts, and puts this info next to the patient's bed for the doctor to see. If the patient needs to be admitted into an inpatient bed, he/she is seen again by PFS. At the end of the ED visit, or if additional service is needed, such as admission to an inpatient bed, then the patient receives "financial screening" from PFS.

The 1350 area includes emergency admission on a 24-h, 7-day-a-week, basis. It includes trauma care and services for other acute patients. When the other two ER areas are closed, it serves ambulatory patients as well.

Trauma patients are the most severe of the emergency patients that come into 1350. Most are victims of automobile accidents or violent crimes, such as gunshots. Trauma represented about 7,500 patients per year at the time of this study. Virtually all trauma patients eventually become inpatients and represent a significant

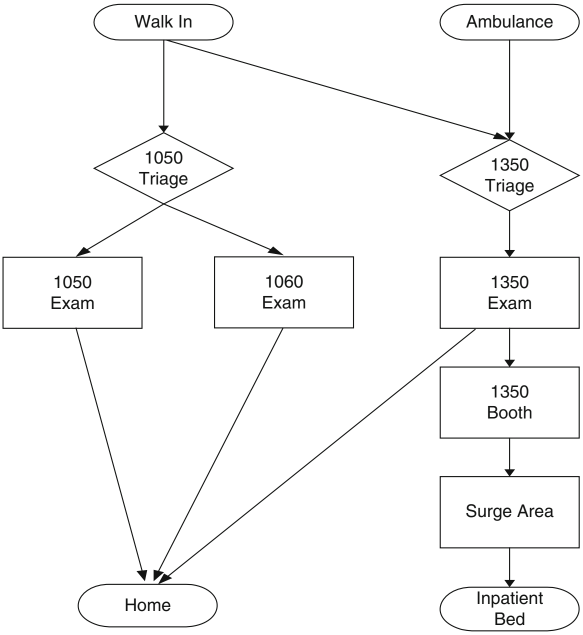


Fig. 1.7 Typical patient flows in ER

proportion of the total inpatient population. The 1350 ER had one of the largest trauma centers in the USA. The trauma staff included about eight faculty, nine residents, and five physician assistants.

In the center of the 1350 area is the C-booth that serves critical trauma patients (Fig. 1.8). The C-booth is a central resuscitation area that can hold four critical patients at one time. Surrounding the C-booth area are about 22 patient care booths that are used for less critical patients. When necessary, additional patients are handled on gurneys and chairs in the same physical area. Surges in demand occur, perhaps several days each month, when the capacity of the booths is exceeded.

The 1060 area includes minor trauma care on an ambulatory basis. This area is a sort of "mini ER" that focuses on skin and bone emergencies, lacerations, boils, fractures, sutures, etc. Patients cannot come directly into 1060 as they need to be referred from either 1050 or 1350. These patients have undergone triage in one of these two areas. However, registration takes place in 1060. There is a special hold area in 1060 for ambulatory patients. From here, patients either go home or go to an inpatient bed.

1060 handles a large volume throughout the year since it is open 24 × 7 while 1050 is not. It handled the most ED patients: 150–200 patients per day at the time of this study and about 15–20 %, or 20–30 patients, go to inpatient beds.

The 1050 area is for walk-in patients that believe that they need immediate clinical help (Fig. 1.9). Everyone goes through a common initial meeting with an RN for triage at a window and then may be seen right away or asked to wait or sent

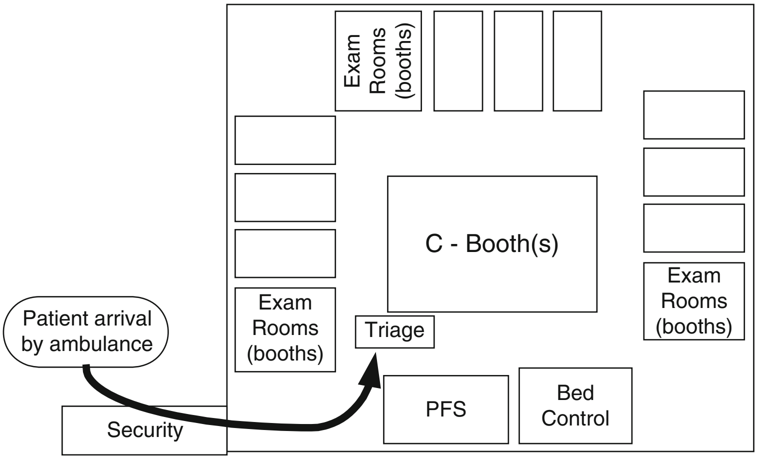


Fig. 1.8 1350 Area

to 1350 if very critical. This initial window triage is followed by an additional triage with a nurse. If required, the patient will be moved to a booth in 1050A for a meeting with a doctor who will diagnose the patient and provide orders. Some patients are admitted from 1050 to inpatient beds, but this is relatively rare.

There are about 15 individual examination booths in 1050. Doctors see two or more patients per hour, and patients spend about 10–15 min per booth visit. A wait can be as long as 8 h. A standard triage form is used where criticality is determined as one of five levels, which defines the path of care to be provided.

The 1050 patient volume was about 150 patients per day. About 10–15 % left without being seen. About 1 % of the 1050 arrivals were transferred to other inpatient services on the same day, about 10 % were admitted to an inpatient bed and the remainder went home after being seen, some with an appointment for a return. If a 1050 patient cannot be seen by the end of the day, they are sent to 1350.

Patients are sent to exam booths to meet with a doctor and doctors select which patient to see next. Doctors can select from the queue if they wish. A separate area (1050B) is for return and/or “Fast Track” patients. It’s a type of primary care clinic where a patient returns for ongoing care or medication, such as follow-up for a broken limb. About 5 % of 1050B patients are for medication refills. A Physician Assistant (PA), rather than an MD, staffs 1050B.

2.2.2 Radiology

Clinicians generate the patient flow into the radiology department. The clinician’s orders are delivered to the radiology department much before the patient arrives. The various facilities in the radiology department include diagnostic X-ray, nuclear medicine, diagnostic ultrasound, off-site MRI, CT scan, various interventions, and

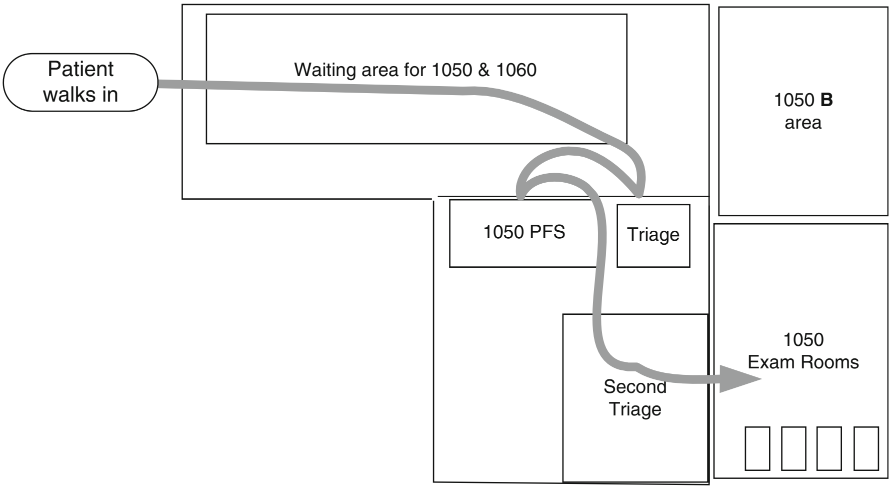


Fig. 1.9 1050 and 150B Areas

more. In addition to the services on the third floor, three X-ray rooms are allotted to the ground floor ED. There is also a portable X-ray machine placed in the 1350 ER, which can X-ray the body parts that are most likely to need X-rays. For specialized X-rays, patients are taken to the third floor. OPD also has its own radiology department.

Separate scheduling systems are used for each of the radiology services. Outpatient scheduling is separate from inpatient. Priority is given to ED patients then outpatients. Since inpatients are available for a greater time period they are given a lower priority. To avoid long delays arising from the radiology department waiting for an inpatient to be brought in, they call more inpatients than required. The inpatients are served on a first-come, first-served basis. The radiology waiting area is open and the inpatients and outpatients must wait along with the jail patients.

Figure 1.8 provides a detailed process map for the flow of inpatients through radiology (separate process maps have been created for outpatients and ED patients). Key steps in the chart include: (1) writing physician order, (2) review of request within radiology department, and review of available slots, (3) placing order for patient transport, (4) waiting for service upon arrival in radiology, (5) completion of scan, (6) transport of patient back to a ward, and (7) review of results. Through review of the full set of radiology processes, we identified a set of bottlenecks and problems, which we are in the process of resolving:

- Outpatients arrive hours before their appointment time, hoping to be served earlier. Thus, waiting rooms are full and the patient spends a longer time at the hospital.

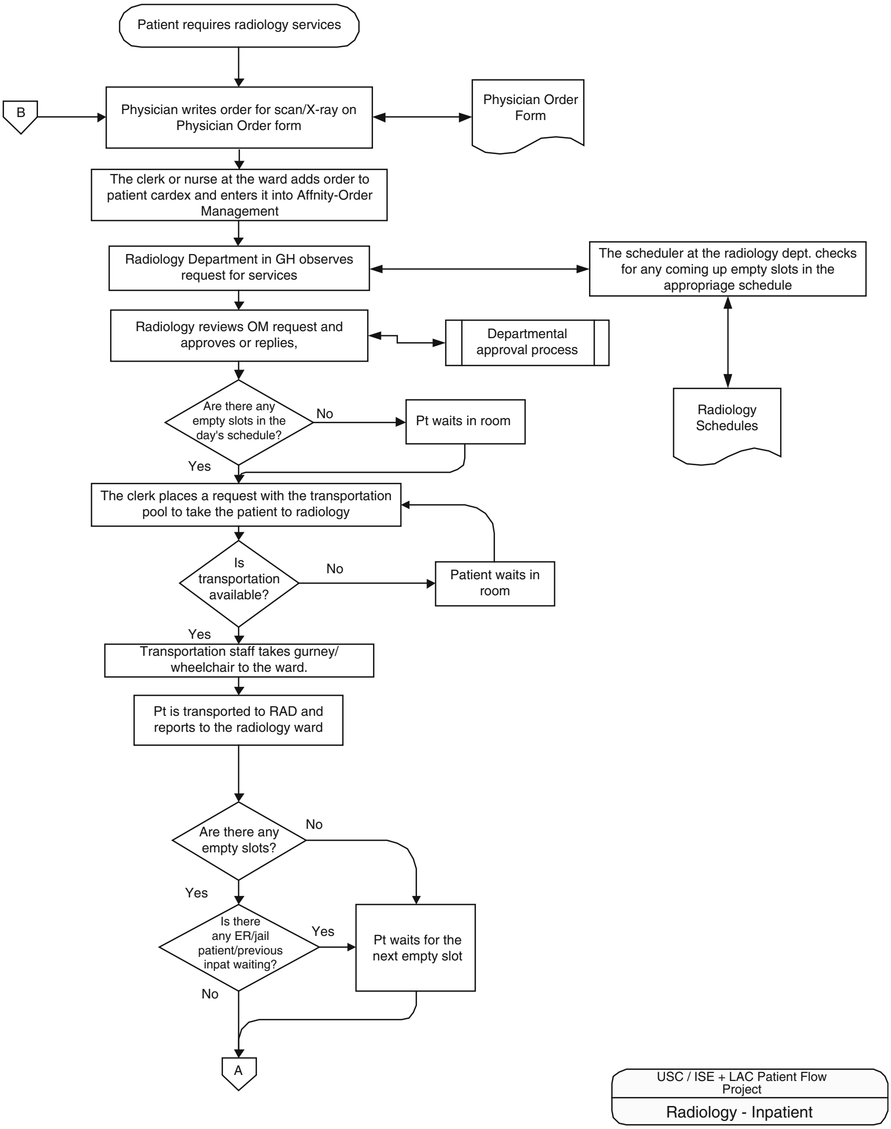
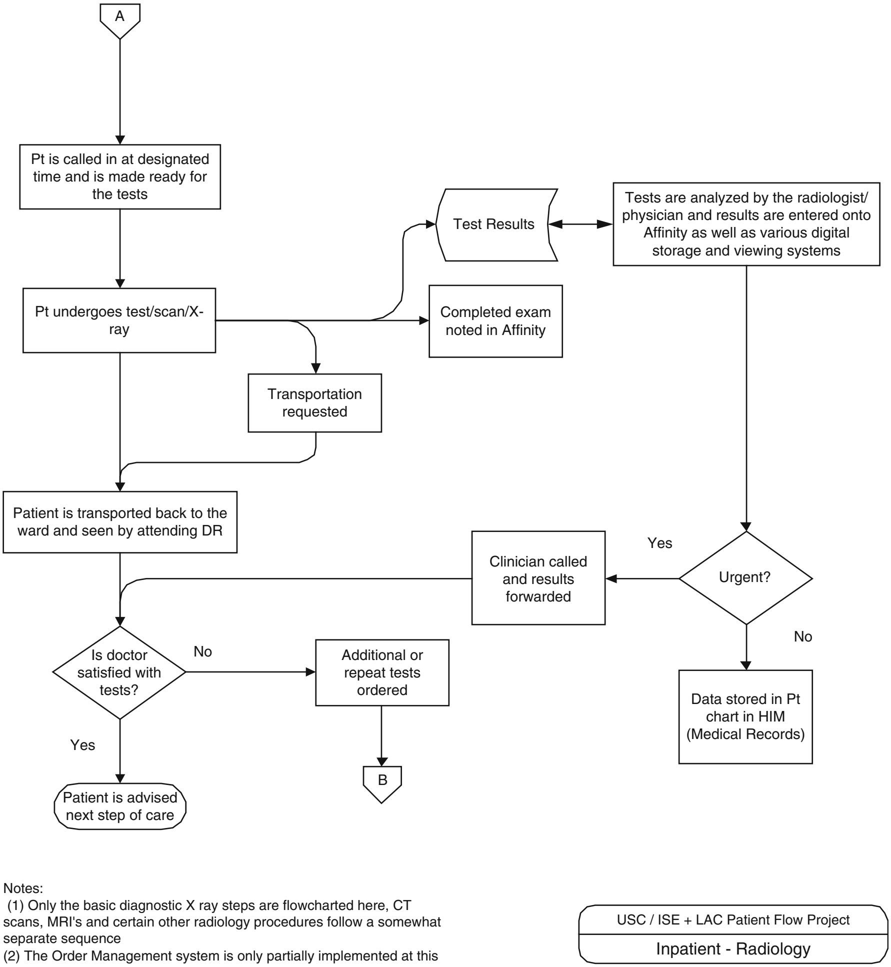

Patient requires radiology services
B
Physician writes order for scan/X-ray on Physician Order form
Physician Order Form
The clerk or nurse at the ward adds order to patient cardex and enters it into Affnity-Order Management
Radiology Department in GH observes request for services
The scheduler at the radiology dept. checks for any coming up empty slots in the appropriage schedule
Radiology reviews OM request and approves or replies,
Departmental approval process
Radiology Schedules
Are there any empty slots in the day's schedule?
No
Pt waits in room
Yes
The clerk places a request with the transportation pool to take the patient to radiology
Is transportation available?
No
Patient waits in room
Yes
Transportation staff takes gurney/ wheelchair to the ward.
Pt is transported to RAD and reports to the radiology ward
Are there any empty slots?
No
Yes
Is there any ER/jail patient/previous inpat waiting?
Yes
Pt waits for the next empty slot
No
A
USC / ISE + LAC Patient Flow Project
Radiology - Inpatient

Fig. 1.10 (continued)

- Rather than send orders, physicians come to the third floor (where the radiology department is located) to hand over the paper requisitions and check with the radiologist whether all the required information is present on the requisition sheet.
- Since there are many residents, considerable time is spent in teaching tasks that slow the availability of results.

- Staffing shortages and insufficiently experienced staff create idle equipment, even when there are patients waiting.
- Time between a doctor's order for a test and receipt of the results is lengthy.
- Patients are sometimes admitted to wards to gain priority over outpatients for tests

We have found in radiology, and in other departments, that creating a process map, such as the one in Fig. 1.10, helps reveal the bottlenecks to all participants, and leads to creative solutions.


A
Pt is called in at designated time and is made ready for the tests
Test Results
Tests are analyzed by the radiologist/ physician and results are entered onto Affinity as well as various digital storage and viewing systems
Pt undergoes test/scan/X-ray
Completed exam noted in Affinity
Transportation requested
Patient is transported back to the ward and seen by attending DR
Yes
Clinician called and results forwarded
Urgent?
No
Is doctor satisfied with tests?
No
Additional or repeat tests ordered
Data stored in Pt chart in HIM (Medical Records)
Yes
Patient is advised next step of care
B
USC / ISE + LAC Patient Flow Project
Inpatient - Radiology

Notes:
(1) Only the basic diagnostic X ray steps are flowcharted here, CT scans, MRI's and certain other radiology procedures follow a somewhat separate sequence
(2) The Order Management system is only partially implemented at this time. It will eliminate paper orders.

Fig. 1.10 (**a**) Process map for inpatient radiology, part 1. (**b**) Process map for inpatient radiology, part 2

2.2.3 Pharmacy

Pharmacy services are provided to inpatients, outpatients, and patient discharges. The GH has a first floor pharmacy, which operates 24 × 7; a 24 × 7 eighth floor pharmacy, which is for inpatients, and an outpatient pharmacy in the OPD building. The first floor pharmacy also serves outpatients when the OPD pharmacy is closed and provides medications for patients when they are discharged.

The first floor pharmacy filled about 900 prescriptions per day, about 700 during the day shift, at the time of this study. They received some prescriptions by fax, but most were on paper. A ward nurse, a patient or a patient's family, brings prescriptions to the pharmacy. Refill orders are received by phone. The eighth floor pharmacy fills cassettes, which are 1 day supplies for inpatient beds. Many orders are received from wards to this pharmacy by fax.

Patients don't pay at the pharmacy in advance for their medicine. After the order is ready, patients are given a cash receipt and are sent to the cashier to pay, After payment, they come to pick up their medicines. Some are given a mail-in envelope for payment.

Pharmacy staff and other personnel noted bottlenecks and problems:

- Staffing shortages and insufficiently experienced staff.
- Waiting for medications was said to contribute to delays in discharging patients from hospital inpatient beds.
- On IPD discharge, the doctor is supposed to provide prescriptions in advance, on the day before, but they often do not write it until their morning rounds.

2.2.4 Lab

The GH lab provides a centralized service for a wide variety of tests. The primary flow is: specimens, mostly blood in tubes, are received in the lab area by pneumatic tube, hand carried to a receiving window or gathered by an outside transportation contractor to gather samples from CHC locations and various satellite locations. An initial set of steps involves receiving the material and paperwork; a second phase involves organizing the samples (for which they have automated equipment) and then doing the test itself. The tubes are generally bar coded and other information is printed as text on the tube.

The exact volume of work was said to be unknown, but in the central lab they processed about 1,000 chemistry tests a day and about 700 of other tests a day. About 200 people worked in the department, with about 80 test technicians and a number of open positions. Electronic orders from doctors were received through an order management system.

Some "outlying" lab people work in the ED and elsewhere, which do some of the receiving tasks. Lab turnaround time was targeted at 1 h for stat work (30–40 % of orders) and 4 h for routine work. Results are often delivered electronically, and paper results are sent to medical records for the patient's chart (7 days later for inpatients). Doctors must take the initiative to check lab results, which does not always occur.

2.2.5 Surgery

The hospital had about 20 surgery suites (operating rooms) on several floors of the hospital. The exact number of rooms used on any day varied with staffing availability. Surgeries are of three basic types: emergency, inpatient, and outpatient (or day surgeries). Nonemergency surgeries were scheduled 1 day in advance with a homegrown stand-alone computer system and priorities were set by various physicians responsible for their respective specialty. Surgery days are blocked out for various specialties on a 2-week rotation pattern. Thus, a room is scheduled weeks in advance for a specific type of surgery (such as "cardiac") and may not be available for that type of surgery again for days or weeks. Doctors from each specialty define the sequence of patients within their specialty. Queues for each specialty may be weeks or months in length.

On the day before each surgery, the surgeon estimates duration of surgery, which is used to schedule the next day's use of the operating rooms. Surgeries often take longer than their estimated time. As a result, the last scheduled surgery for the day often is not done on the day scheduled. This missed surgery might not be done the next day due to what has already been blocked out and the surgery may be scheduled for sometime later in the month. This practice sometimes prolongs the number of days an inpatient occupies a bed, because an inpatient must stay in the GH until their surgery is completed. In the case of outpatients, they must then go home and come back for their rescheduled surgery and repeat their pre op visit. Also, some scheduled surgery patients do not show up for their appointment. If a patient doesn't turn up at the scheduled time then another patient must be identified and prepared for surgery, which results in a delay.

The GH's 20 OR suites handled about 28–30 surgeries per day, out of which about 35 % were outpatient and 65 % inpatient. They generally used three ORs in the evening—one for red blanket patient (trauma), other two for scheduled or ED surgeries. A "white board" on the surgery floor lists pending emergency surgeries throughout the day.

Bottlenecks related to surgery include the following:

- Inpatient beds unavailable, which causes a backup of patients completing surgery.
- Scheduling which does not fully utilize the available surgery time.
- Patients for day surgeries who do not show up as expected.
- Incorrect or unavailable ancillary service results (Radiology, lab reports, Medical records).
- Staffing shortages resulting in fewer rooms or services available.

- Frequent rescheduling and bumping of surgeries for a variety of causes.
- Late start times and an early shift cutoff time.
- Slow cleanup between surgeries.
- Delays in transport service and waits for elevators.
- Allocation of rooms to specialties may not match the relative demand among specialties. Paperwork is not always available or correctly completed on time.

3 Evaluation and Improvement Strategies

We now turn to some of the methods for measuring system performance. It is important for every patient flow improvement project to develop quantitative measures of both problems and successes, to guide implementation of changes, and to create ongoing monitoring systems to make continual improvements. Unfortunately, desired data are not always collected and are frequently not presented in a meaningful form. In this section we describe some of the more important measurements for patient flow, and describe our challenges in obtaining data.

3.1 Understanding Patient Arrival Patterns

Patient arrival patterns drive systems for scheduling staff and other resources. The patterns are somewhat predictable, even for the unscheduled ED. Hospital scheduling controls most of the other arrivals. The following time-of-day graphs for the 1350 and 1060 ER areas (Figs. 1.11 and 1.12) show a strong peak early in the day (particularly 9:00 a.m. to 11:00 a.m.). The arrival pattern is somewhat different in 1350, with a peak in the evening hours, from about 5:00 p.m. until 1:00 a.m., which may reflect the severity and incidence of injuries resulting from accidents and violence.

ED timing is important because 85 % of inpatients arrive from the ED. ED arrival time has regular patterns over the typical day and week. However, all of these measures may be influenced by the ED being overly busy (possibly on diversion) and by long waiting time in the less acute area of the ED discouraging additional patient arrivals. Typically, during the course of a week, Monday and Tuesday are busiest, and Sunday is the least busy, in the total ED (Fig. 1.13). Overall inpatient volume has been decreasing during recent years with relatively flat seasonality (Fig. 1.14).

The inpatient arrival time pattern is not a very meaningful measure since other issues, such as discharge times, which are more important for performance measurement, affect it. Thus, arrivals to inpatient are more a reflection of output than input. Admission to the inpatient area depends on bed availability and thus

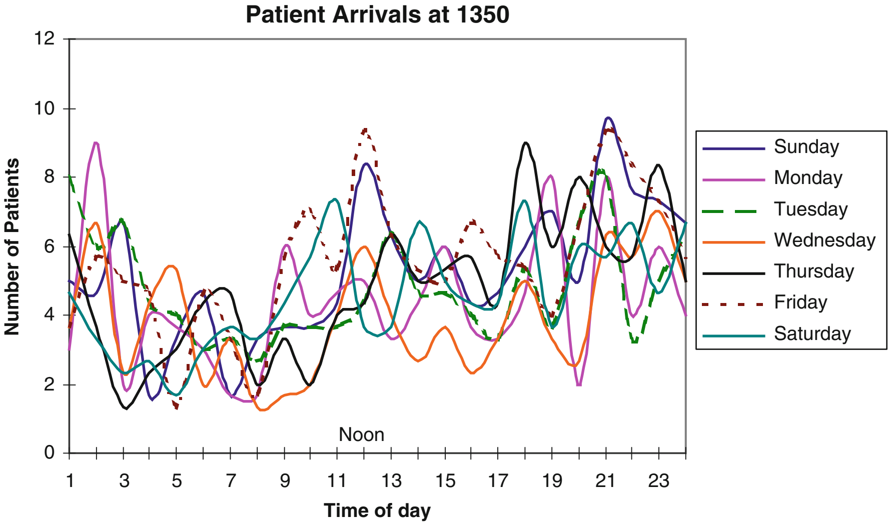


Fig. 1.11 Hourly arrivals in 1350

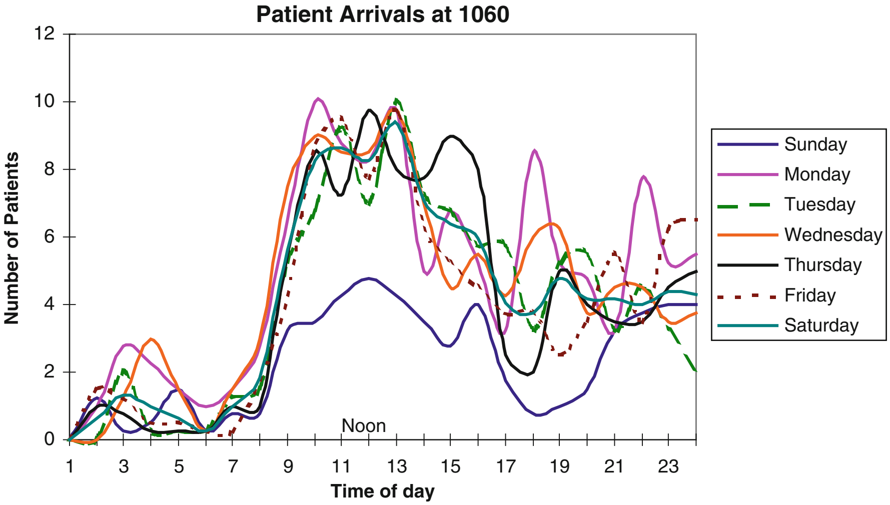


Fig. 1.12 Hourly arrivals in 1060

admission time depends on the previous patient's discharge time and the time to make the room and bed ready for a new patient. The transfer from ED to an inpatient ward may also be via a holding or surge area used when inpatient beds are unavailable.

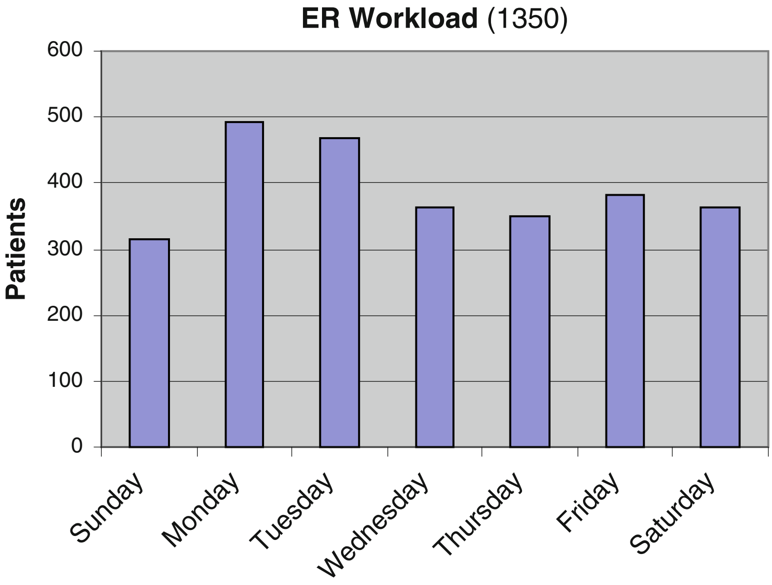


Fig. 1.13 ER by day of the week

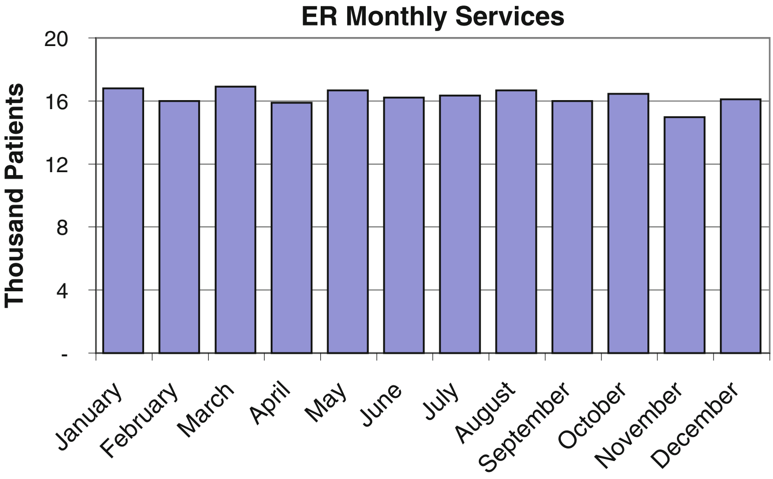


Fig. 1.14 ER by month of the year

3.2 Tracking Patient Flow by Area

LAC/USC served 205,000 patients in its ED in 2003–2004, as well as 522,000 outpatients and 23,000 other services. Sixteen thousand patients were admitted to surgical beds and 17,000 to medical beds. The hospital has detailed data for admissions into various areas, but patient flow between areas was not available. To quantify patient movement between major areas of the hospital, we developed a flow matrix by interpreting the sequence of transactions for a sample of 400 patients in April of 2004. This resulted in the input–output matrix in Table 1.3, which was extrapolated to annual flows between departments in Fig. 1.15. The system flow chart helps identify where the focus should be with respect to improving department-to-department transfer of patients. It can also lead to subsequent analysis of patient delays on an input–output basis. Last, it is an example of a nonroutine analysis, which could be imbedded in daily, weekly or monthly reporting through suitable modification of the center's MIS.

3.3 Defining Performance Measures

At this point we turn to measuring specific performance outcomes, such as waiting time, number of patients waiting, denied days, and utilization.

3.3.1 Time in System and Waiting Time

Time in system and wait time are reported in certain areas of the hospital but their accuracy is uncertain and may not provide a useful picture regarding patient flow. The basic problem is that the hospital does not time stamp events at the exact time when they occur. Event times are often recorded retrospectively (if at all), and exact times may get rounded to the nearest day, which is insufficient for tracking delays. Ideally, the time of every key event in the patient's stay should be recorded automatically as it happens, for instance with a simple bar code scan. With these data, it is possible to track waits by location and activity. It is also possible to optimize resource utilization, as staff can be alerted immediately when a resource is made available, thus eliminating idle time. We now turn to the actual data that we had available at LAC/USC.

ER Wait Time

ED wait time is an important patient flow measure since this is where many patients enter the hospital and delays can affect the clinical results. The hospital has only recently started recording the time of patient ED transactions. A problem is that the

Table 1.3 Input–output table of patient movement

	From 1050 B	1060	1350	Wards	ICU	CMA	Surgery ward	OPD/direct admissions	Scheduled admissions/day surgery
To 1050 and 1050 B		0.0 %	3.2 %						
1060	2.2 %		1.9 %						
1350	2.2 %	0.8 %		1.1 %			1.4 %		
Wards	8.6 %	5.3 %	19.6 %	11.1 %	16.7 %	75.0 %	17.4 %	62.5 %	42.9 %
ICU	0.0 %	0.8 %	1.3 %	2.2 %	0.0 %	0.0 %	0.0 %		14.3 %
CMA	0.0 %	0.0 %	1.3 %	1.1 %	33.3 %	0.0 %	0.0 %	0.0 %	0.0 %
Surgery ward	1.1 %	5.3 %	13.3 %	5.6 %	33.3 %	0.0 %	2.9 %	37.5 %	42.9 %
OPD	4.3 %	3.8 %	2.5 %	5.6 %					
Day surgery									
Discharge	81.7 %	84.2 %	57.0 %	73.3 %	16.7 %	25.0 %	78.3 %		

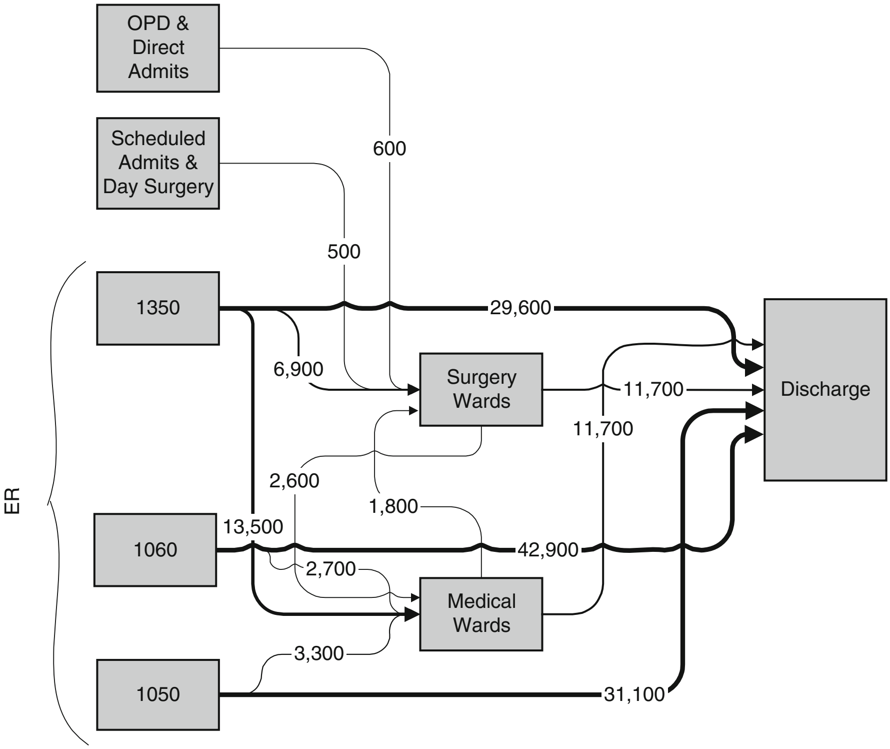


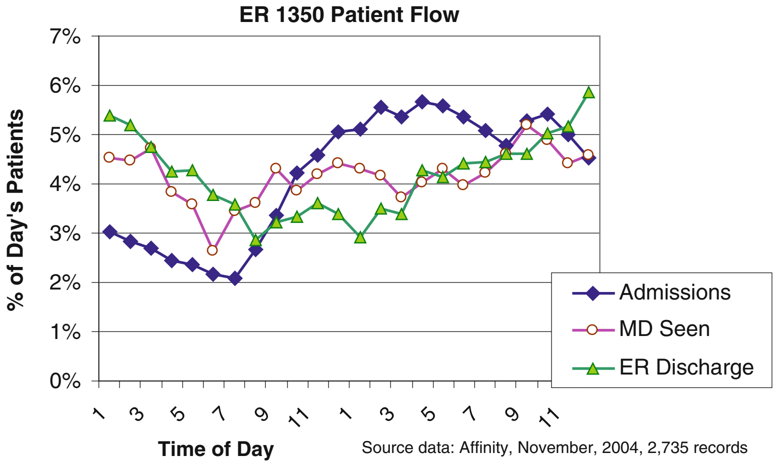
Fig. 1.15 Patient flows between areas on an annual basis

recording methodology requires additional effort by nurses, doctors and clerical staff. Since they are already very busy with a backlog of patients, it is difficult to add data recording tasks or to assure that they are done accurately.

Some information on waits can be inferred from Fig. 1.16, which shows the distribution of three events—admissions, MD seen, and discharge—over the course of a day, for the month of November, 2004 for 1350. Wait times are reflected in the differences between event times. For instance, it appears that the backlog in patients waiting to be seen starts growing at 10:00 a.m. and reaches a maximum around 9:00 p.m., at which time the backlog shrinks. This is because the admissions percentages exceed the MD seen percentages during this time of day. The queue diminishes after midnight, and continues dropping until early morning. Adding staffing between 10:00 a.m. and 9:00 p.m. would be especially beneficial in reducing waits.

Outpatient Wait Time

Wait time in the Outpatient area involves two types of waiting: waiting for an appointment to see a doctor in a clinic and the wait time that the patient experiences once arriving at the clinic.



Source data: Affinity, November, 2004, 2,735 records

Fig. 1.16 Patient movement within 1350 ER

Table 1.4 Estimated time to wait for an OPD appointment

OPD clinic	Days Average time until an appointment is available, calendar days, fall 2004
Medical	15
Orthopedic	24
Primary care, new outpatient	17
Primary care, new outpatient	9
ENT	19
Dental	0 (walk in, no wait)
Surgery—nonemergency	Over 180 days
Ophthalmology	31

Wait time for an appointment is summarized in Table 1.4. The hospital works to keep these times as short as possible. A shorter wait time will often result in a better clinical result. Also, when wait times get longer, more missed appointments can be expected and the effect of the wait is compounded.

Patient wait time for an appointment is important and is often used by other hospitals as a measure of the performance of the outpatient function. However, there are important problems related to this measure.

In some sense the wait time of the clinic is different than what might be expected. If a clinic provides poor service few patients will patronize the clinic and the wait time will be short. On the other hand, if the wait time is long, it may be an indication of good quality of service. Another factor is the practically infinite outpatient demand at LAC GH. If the community senses good quality and a reasonable level of service, there will be a surge in demand and wait time will increase.

Wait time within the clinics themselves is not measured. Opinions seem to indicate 3–4 h wait at the largest clinics (Medicine and Orthopedics), but it can vary greatly by type of clinic and time of day.

Individual patients may experience longer waiting times. Some patients arrive at the OPD building much earlier than their appointments with the expectation that this will improve their chances of being seen that day or reduce their total waiting time. In some cases early arrival at the clinic may actually result in their being seen earlier.

3.3.2 Number of Patients Waiting

The average number of patients waiting in ER, OPD and other parts of the hospital is generally unknown, as is the average wait time (though it could easily be derived if all critical events were time stamped and recorded). Based on visual observation, the number of ambulatory persons waiting in the 1050 and 1060 waiting room ranged from 50 to 150 during most of the daytime. The number waiting in 1350 was more difficult to evaluate since most of the waiting is in a bed and varies over the 24 × 7 operation. Also, it is impossible to determine which of the waiting people are patients and which are family members or friends.

The number waiting in OPD is also highly variable. Some large clinics have many patients waiting at times but they have a number of doctors working simultaneously and the wait time and service times may be less than in other clinics where far fewer patients are waiting.

Waiting in the ED area is not simply at the entrance but occurs in several locations during the ED visit. These are shown in the following map with the locations for queues measured. It should also be noted that the queues are not independent. Waiting in one queue, such as waiting for PFS, reduces the flow to subsequent processes.

To understand where delay occurs within the ED, we conducted a special study in which we hand counted the number of patient charts by location over a 2-day period. For this sample, the number of people waiting in booths in 1350 remained nearly constant from 10:00 a.m. to 3:00 p.m., as did the number of patients waiting for an inpatient bed assignment from the BCU. The waiting in 1350 for PFS, on the other hand, grew from about 5–30 patients during this period. However, this alone does not indicate that PFS is a system bottleneck, as the backup may be more the result of waiting for beds or waiting to be seen by doctors, which spills back into other areas.

Based on these observations, Fig. 1.17 was created to depict queue sizes at ten locations over time. This graph shows that the queues grow over the course of the day, primarily reflected in an increasing backlog for PFS. Queues in booths do not grow, largely because there are a finite number of booths, which limits the total number of patients who can be waiting at this stage. It is unclear whether an

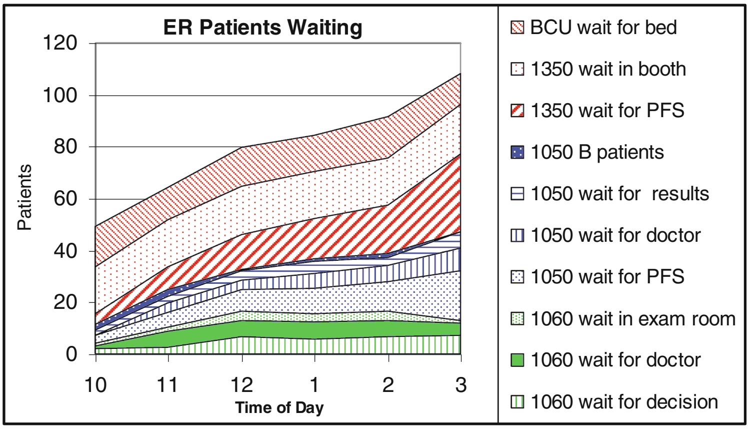


Fig. 1.17 Patient queue size by ED location

increase in PFS staffing would reduce overall delay because booths may still be the constraining bottleneck in the system.

3.3.3 Productivity and Service Time

Productivity is defined as useful output divided by input or as work completed divided by resources used, such as the resource of labor. In the case of hospitals, output can be measured in terms of patient admissions, patient days or the number of specific services done. Input can be measured by staff size in FTEs or labor hours. Relative value units (RVUs) are also a measure of the value of the work performed, and thus measure output.

An alternative but indirect measure of productivity is hospital cost per patient day. This may be helpful on an overall basis in comparing the LAC GH to other similar institutions but probably not helpful when evaluating a functional area or a specific process. Service times, such as the time for a complete outpatient or ER visit, are a measure of how well the hospital is organized to provide prompt care. They can be compared to benchmarks from other hospitals or organizations. Service times at LAC GH were generally not measured with any useful precision by the hospital's HIS. Cycle times are known on a detailed level by technicians and other staff familiar with specific procedures. For example, the average scan time of 30–45 min for an MRI was known. Of greater interest is the time for an MRI appointment, and how this compares to benchmarks and historical data to measure trends in improvement.

Some anecdotal service times were reported but rarely were they based on a true independent measurement. Such details are particularly helpful in managing patient flow if they can be compared to benchmark times or be compared to the hospital's own past performance. Tracking of such time would be helpful as a component in managing and improving patient flow.

3.3.4 Denied Days

An important financial consideration of the hospital is “denied days.” These represent days that are not reimbursed because the insurance provider does not view excessive inpatient days as appropriate. An example is inpatient days spent waiting for a surgery due to scheduling delays. This is a costly event exceeding 15 % of total days and the hospital works to avoid such a situation. Improved patient flow will inevitably reduce denied days by assuring that patient movement is appropriate and prompt. Patient flow improvement is also important for patient satisfaction and the extent to which the hospital can serve the community. Long wait times and crowding are avoided by efficient patient flow throughout the hospital.

3.3.5 Utilization (Beds, OR, Staff) and Length of Stay

Utilization of beds in the GH is generally very high. The demand for beds exceeds the supply, which is determined by how many beds the hospital can open based on the availability of staff. The supply also depends on the ability to discharge patients rapidly, as well as to rapidly prepare the bed for the next patient, as well as transport that patient. Efficient utilization also depends on achieving a good match between the types of beds available and the population of patients waiting to be admitted.

More nursing staff also means more beds can be made available. Where there is a shortage of nursing staff, the hospital uses outside services, but this is costly and inefficient, so the use is limited. If staff cannot be gathered, beds are closed and patients must wait for beds to become available. However, bed utilization can be a misleading measure, as many inpatients are queued, waiting for surgery or waiting for tests. If these processes were completed more rapidly, the need for beds would diminish, which could in turn reduce delays in the ED. They could also significantly reduce the length of stay, which is in itself in part a measure of waiting time (i.e., wait for surgery, wait for test, wait for discharge). Though it is impossible to reduce length of stay to zero (as minimum times are needed for procedures and recuperation) much of the stay is devoted to costly waiting that should be eliminated.

Utilization is important in other areas, such as surgery and radiology. While an effort is made to keep these resources scheduled, resources are often left underutilized due to difficulties in predicting service times, cancellations or late arrivals, and slack times in preparing equipment or rooms. Thus, it is possible for a department to have a long backlog while simultaneously working below its capacity. In some instances long backlogs can cause efficiency to drop, as additional attention is needed to support waiting patients and service processes become disrupted.

Thus, an objective at LAC GH is to operate with prompt and efficient patient flow to process the maximum number of patients while maintaining high quality. Utilization of resources must be considered in regard to the inevitable backlog with no likelihood of idle time.

The average length of stay at the GH was about 6.5 days. This was considered long by the hospital administration, and was in fact reduced over time. Utilization of hospital beds is high relative to other hospitals. At most times all beds are in use, in preparation to be used by the next patient, or closed due to shortage of staff.

3.3.6 Ancillary Performance Measures

In addition to patient wait time, nonlabor resources should also be measured. For example, managers of ancillary services should have accessible and relevant measures of performance of their diagnostic equipment. The lab should monitor equipment related data such as downtime (time equipment is unavailable due to maintenance, failure, etc.) and idle time. The extent of downtime should be tracked and compared to past levels, trends and to industry benchmarks. Maintenance vendors should be required to provide such data if it is not readily available from the equipment itself.

Most service areas (Radiology, Pharmacy, etc.) have industry standards available to represent typical performance numbers. Error rates, cycle time, and throughput by equipment type should be compared to past performance and to figures typical for each type of equipment.

3.4 Improving System Performance

Our study of LAC/USC resulted in numerous recommendations. We summarize some of the major findings.

3.4.1 Improvement Process

Create a team of motivated, knowledgeable, and empowered individuals to make needed changes related to patient flow. Provide them training and guidance in process mapping, analysis, and operations improvement tools. Fact based data analysis must be a key ingredient and starting point. Give teams specific goals, deadline dates and incentives. Include necessary disciplines (nurses, clerical, administration, and physicians) needed to implement changes.

3.4.2 Use Existing Data to Track Patient Flow

Patient arrivals, waiting time, service time, and other measures can be created from existing data. Patient flow management requires facts that are best gathered from processes already in place. The current information system is not ideal, but it can be mined for additional information.

3.4.3 Discharge Waiting Unit Expansion

Continue expansion of this function so that it better serves the entire hospital and maximizes the utilization of inpatient beds while considering patient satisfaction. By evaluating the discharge workflow and facts concerning patient flow it is possible to determine the optimal size of the discharge waiting unit. Also, a training program for nursing staff focusing on the capabilities of the discharge unit and the steps that can be followed by wards to expedite discharge would be very useful.

3.4.4 Transportation Level Optimized

Determine a cost effective level of internal patient transportation and implement it. This must include the necessary staff, equipment, and scheduling. It does not represent an added cost because the staff is already doing the transportation. The recommendation is to make it more efficient by analyzing the workflow and define the optimal assignment of staff.

3.4.5 Appointment/Scheduling Systems Using Simulations

Tools such as computer simulation and an optimized scheduling system will support better decisions. Scheduling requires complex trade-offs and such decisions should be based on forecasts and determination of their likely impact before they occur.

3.4.6 Bed Management System

The bed control function is particularly critical to patient flow. As noted, the system used at the time needed good information and was been criticized by users from all sides. Various parties, including the inpatient wards and the emergency room staff, regularly criticized bed assignment decisions. Thus, a strong, clear and well publicized set of rules is needed to support prompt decisions. Also, part of the improvement to bed control is better information regarding bed availability.

Discharge orders must be promptly entered into the information system and planned discharges should be frequently reported the day before the discharge is to occur.

3.4.7 Patient Tracking System and ID

By enforcing a system with a clear patient ID, costs can be avoided. Many computerized patient tracking systems, such as those the hospital is considering, have capabilities in electronic tracking. Operational data related to patient flow requires a clear, consistent, and efficient patient identification system. A variety of alternatives is available, such as barcodes and radio frequency ID, which will save operating costs and offer very useful data on patient flows.

3.4.8 Hospital Portal

Patient appointments and referrals are received by the hospital at a variety of points. This complicates the scheduling process and harms patient satisfaction through a lack of consistency and control. A single centralized point of access and a strengthened CSC would support a more efficient hospital operation and improve customer service.

4 Conclusions and Extensions

Clinicians and administrators can form collaborations to reduce health care delays. Success depends on an ability to understand health care as a system, including the many interactions between patients, clinicians, support services and other resources. Success also depends on an ability to pinpoint the bottlenecks and system failures, particularly with respect to interactions among departments as patients flow through the system.

This chapter presented process charting and performance measurement approaches, which have been used to model and evaluate patient flow delays at the LA County/USC health center. These tools can be used elsewhere, provided that hospital management is committed to improvement, and that it carries that message to its staff.

Acknowledgment We gratefully acknowledge the Los Angeles County Department of Health Services, whose financial support enabled the completion of this study. Our work relied on the close cooperation and assistance of the staff of the LA County/USC Hospital. In particular, we would like to acknowledge the dedication and support of Mr. Pete Delgado, Mr. Cecil Clark, Dr. David Altman, Ms. Irene Recendez, RN, Ms. Shiela Mallete, RN, and Ms. Behnaz Hekmatnia. We also acknowledge the support and guidance of Dr. Thomas Garthwaite and Ms. Paula Packwood of the Department of Health Services.

References

American College of Emergency Physicians, ACEP. (1999). Ambulance diversion policy statement. http://www.acep.org/Clinical—Practice-Management/Ambulance-Diversion/.

American College of Emergency Physicians, ACEP. (2002). *Responding to emergency department crowding: A guidebook for chapters*. Dallas, TX: American College of Emergency Physicians, ACEP.

Belson, D., Hall, R., Murali, P., & Dessouky, M. (2004). *Collaborative to improve patient flow at Los Angeles County/University of Southern California Hospital* (Final Report). Epstein Department of Industrial and Systems Engineering.

Bindman, A. B., Grumbach, K., Keane, D., Rauch, L., & Luce, J. M. (1991). Consequences of queueing for care at a public hospital emergency department. *Journal of the American Medical Association, 266*, 1091–1096.

Buesching, D. P., Jablonowski, A., & Vesta, E. (1985). Inappropriate emergency department visits. *Annals of Emergency Medicine, 14*, 672–676.

Derlet, R. W., & Nishio, D. A. (1990). Refusing care to patients who present to an emergency department. *Annals of Emergency Medicine, 19*, 262–267.

Derlet, R. W., & Richards, J. R. (2000). Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. *Annals of Emergency Medicine, 35*, 63–68.

Dershewitz, R. A., & Paichel, W. (1986). Patients who leave a pediatric emergency department without treatment. *Annals of Emergency Medicine, 15*, 717–720.

Greene, J. (1995). Challenges entering the ER. *Modern Healthcare*, 31–34.

Hall, R. W. (1991). *Queueing methods for services and manufacturing*. Englewood Cliffs, NJ: Prentice Hall.

Institute for Healthcare Improvement. (1996). *Reducing delays and waiting times throughout the healthcare system*. Boston, MA: IHI.

Litvak, E., Long, M. D., & Cooper, A. B. (2001). Emergency department diversion: Causes and solutions. *Academic Emergency Medicine, 8*, 1108–1110.

McCaig, L. F., & Burt, C. W. (2001). National Hospital Ambulatory Medical Care Survey: 1999 emergency department summary. *Advance Data*, (320), 1–34.

Schneider, S., Zwemer, R., & Doniger, A. (2001). New York: A decade of emergency department overcrowding. *Academic Emergency Medicine, 8*, 1044–1050.

Schull, M. J., Szalai, J.-P., & Schwartz, B. (2001). Emergency department overcrowding following systematic hospital restructuring: Trends at twenty hospitals over ten years. *Academic Emergency Medicine, 8*, 1037–1043.

Shaw, K. M., Selbst, S. M., & Gill, F. M. (1990). Indigent children who are denied care in the emergency department. *Annals of Emergency Medicine, 19*, 59–62.

Chapter 2
Interdependency of Hospital Departments and Hospital-Wide Patient Flows

Alexander Kolker

Abstract This chapter presents a quantitative analysis of patient flows for a typical hospital-wide system that consists of a set of interdependent subsystems: Emergency Department (ED), Intensive Care Unit (ICU), Operating Rooms (OR), and Inpatient Nursing units (NU) including an effect of patient readmission within 30 days of discharge. It is quantitatively demonstrated that local improvement of one subsystem (ED) does not necessarily result in performance and throughput improvement of the entire system. It is also demonstrated that local improvement targets should be aligned to each other in order to prevent unintended consequences of creating another system bottleneck, and worsening the performance of downstream units.

Keywords System flows • Simulation • Throughput

1 Introduction

Modern medicine has achieved great progress in treating individual patients. This progress is based mainly on life science (molecular genetics, biophysics, biochemistry) and development of medical devices and imaging technology. However, relatively few resources and little technical talent have been devoted to the proper functioning of the overall health care delivery as an integrated system in which access to efficient care is delivered to many thousands of patients in an economically sustainable way. According to the report published jointly by the Institute of Medicine and National Academy of Engineering, a big impact on quality, efficiency, and sustainability of the health care system can be achieved using health care delivery engineering methods (Reid et al. 2006).

A. Kolker (✉)
API Healthcare, Hartford, WI 53027, USA
e-mail: alexanderkolker@yahoo.com

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_2, © Springer Science+Business Media New York 2013

A system is a set of interdependent elements (subsystems) that form a complex whole that behaves in ways that these elements acting alone would not. Validated models of a system enable one to study the impact of alternative ways of running the system, alternative designs, different configurations and management approaches. System models enable one to experiment with systems in ways that cannot be used with real systems. A mathematical model of the system reveals important hidden and critical relationships that can be leveraged to influence the system's behavior in a desired direction.

Large systems are usually deconstructed into smaller subsystems using natural breaks in the system. The subsystems can be modeled and analyzed separately, but they should be reconnected back in a way that captures the most important interdependency between them. Goldratt and Cox (2004) state "A system of local optimums is not an optimum system at all; it is a very inefficient system." Similarly, Lefcowitz (2007) summarized "...maximization of the output of the various subsystems should not be confused with maximizing the final output of the overall system." Thus, analysis of a complex system is usually incomplete and can be misleading without taking into account subsystems' interdependencies.

An insight that systems behave differently than a combination of their standalone independent components is a fundamental management principle. A summary of other fundamental management principles is presented by Kolker (2012) and Hopp and Lovejoy (2013). The latter authors note "To qualify as a principle an insight must be both highly general (applicable to many settings) and stable (relevant now and in the future)... Overlooking the things that can be captured as principles can lead to fundamental errors. Hence, understanding management principles is extremely valuable as a starting point for managing operations."

The objective of this chapter is to present a quantitative illustration of the mentioned above fundamental principle using, as an example, patient flow for a typical hospital-wide system. The system consists of a set of interdependent subsystems: Emergency Department (ED), Intensive Care Unit (ICU), Operating Rooms (OR), and Inpatient Nursing units (NU). An effect of patient readmission within 30 days of discharge is also included in the system's overall patient flow. It will be quantitatively demonstrated that ED improvement targets (local improvements) should be aligned with capacity of the downstream units to handle increased patient flow out of ED in order to prevent unintended consequences of creating another system bottleneck, and worsening the performance of downstream units. This type of problem is a particular case of dynamic supply and demand balance (Kolker 2012).

Three basic components should be accounted for in these types of problems: (1) the number of patients (or, generally, any entities) entering the system at any point of time; (2) the number of entities leaving the system at any point of time after spending some variable time in the system, and (3) the limited capacity of the system that restricts the flow of entities through the system. All three components affect the flow of entities that the system can handle. A lack of proper balance between these components results in the system over-flow, bottlenecks or, sometimes, underutilization. It is widely acknowledged that the most powerful and versatile methodology for quantitative analysis of the proper balance and dynamic variability in complex systems is discrete event simulation (DES). This methodology is used in this chapter for the analysis of a hospital-wide patient flow.

One notable example of a model that describes patient flow through a large hospital was presented by Cochran and Bharti (2006). Their model of a 400-bed hospital included surgical case, emergency and direct admissions. Also, they included blocking of patients caused by the finite bed capacities of each unit, two classes of patients (emergency and regular), and probability distributions that varied with the time of day and the day of the week. The authors proceeded to develop the optimal bed allocation to balance patient load. They also showed how blocking could be decreased if elective procedures were scheduled during off-peak times.

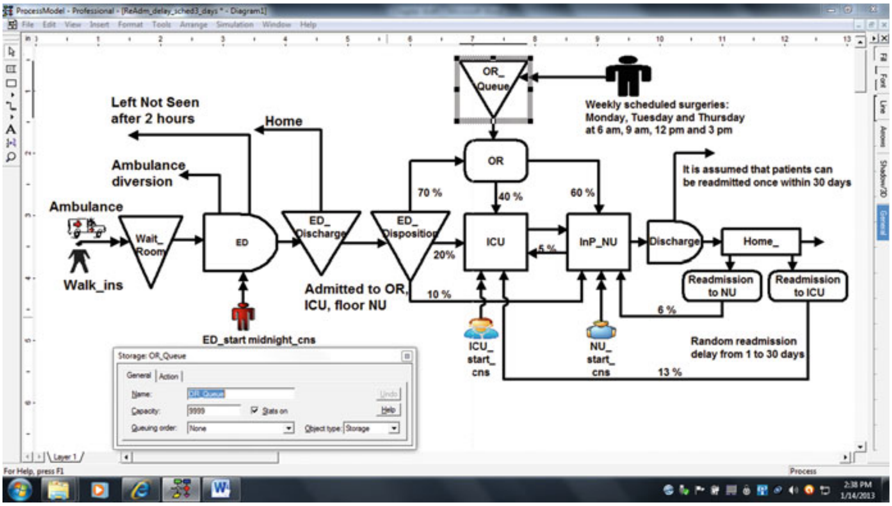
The remainder of this chapter will illustrate principles of patient flow across a hospital system by simulating a representative case-study hospital. The simulations will demonstrate dependencies between subsystems and their impact on overall system performance.

2 Hospital System Description

This section introduces a case-study hospital system that will later be simulated under different conditions. The hospital system is a tertiary referral center and a primary teaching community hospital for Southeast Wisconsin. It was chosen to represent a typical hospital system that includes the following interdependent high-level subsystems: (1) subsystem 1—Emergency Department (ED), overall bed capacity of 25 beds; (2) subsystem 2—Intensive Care Unit (ICU), overall bed capacity of 49 beds; (3) subsystem 3—Operating Rooms (OR), overall room capacity of 6 rooms; (4) subsystem 4—Inpatient Nursing units (NU), overall bed capacity of 360 beds. A high-level flow map (layout) of the entire hospital system is shown in Fig. 2.1.

Patients transported into the ED by ambulance (~18 %) and walk-in patients (~82 %) form an ED input flow. Two months arrivals patient volume (total 8,411 patients) is included. Some patients are treated, stabilized and released home (~74 %). ED patients admitted into the hospital (~26 %) form an inpatient input flow into the ICU, OR and/or NU. We assume in our simulation that patients waiting longer than 2 h in the ED waiting room leave the ED without being seen (LNS: lost-not-seen patients). About 70 % of admitted patients are taken into operating rooms (OR) for emergency surgery, about 20 % of admitted patients move into the ICU, and about 10 % of patients are admitted from ED into the inpatient nursing units (NU).

A diversion status is declared when ED, OR, ICU, or NU are at full bed capacity. The unit diversion is defined here as the percentage of operational time when the unit is at full bed capacity and can no longer accept new patients.


ProcessModel - Professional - [ReAdm_delay_sched3_days * - Diagram1]
File Edit View Insert Format Tools Arrange Simulation Window Help
OR_ Queue
Weekly scheduled surgeries: Monday, Tuesday and Thursday at 6 am, 9 am, 12 pm and 3 pm
Left Not Seen after 2 hours
Home
OR
It is assumed that patients can be readmitted once within 30 days
Ambulance diversion
70 %
40 %
60 %
Ambulance
Wait_ Room
ED
ED_ Discharge
ED_ Disposition
20%
ICU
5 %
InP_NU
Discharge
Home_
Walk_ins
Admitted to OR, ICU, floor NU
10 %
Readmission to NU
Readmission to ICU
6 %
ED_start midnight_cns
ICU_ start_ cns
NU_ start_ cns
Random readmission delay from 1 to 30 days
13 %
Layer 1
For Help, press F1
Process
2:38 PM 1/14/2013

Storage: OR_Queue

General | Action

Name: OR_Queue

Capacity: 9999 ☑ Stats on

Queuing order: None Object type: Storage

Fig. 2.1 Layout of the high-level simulation model of patient flow for a typical hospital system

About 40 % of postsurgical patients are admitted from OR into the ICU (direct ICU admission), while 60 % are admitted into the inpatient NU. However, some patients (about 5 %) are readmitted from the NU back to the ICU (indirect ICU admission).

The flow map includes 30-days of patient readmission feedback loops. These loops represent a uniformly distributed random delay in the range from 1 day to 30 days. It was reported that almost one-fifth (19.6 %) of patients nationwide who had been discharged from a hospital were rehospitalized within 30 days (MPAC 2007; Jencks et al. 2009). This case study system assumes 13 % readmission into ICU and 6 % readmission into inpatient nursing units (NU), totaling the overall 19 % of patients readmitted within 30 days after discharge.

The performance of the hospital needs significant improvement. The ED is on ambulance diversion a large percentage of time and a significant percentage of patients left not seen (LNS). The ICU frequently does not have beds for ED patient admissions or delays admission of postsurgical patients. The Surgical Department is often at capacity, and elective surgeries are frequently rescheduled. The hospital management needs to decide on the following: what unit/department to start with for process improvement projects; what type of projects to select; and process improvement performance metrics.

Because patient crowding is most visible in the ED, the hospital believed that inadequate ED throughput capacity was an issue. One way of increasing ED throughput capacity is by reducing ED patient length of stay (ED LOS) (Hopp and Lovejoy 2013). This might be accomplished in several ways. For example, Cho et al. (2011) constructed a computerized consultation management system in the ED of a tertiary care teaching hospital and evaluated the influence of the consultation management system on ED length of stay (LOS) and the throughput process.

ED personnel selected the department and on-call physician in the specialty department using the consultation management software and activated the automatic consultation process when specialty consultation was necessary. If the treatment plan had not been registered for 3 h, all of the residents in the specific department are notified of the delay in the treatment plan with a SMS message. If an admission or discharge order had not been made in 6 h, all of the residents and faculty staff in the specific department receive SMS messages stating the delay in disposition. The authors report significant reductions of ED LOS after implementing the system: the median ED LOS decreased from 417.5 min in the pre-system period to 311.0 min in the post-system period. The automated consultation and monitoring process formalized communication between physicians in ED with high consultation and admission rates.

Wang (2012) developed a simulation model of an emergency department (ED) at a large community hospital, Central Baptist Hospital in Lexington, KY aimed at determining the most critical process for improvement in quality of care in terms of patient length of stay. The author identified that floating nurse, combining registration with triage, mandatory requirement of physician's visit within 30 min, and simultaneous reduction of operation times of some most sensitive procedures can all result in substantial LOS reduction.

Oredsson et al. (2011) have undertaken a systematic literature review to explore which interventions improve patient flow in ED (33 studies with over 800,000 patients in total were included, mostly in European hospitals). The authors concluded that fast track for patients with less severe symptoms results in shorter waiting time, shorter length of stay, and fewer patients leaving without being seen. Team triage, with a physician in the team, will probably result in shorter waiting time and shorter length of stay and most likely in fewer patients leaving without being seen. There is only limited evidence that streaming of patients into different tracks, performing laboratory analysis in the emergency department or having nurses to request certain x-rays results in shorter waiting time and length of stay.

The next section analyzes the effect of various targets LOS on throughput and ambulance diversion in the ED as a separate subsystem.

3 ED as a Separate Subsystem: Effect of Patient Length of Stay on ED Ambulance Diversion

Emergency Department (ED) ambulance diversion due to "no available beds" has become a common problem in most major hospitals in the USA. A diversion status due to "no available ED beds" is usually declared when the ED census is close to or at the ED bed capacity. An ED remains in this status until beds become available when patients are moved out of ED (discharged home, expired, or admitted into the hospital as inpatients). The percentage of time when ED is on diversion is one of the important ED performance metrics, along with the number of patients in queue in the ED waiting room, ED patient waiting time, and the percentage of patients left before they are seen (LNS). ED diversion results in low quality of care, dissatisfaction of patients and staff, and lost revenue for hospitals.

Patients' length of stay (LOS) in ED is one of most significant factors that affect the overall ED throughput and ED diversion (Blasak et al. 2003; Gunal and Pidd 2006; Miller et al. 2003; Simon and Armel 2003). There are generally two major groups of ED patients with different LOS distributions: (1) patients who are subsequently admitted as inpatients into the hospital (OR, ICU, floor nursing units), and (2) patients stabilized, treated, and discharged home without admission. Mayhew and Smith (2008) and Hopp and Lovejoy (2013) also recognized a key difference between these two groups. The latter authors note "Operational (ED) metrics can be divided into two categories: time and volume… The most basic time measure is LOS, which is usually measured separately for patients who are admitted to the hospital, patients who are kept for observation, and patients who are released." In order to effectively reduce ED diversion, the LOS of two basic patient groups should be quantitatively linked to ED diversion. Then the target LOS limits can be established based on ED patient flow analysis that significantly reduces or eliminates diversion.

Kolker (2008) provided a detailed analysis of the literature on ED LOS. One instructive article, Mayhew and Smith (2008), evaluates the consequences of a 4 h LOS limit mandated by the UK National Health Services (NHS) for the UK hospitals' Accident and Emergency Departments (A&ED). Because of significant difficulty to meet this standard, the target was later relaxed, allowing that not more than 2 % of patients could exceed 4 h LOS. However, Mayhew and Smith (2008) note that this relaxed standard was not sufficient to take the pressure of conformance from A&ED. These authors conclude "…a target should not only be demanding but that it should also fit with the grain of the work on the ground… Otherwise the target and how to achieve it becomes an end in itself." Further, "…the current target is so demanding that the integrity of reported performance is open to question." Another conclusion was "…the practicality of a single target fitting all A&ED will come under increasing strain." This work vividly illustrates the negative consequences of administratively mandated LOS targets that have not been based on the objective analysis of the patient flow and A&ED capabilities.

Another example of an administrative LOS target for ED department was the Position Statement on Emergency Department Overcrowding published by the Canadian Association of Emergency Physicians (CAEP 2007). The ED LOS benchmark suggested by CAEP was not to exceed 6 h in 95 % of cases for level 1, 2, and 3 patients. CAEP recommends the establishment of the national benchmark for total ED LOS that should be linked to objective ED performance.

Despite the considerable number of publications on ED patient flow and its variability (e.g., Carr and Roberts 2010; Jacobson et al. 2006), not much in the literature provides a practical solution for the target patient LOS: what it should be and how to establish it in order to reduce ED diversion to an acceptable low level, or to prevent diversion at all?

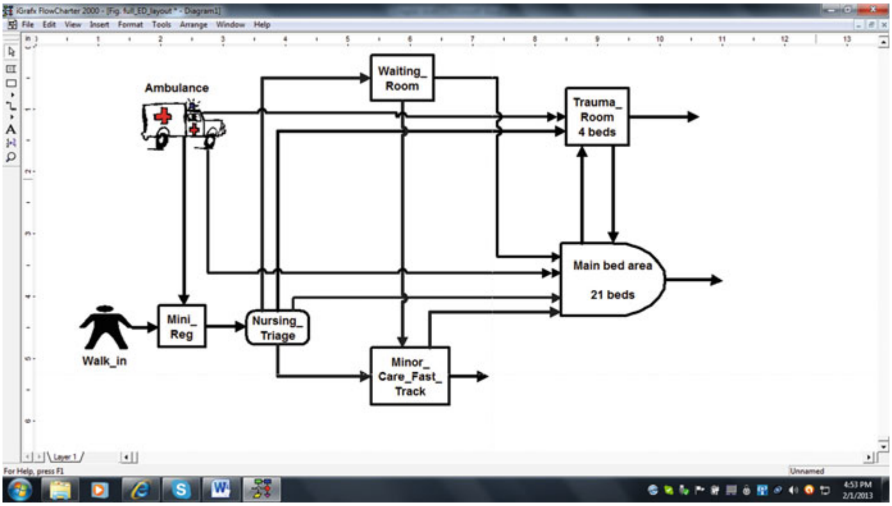


Fig. 2.2 ED structure of the study hospital. ED includes: mini-registration, nursing triage, waiting room, minor care/fast-track lane, trauma rooms, and the main patient bed area

To provide guidance on target LOS, the ED structure of the study hospital presented in Fig. 2.2 is analyzed (Kolker 2008). It includes a fast-track lane, minor care, trauma rooms, and the main patient beds area.

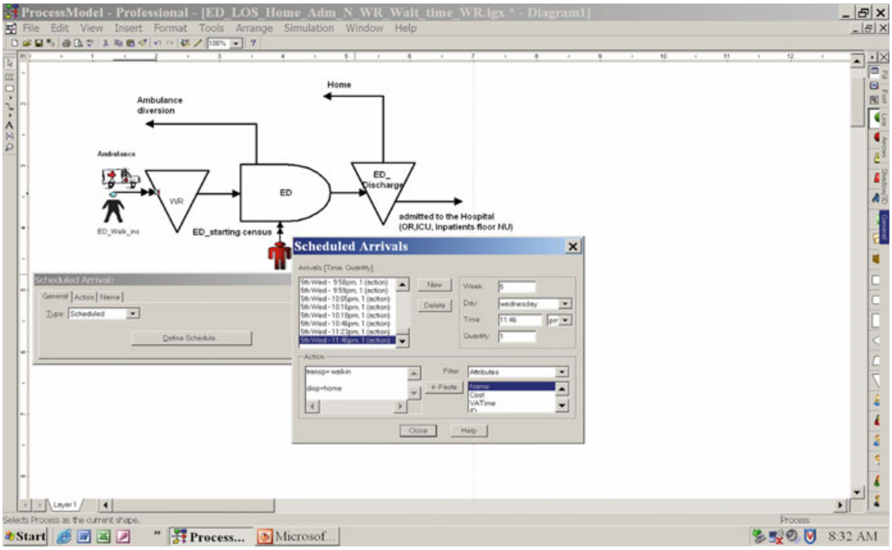
To focus on the effect of patient LOS on diversion for the entire ED, the detailed model layout was simplified in Fig. 2.3, keeping the model as simple as possible while capturing the objectives of the analysis (Law 2007).

Patients arrive into the ED by two modes of transportation: walk-in and ambulance. The week number, day of the week and arrival time characterize each patient in the arrival flow.

Discharged patients (released home or admitted as inpatients) moved out of the system according to their disposition routings. Patient flow "in and out" of the ED forms a dynamic supply and demand balance. The patient volume 8,411 for the 2-month period is included in the analysis. This patient volume is representative of subsequent months and years.

The critical element of the dynamics of the supply and demand balance is the time that patients spend in ED. This time was probabilistically fitted by continuous LOS distribution density functions, separately for admitted inpatients and discharged home patients. The ED length of stay distribution best fit for patients released home was Pearson 6 and for patients admitted to the hospital was log-logistic, as indicated in Fig. 2.4.

Because these LOS distributions represent a combination of many different steps of the patient move through the entire ED, from registration to discharge, they are simply the best analytical fit used to represent actual patient LOS data. Random numbers drawn from these distributions were used to perform multiple replications



Scheduled Arrival:

General | Action | Name

Type: Scheduled

Define Schedule...

Scheduled Arrivals

Arrivals [Time, Quantity]

5th Wed - 9:58pm, 1 (action)
5th Wed - 9:59pm, 1 (action)
5th Wed - 10:05pm, 1 (action)
5th Wed - 10:16pm, 1 (action)
5th Wed - 10:18pm, 1 (action)
5th Wed - 10:46pm, 1 (action)
5th Wed - 11:23pm, 1 (action)
5th Wed - 11:46pm, 1 (action)

New | Delete

Week: 5
Day: wednesday
Time: 11:46 pm
Quantity: 1

Action

transp= walkin

disp=home

<- Paste

Filter: Attributes

Name
Cost
VATime
ID

Close | Help

Selects Process as the current shape.

Fig. 2.3 Simplified ED structure included in the high-level hospital-wide simulation layout

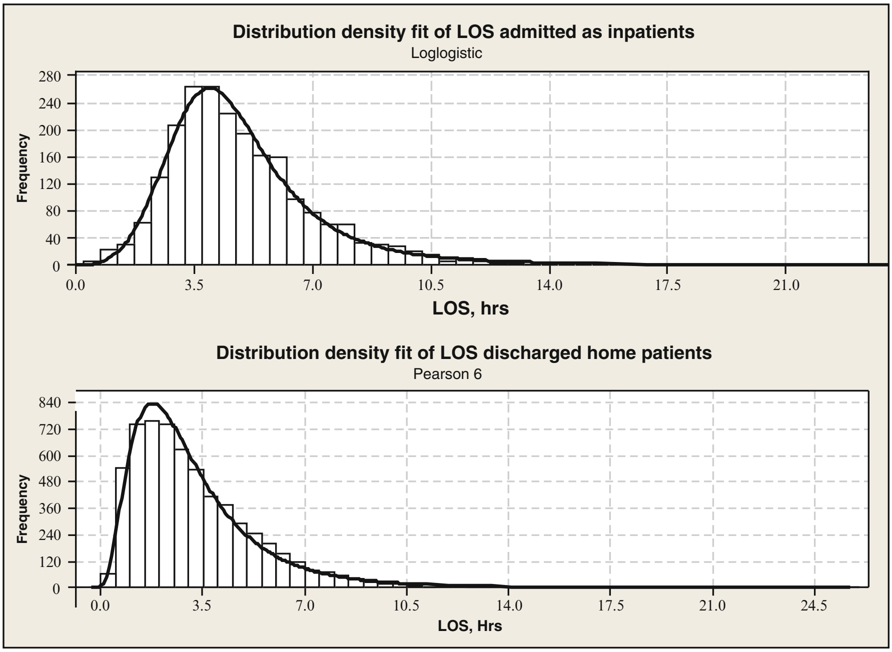


Fig. 2.4 Patient length of stay and best fit distribution density functions. *Top panel*: LOS for admitted inpatients. *Bottom panel*: LOS for discharged home patients

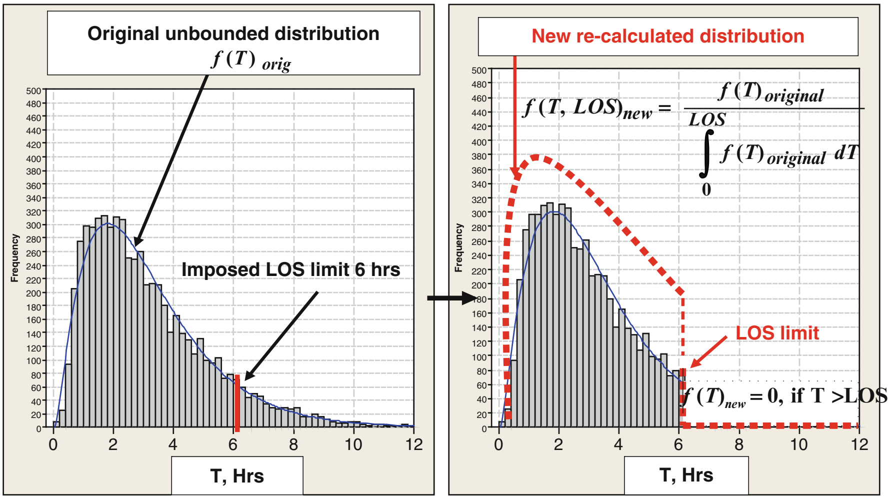


Fig. 2.5 Original LOS distribution density (*left panel*) and recalculated LOS distribution density with the imposed LOS limit (*right panel*)

using discrete event simulation (DES). Because the objective was to quantify the effect of the LOS limits (both for discharged home patients and admitted as inpatients) on the percent diversion, these limits were used as two variable simulation parameters. The original LOS distribution densities should be recalculated for each simulation scenario as functions of these parameters using the concept of conditional probability. Given the original LOS distribution density, $f(T)_{\mathrm{orig}}$, and the limiting value, $\mathrm{LOS}_{\mathrm{lim}}$, the conditional LOS distribution density function of the new random variable restricted to $\mathrm{LOS}_{\mathrm{lim}}$ is

$$f(T)_{\mathrm{new}} = \frac{f(T)_{\mathrm{orig}}}{\int\limits_{0}^{\mathrm{LOS}_{\mathrm{lim}}} f(T)_{\mathrm{orig}}\,\mathrm{d}T}, \qquad \text{if } T \text{ is less or equal to } \mathrm{LOS}_{\mathrm{lim}}$$

$$f(T)_{\mathrm{new}} = 0, \qquad \text{if } T \text{ is greater than } \mathrm{LOS}_{\mathrm{lim}}$$

This is depicted in Fig. 2.5 (right panel, dotted bold line).

The conditional distribution density is a function of both the original distribution density and the simulation parameter $\mathrm{LOS}_{\mathrm{lim}}$ (upper integration limits of the denominator integrals). These denominator integrals were first calculated and then approximated by these third order polynomials:

For discharged home patients:

If $\mathrm{LOS}_{\mathrm{lim}} \leq 10$ h, then

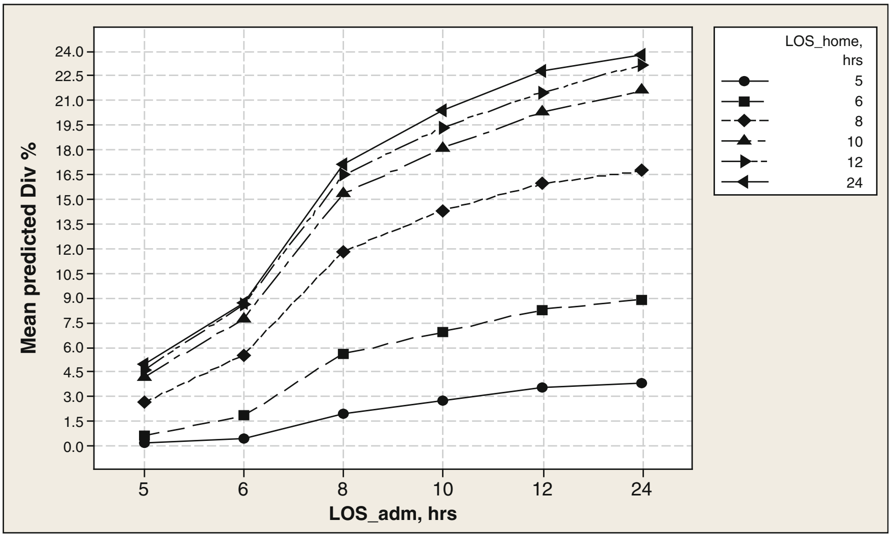


Fig. 2.6 Summary of the percentage diversion as a function of two parameters $\mathrm{LOS_{lim}}$(home) and $\mathrm{LOS_{lim}}$(adm)

$$\int_0^{\mathrm{LOS_{lim}}} f(T)_{\mathrm{orig}}\,\mathrm{d}T = -0.2909 + 0.4013 \times \mathrm{LOS_{lim}} - 0.04326 \times \mathrm{LOS_{lim}^2} + 0.001599 \times \mathrm{LOS_{lim}^3}$$

else, the integral is approximately equal to 0.997.

For patients admitted into the hospital as inpatients:

If $\mathrm{LOS_{lim}} \leq 10$ h, then

$$\int_0^{\mathrm{LOS_{lim}}} f(T)_{\mathrm{orig}}\,\mathrm{d}T = -0.7451 + 0.3738 \times \mathrm{LOS_{lim}} - 0.02188 \times \mathrm{LOS_{lim}^2} + 0.000157 \times \mathrm{LOS_{lim}^3}$$

else, the integral is approximately equal to 0.994.

The model's adequacy was checked by running the simulation of the original baseline patients' arrival. The model's predicted percent diversion (~23.7 %) and the reported percent diversion (22.5 %) are close (in the range of a few percentage points). Thus, the model captures dynamic characteristics of the ED patients' flow adequately enough to mimic the system's behavior. A summary of results is presented in the plot Fig. 2.6.

It follows from this plot that several combinations of parameters $\mathrm{LOS}_{\mathrm{lim}}$(home) and $\mathrm{LOS}_{\mathrm{lim}}$(adm) would result in a low percent diversion. For example, if $\mathrm{LOS}_{\mathrm{lim}}$(home) is 5 h (low curve) then $\mathrm{LOS}_{\mathrm{lim}}$(adm) could be about 6 h with practically negligible diversion. Notice that Clifford et al. (2008) established the goal for ED LOS 6 h for inpatients to eliminate ambulance diversion and this metric is considered exceptional if less than 5 % of patients exceed this limit. Any other combination of $\mathrm{LOS}_{\mathrm{lim}}$(home) and $\mathrm{LOS}_{\mathrm{lim}}$(adm) could be taken from the graph to estimate a corresponding expected percent diversion. Thus, simulation helped to establish a quantitative link between an expected percent diversion and the limiting values of LOS. It has also suggested reasonable targets for the upper limits $\mathrm{LOS}_{\mathrm{lim}}$(home) and $\mathrm{LOS}_{\mathrm{lim}}$(adm).

Analysis of the actual LOS pattern in the study hospital indicated that a significant percentage of ED patients stayed much longer than the LOS targets required for low or no ambulance diversion. For example, ~24 % patients of a study hospital exceeded $\mathrm{LOS}_{\mathrm{lim}}$(adm) of 6 h, and stayed up to 24 h; ~17 % of patients exceeded $\mathrm{LOS}_{\mathrm{lim}}$(home) of 5 h, and also stayed up to 24 h (Fig. 2.4). These long LOS values were a root cause of ED closure and ambulance diversion.

Established $\mathrm{LOS}_{\mathrm{lim}}$ targets could be used to better manage a daily patient flow. The actual current LOS is being tracked down and known for each individual patient. If the current LOS for the particular patient is close to the target $\mathrm{LOS}_{\mathrm{lim}}$ a corrective action should be implemented to expedite a move of this patient. Multiple factors could contribute to the looming delay over the target LOS, such as delayed lab results or X-ray/CT; consulting physician is not available; no beds are downstream on hospital floor (ICU) for admitted patients, etc. Analysis and prioritizing the contributing factors to the over-the-target LOS is an important task. Notice that the average LOS that is frequently reported as one of the ED patient flow performance metric is not adequate to manage daily patient flow.

In order to calculate the average LOS, the data should be collected retrospectively for at least a few dozen patients. Therefore, it would be too late to make corrective actions to expedite a move of the particular patient if the average LOS becomes unusually high (whatever "high" means). In contrast, if the established upper limiting LOS targets were not exceeded for the great majority of patients, it would guarantee a low ED percent diversion, and the average LOS would be much lower than the upper limiting LOS lim. Marshall et al. (2005) and de Bruin et al. (2007) also discussed the shortcomings of reporting LOS only as averages (the flaw of averages) for the skewed (long tailed) data (as wells as Costa et al. (2003) and Savage (2009).

Emergency Departments of different hospitals differ by their structure, patient mix, LOS distribution, and bed capacity. However, the overall simulation methodology presented here will be valid regardless of the particular hospital ED.

4 Intensive Care Unit (ICU) as a Separate Subsystem: ICU Diversion

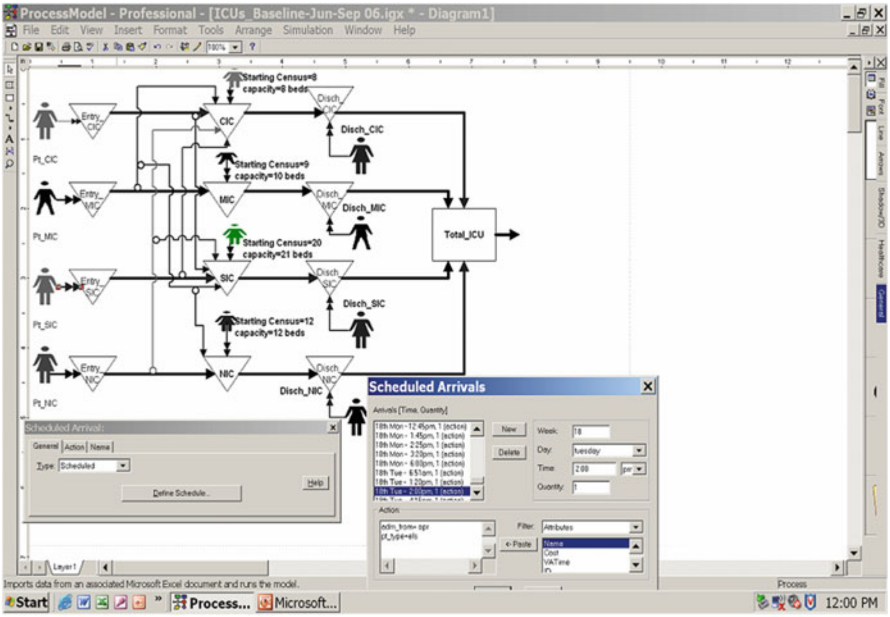
An Intensive Care Unit (ICU) is often needed for patient care. Demand for ICU beds comes from emergency, add-on and elective surgeries. Emergency and add-on surgeries are random and cannot be scheduled in advance. Elective surgeries are scheduled ahead of time. However, they are often scheduled for the daily block-time driven mostly by physician priorities. (Daily block time is the time in the operating room that is allocated to the surgeon or the group of surgeons on particular days of the week to perform a particular type of surgical service.) Usually elective surgery scheduling does not take into account the competing demand for ICU beds from the emergency and add-on cases.

Because of the limited capacity of ICU beds, a mismatch between bed availability and the flow of unscheduled patients can result in the Emergency Department (ED) diversion. This is an example of a system disconnect caused by the interdependent and competing demands among patient flows in a complex system: the upstream problem (ED closure) is created by the downstream problem (no ICU beds).

Usually two types of variability affect the system's patient flow: natural process flow variability and scheduled (artificial) flow variability (Litvak et al. 2001; Haraden et al. 2003). Patients can be admitted into an ICU from the Emergency Department (ED), other local area hospitals, inpatient nursing units, and/or operating rooms (OR). Patients admitted into ICU from ED, other local area hospitals, and inpatient nursing units are primary contributors to the natural random flow variability because the timing of these admissions is not scheduled in advance and is unpredictable.

Admissions into ICU from the OR include emergency, add-on, and elective surgeries. Elective surgeries are defined as surgeries that could be delayed safely for the patient by at least 24 h (or usually much longer). Emergency and add-on surgeries also contribute to the natural process flow variability. Because this type of variability is statistically random, it is beyond hospital control. It cannot be eliminated (or even much reduced). However, some statistical characteristics can be modeled based on data over a long period of time.

Elective surgeries that require postoperative admission into ICU contribute to the scheduled (artificial) flow variability. Elective surgery scheduling is driven by individual priorities of the surgeons and their availability, which reflects other commitments (teaching, research, etc.). This variability is usually within the hospital control, and it can be reduced or eliminated with proper management of the scheduling system. It is possible to manage the scheduling of the elective cases in a way to smooth (or to daily load level) overall patient flow variability. A daily load leveling would reduce the chances of excessive peak demand for the system's capacity and, consequently, would reduce diversion. There are quite a few publications in which the issues of smoothing surgical schedules and ICU patient flow are discussed. Kolker (2009) provided a detailed analysis of the literature.



Scheduled Arrival:

General | Action | Name

Type: Scheduled

Define Schedule...

Help

Scheduled Arrivals

Arrivals [Time, Quantity]

```
18th Mon - 12:45pm, 1 (action)
18th Mon - 1:45pm, 1 (action)
18th Mon - 2:25pm, 1 (action)
18th Mon - 3:20pm, 1 (action)
18th Mon - 6:00pm, 1 (action)
18th Tue - 6:51am, 1 (action)
18th Tue - 1:20pm, 1 (action)
18th Tue - 2:00pm, 1 (action)
```

New

Delete

Week: 18

Day: tuesday

Time: 2:00 pm

Quantity: 1

Action:

```
adm_from= opr
pt_type=els
```

Filter: Attributes

<- Paste

Name
Cost
VATime

Fig. 2.7 Layout of the ICU patient flow model

Layout of the ICU model of the study hospital is represented in Fig. 2.7. The entire ICU system includes four specialized ICU units: Cardio (CIC), bed capacity is 8; Medical (MIC), bed capacity is 10; Surgical (SIC), bed capacity is 19; and Neurological (NIC), bed capacity is 12. The total ICU bed capacity is 49. Patients admitted into each ICU unit form an arrival flow. The week number, the day of the week, and the admitting time characterize each patient in the arrival flow. Each discharged patient is also characterized by the week number, the day of the week, and the discharge time.

Patient flow "in and out" forms a dynamic supply and demand balance (supply of ICU beds and patient demand for them). ICU length of stay is assumed to be in the range from 1 day to 3 days, with 1.5 days most likely, represented by a triangle distribution. If there is no free bed at the time of admission in the particular primary ICU unit, then the patient is moved into another ICU unit using alternate type routings (depicted by the thin lines between the units, Fig. 2.7). Patient moves followed the following hospital's rules to deal with the excess capacity of the particular ICU units: (1) if no beds are available in CIC then move to SIC; (2) if no beds are available in MIC then move to CIC else move to SIC else move to NIC; (3) if no beds are available in NIC then move to CIC else SIC.

When the patient census of the ICU system hit its bed capacity limit, then an ICU diversion is declared due to "no ICU beds." In the study hospital the number of elective cases was about 21 % of all ICU admissions.

The model adequacy check was performed by comparing the predicted percent diversion for the different time periods and the actual percent diversion. It could be concluded (Kolker 2009) that the model captures dynamic characteristics of the ICU patient flow adequately (within 1–2 % from the actually reported values) to mimic the system's behavior and to compare alternative ("what-if") scenarios.

5 OR as a Separate Subsystem

An OR suite has six interchangeable operating rooms used both for ED emergency and scheduled surgeries. There are two general surgery operating rooms, one operating room each for trauma, cardiovascular, orthopedic, and neurosurgery. The operating rooms are interchangeable, so if the primary surgical room is busy, then the patient can be moved into another room if it is available. Emergency cases have higher priority than scheduled ones. Typically four OR cases are scheduled three times a week on Monday, Tuesday and Thursday at 6 am, 9 am, 12 pm and 3 pm. Usually there are no scheduled surgeries on Wednesday, Friday and weekends because surgeons have other commitments, such as teaching, research, manuscripts preparation, consulting, training, etc. However, more elective cases are occasionally added if needed and can be included in the simulation model. This artificial scheduling variability illustrates an observation (McManus et al. 2003) that "…variability is particularly high among patients undergoing scheduled surgical procedures, with variability of scheduled admissions exceeding that of emergencies." Further, "One result of this variability is a widely ranging demand for critical care services (ICU) that, in units operating at high capacity, frequently responsible for patients being placed off-service or denied access altogether."

Scheduled cases form a separate OR admissions flow, as indicated on the diagram Fig. 2.1. Elective surgery duration depends on surgical service type, such as general surgery, orthopedics, neurosurgery, etc. For model simplicity, elective surgery duration was weighted by each service percentage, and the best statistical distribution fit was identified (inverse Gaussian in this case). Emergency surgery duration was best fit by Pearson 6 statistical distribution.

About 40 % of postsurgical patients are admitted from OR into ICU (direct ICU admission), while 60 % are admitted into inpatient nursing units (NU).

6 Inpatient Nursing Units as a Separate Subsystem

Total inpatient nursing unit (NU) bed capacity was 360 beds. Patient length of stay (LOS) in inpatient NU was assumed to be in the range from 2 days to 10 days, with the most likely of 5 days, represented by a triangle distribution. Simulated census for a typical week is represented in Fig. 2.8. It is clear that the bed capacity limits are consistently hit on a daily basis, usually at the middle of the day, except on weekends.

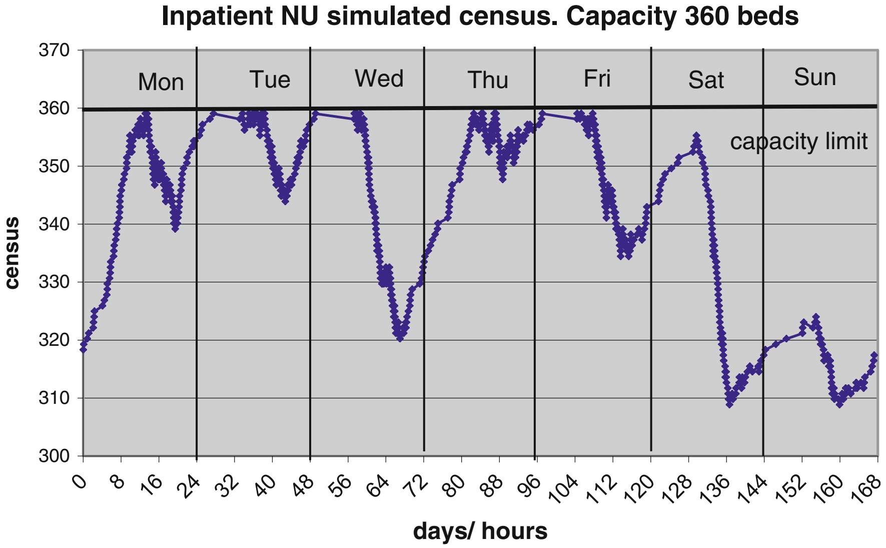


Fig. 2.8 Inpatient nursing units NU. Simulated census of a typical week

7 Reconnecting Separate Units: The Entire Hospital System

As was discussed in the introduction, large complex hospital systems and multi-facility clinics are usually analyzed as deconstructed smaller subsystems or units. Most published simulation models focus on the separate analysis of these individual units. However, according to the principles of complex systems analysis, these separate subsystems (units) should be reconnected back in a way that captures the most important interdependency between them. Simulation models that capture interaction of major units in a hospital, and the information that is obtained from analysis of the system responses as a whole can be invaluable to hospital planners and administrators.

This section illustrates a practical application of this system-engineering principle. High-level simulation models of the separate main hospitals units, i.e., ED, ICU, OR, and inpatient NU patient flow, have been described in the previous sections. These units are not stand-alone systems but they are closely interdependent, as indicated in Fig. 2.1.

The output of the ED model for patients admitted into the hospital (ED discharge) now becomes an ICU, OR and NU input through ED disposition. In our case study, about 70 % of admitted ED patients are taken into operating rooms (OR) for emergency surgery; about 20 % of admitted ED patients move directly into ICU; and about 10 % of patients admitted from ED are taken into combined inpatient nursing units.

At the simulation start on week 1, at the Monday midnight all units are empty, while in reality they are not. We are interested in this analysis in a long-term steady-state period rather than a transient period. Therefore, at the simulation start, the empty units should be prefilled to the typical midnight census values, which are 15, 46 and 350 patients for the ED, ICU and NU, respectively.

A summary of simulations for the various performance metrics is shown in Table 2.1.

Eight performance metrics (95 % Confidence Intervals-CI) are indicated in column 1. Baseline metrics that correspond to patient ED LOS up to 24 h are presented in column 2. It was demonstrated in Sect. 3 that the ambulance diversion for stand-alone ED becomes very low if improvement efforts reduced LOS for patients admitted into the hospital to less than 5 h and LOS for released home patients to less than 6 h (from ED registration to ED discharge). However, because of interdependency of the ED and the downstream units, four out of eight metrics became much worse (columns 3 and 4). The ED bottleneck just moved downstream into the OR and ICU because of their inability to handle the increased patient volume from ED. Thus, aggressive process improvement in one subsystem (ED) resulted in a worse situation in other interrelated subsystems (OR and ICU). ED improvement is not necessarily translated into the goal of increasing the throughput of the entire hospital system. It turns out that patient flow is a property of the entire hospital system rather than the property of the separate departments/units. A detailed analysis is required of the overall hospital system patient flow and the interdependency of subsystems/units in order to establish the system's weak link and the right units for process improvement projects priority.

If, instead of too aggressive ED LOS reduction, a less aggressive improvement is implemented, e.g., ED LOS is not more than 9 h for patients admitted to the hospital, then none of the eight metrics become much worse than the baseline state (columns 5 and 6). While in this case ED performance is not as good as it could be, it is still better than it was at the baseline level. At the same time, a less aggressive local ED improvement does not make the ICU, OR, and NU much worse. In other words, the less aggressive ED improvement is better aligned with the ability of the downstream units to handle the increased patient volume.

Thus, from the entire hospital system standpoint, the primary focus of process improvement should be on the ICU because it has the highest percent of patients waiting for admission more than 1 h and the highest diversion, followed by the NU and ED. If process improvement aimed at reducing patient LOS starts in the upstream unit—ED without addressing first capacity to handle increased patient flow of the downstream units—ICU and NU—it will only result in more patients that are formally discharged from ED but boarded there waiting for admission to ICU and NU, as indicated by the increased diversion and waiting time for latter units in Table 2.1. Otherwise, even if the ED makes significant progress in its patient LOS reduction program based on formal discharge time, this progress will not translate into improvement of the overall hospital-wide patient flow. Of course, many other scenarios could be analyzed using the simulation model to find out how to improve the entire hospital-wide patient flow rather than that for each separate

Table 2.1 Summary of simulation results for the hospital system patient flow performance metrics

	1	2	3	4	5	6
	Performance metrics	Baseline state	Aggressive ED improvement: admitted LOS 5 h; released home LOS 6 h	Better or worse than baseline?	Less aggressive ED improvement: admitted LOS 9 h; released home LOS 10 h	Better or worse than baseline?
1	95 % CI of ED diversion	23.6–23.9 %	3.0–3.1 %	Much better	20.9–21.1 %	Better
2	95 % CI of the percentage of patients left not seen	7.7–8.1 %	0 %	Much better	4.8–5.0 %	Better
3	95 % CI of the percentage of patients waiting admission to OR from ED longer than 1 h	0.05–0.1 %	0.4–0.7 %	Much worse	0.1–0.3 %	A little bit worse
4	95 % CI of OR diversion	0.7–0.8 %	1.9–2 %	Much worse	0.9–1.1 %	A little bit worse
5	95 % CI of the percentage of patients waiting admission to ICU from ED longer than 1 h	25.4–28.2 %	33.4–36.6 %	Much worse	28.3–31.3 %	A little bit worse
6	99 % CI of ICU diversion	16.2–17.9 %	23.4–25.5 %	Much worse	18.9–20.7 %	A little bit worse
7	95 % CI of the percentage of patients waiting admission to NU from ED longer than 1 h	29.5–31.4 %	29.4–31.3 %	Not much different	28.9–30.8 %	Not much different
8	95 % CI of NU diversion	8.5–8.6 %	9.0–9.2 %	Slightly worse	8.7–8.8 %	Not much different

local subsystem/unit. This illustrates one of the fundamental principles of system analysis.

In order to improve ICU throughput performance, more rigorous ICU admission and discharge criteria could be applied. If, for example, ICU admission volume is reduced to 15 % of the total ED disposition patient volume, then the simulation model indicates that the percentage of patients waiting more than 1 h will be about 21 % (down from about 28 to 31 %), and ICU diversion will be about 11 % (instead of 19–20 %).

Another option is reducing the maximum ICU length of stay from 3 days (72 h) to, for example, 2.75 days (66 h) instead of limiting the ICU admission volume. In this case, the percentage of ICU patients waiting more than 1 h will be about 22 % and ICU diversion will be about 13 %. These ICU performance metrics are very close to the above with the reduced admission volume.

Of course, a combination of the above scenarios is possible for further improvement. Many other scenarios could also be modeled to find out how to improve the entire hospital system patient flow rather than each separate hospital department.

8 Effect of Reduced Avoidable 30 Day Readmission Rate

It was already mentioned that nearly one-fifth of patients discharged from a hospital return within 30 days in the USA (MPAC 2007). Identifying and reducing avoidable readmissions will improve patient safety, enhance quality of care, and lower health care spending. That is why policymakers, consumers, hospital leaders and the medical community are focused increasingly on readmissions to hospitals. Most recently, in the Patient Protection and Affordable Care Act (ACA), the US Congress enacted the Hospital Readmissions Reduction Program (HRRP) under which Medicare will penalize hospitals for higher-than-expected rates of readmissions, beginning in 2013. Some hospitals are moving forward with efforts to reduce readmissions and improve quality of care. For example, Metro Health Hospital in Wyoming initiated its Congestive Heart Failure (CHF) readmissions program and cut its avoidable CHF readmission rate to 7.4 % (AHA 2011).

A thirty-day readmission was simulated here as feedback loops of the discharged patients with random uniformly distributed delay in the range from 1 to 30 days (Fig. 2.1). Suppose, for example, that the study hospital analyzed in this chapter cut its total avoidable 30 days readmission rate to about 10 % (including 2 % ICU readmission rate and 8 % inpatient NU readmission rate). Simulation modeling with this lower readmission rate indicated that the ICU performance would markedly improve: the ICU percentage of patients waiting more than 1 h dropped to about 17 % (down from about 28 to 31 %) and ICU diversion is down to 6 % (rather than 19–21 %). Thus, reduction of the avoidable readmission rate not only reduces the monetary penalty but also significantly improves performance characteristics.

9 Conclusions

Analysis of a complex system is usually incomplete and can be misleading without taking into account subsystems' interdependencies. The insight that systems behave differently than a combination of their stand-alone independent components is a fundamental management principle.

It was demonstrated in this chapter that the performance of a hospital-wide system could inadvertently be jeopardized because locally oriented improvement in one process or department worsens performance of the overall system. It may be said "Curing the Process May Kill the System" (Kamanth et al. 2011). It was quantitatively demonstrated using simulation modeling that aggressive process improvements implemented in the ED to reduce patient length of stay (good for the ED) can result in increasing ICU and operating room wait time and percent diversion (bad for the ICU and OR). Thus, improvements in an upstream subsystem may worsen performance of the downstream units and the overall system—at least for some performance measures. Therefore, improvement of the upstream units should be aligned with the ability of the downstream units to handle the increased patient volume. The ability of system analysis and simulation modeling methodology to incorporate a broader system-thinking approach is one of its advantages over some local process-specific improvement methods, such as plan-do-study-act (PDSA) learning cycles (Kamanth et al. 2011).

References

AHA. (2011). *Trend watch. Examining the drivers of readmissions and reducing unnecessary readmissions for better patient care*. Washington, DC: American Hospital Association (AHA). September 2011.

Blasak, R., Armel, W., Starks, D., & Hayduk, M. (2003). The use of simulation to evaluate hospital operations between the ED and medical telemetry unit. In S. Chick et al. (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1887–1893). Washington, DC: IEEE.

CAEP. (2007). Canadian Association of Emergency Physicians. Position statement on emergency department overcrowding, February issue. www.CAEP.ca

Carr, S., & Roberts, S. (2010). Computer simulation in healthcare. Chapter 14. In Y. Yih (Ed.), *Handbook of healthcare delivery systems*. Boca Raton, FL: CRC Press.

Cho, S. J., Jeong, J., Han, S., Yeom, S., Park, S. W., Kim, H. H., et al. (2011). Decreased emergency department length of stay by application of a computerized consultation management system. *Academic Emergency Medicine, 18*(4), 398–402. doi:10.1111/j.1553-2712.2011.01039.x.

Clifford, J., Gaehde, S., Marinello, J., Andrews, M., & Stephens, C. (2008). Improving inpatient and emergency department flow for veterans. *Improvement report*. Institute for Healthcare Improvement. Retrieved from http://www.IHI.org/ihi

Cochran, J., & Bharti, A. (2006). A Multi-staged stochastic methodology for whole hospital bed planning under peak loading. *International Journal of Industrial and System Engineering, 1*, 8–36.

Costa, A., Ridley, S., Shahani, A., Harper, P., De Senna, V., & Nielsen, M. (2003). Mathematical modeling and simulation for planning critical care capacity. *Anesthesia, 58*, 320–327.

De Bruin, A., van Rossum, A., Visser, M., & Koole, G. (2007). Modeling the emergency cardiac in-patient flow: An application of queuing theory. *Health Care Management Science, 10*, 125–137.

Goldratt, E., & Cox, J. (2004). *The goal* (3rd ed.). Great Barrington, MA: North River Press.

Gunal, M., & Pidd, M. (2006). Understanding accident and emergency department performance using simulation. In L. Perrone et al. (Eds.), *Proceedings of the 2006 winter simulation conference* (pp. 446–452). Washington, DC: IEEE.

Haraden, C., Nolan, T., & Litvak, E. (2003). Optimizing patient flow: Moving patients smoothly through acute care setting. Institute for Healthcare Improvement Innovation Series 2003. White Papers 2: Cambridge, MA.

Hopp, W., & Lovejoy, W. (2013). *Hospital operations: Principles of high efficiency health care*. Fontana, CA: FT Press, Upper Saddle River, NJ.

Jacobson, H., Hall, S., & Swisher, J. (2006). Discreet-event simulation of health care systems. In R. Hall (Ed.), *Patient flow: Reducing delay in healthcare delivery* (pp. 210–252). New York, NY: Springer.

Jencks, S., Williams, M., & Coleman, E. (2009). Re-hospitalizations among patients in medicare fee for service program. *New England Journal of Medicine, 360*, 1418–1428.

Kamanth, J., Osborn, J., Roger, V., & Rohleder, T. (2011). Highlights from the third annual mayo clinic conference on systems engineering and operations research in health care. *Mayo Clinic Proceedings, 86*(8), 781–786.

Kolker, A. (2008). Process modeling of emergency department patient flow: Effect of patient length of stay on ED diversion. *Journal of Medical Systems, 32*(5), 389–401.

Kolker, A. (2009). Process modeling of ICU patient flow: Effect of daily load leveling of elective surgeries on ICU diversion. *Journal of Medical Systems, 33*(1), 27–40.

Kolker, A. (2012). *Healthcare management engineering: What does this fancy term really mean? Springer_Briefs series in healthcare management & economics*. New York, NY: Springer. 122.

Law, A. (2007). *Simulation modeling and analysis* (4th ed.). New-York: McGraw-Hill.

Lefcowitz, M. (2007, February 26). Why does process improvement fail? Builder-AU by Developers for developers. Retrieved from www.builderau.com.au/strategy/projectmanagement/

Litvak, E., Long, M., Cooper, A., & McManus, M. (2001). Emergency department diversion: Causes and solutions. *Academic Emergency Medicine, 8*, 1108–1110.

Marshall, A., Vasilakis, C., & El-Darzi, E. (2005). Length of stay-based patient flow models: Recent developments and future directions. *Health Care Management Science, 8*, 213–220.

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 Management Science, 11*, 11–21.

McManus, M., Long, M., Cooper, A., Mandel, J., Berwick, D., Pagano, M., et al. (2003). Variability in surgical caseload and access to intensive care services. *Anesthesiology, 98*(6), 1491–1496.

Miller, M., Ferrin, D., & Szymanski, J. (2003). Simulating Six Sigma Improvement Ideas for a Hospital Emergency Department. In S. Chick et al. (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1926–1929). Washington, DC: IEEE.

MPAC (Medicare Payment Advisory Commission). (2007). Payment policy for inpatient readmissions. *Report to the congress: Promoting greater efficiency in Medicare*. Washington, DC.

Oredsson, S., Jonsson, H., Rognes, J., Lind, L., Göransson, K., Ehrenberg, A., et al. (2011). A systematic review of triage-related interventions to improve patient flow in emergency departments. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 19*, 43. doi:10.1186/1757-7241-19-43.

Reid, P., Compton, W., Grossman, J., & Fanjiang, G. (Eds.). (2006). Building a better system: A new engineering healthcare partnership. Washington, DC: National Academy of Engineering and Institute of Medicine. The National Academy Press

Savage, S. (2009). *The flaw of averages*. Hoboken, NJ: Wiley. 392.

Simon, S., & Armel, W. (2003). The use of simulation to reduce the length of stay in an emergency department. In S. Chick et al. (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1907–1911). Washington, DC: IEEE.

Wang, J. (2012). Reducing length of stay in emergency department: A simulation study at a community hospital. *IEEE Transactions on Systems, Man, and Cybernetics. Part A: Systems and Humans, 42*(6), C1–1309.

Chapter 3
Hospitals and Clinical Facilities, Processes, and Design for Patient Flow

Michael Williams

Abstract This chapter discusses current trends and key drivers that affect patient flow and efficiency and analyzes the most common myths of resource allocation for healthcare. Providers are now finding that simply adding beds or staffs will not solve the commonplace problems of long waits and delays. Contemporary physical design concepts that improve flow for healthcare are described in this chapter, and seven high-leverage steps that can significantly improve flow and expand capacity, and thus limit delays and waiting, are recommended.

Keywords Ambulance diversion • Artificial variation • Bed control • Boarding • Bottlenecks • Capacity • Crowding • Design • Discharge lounge • Emergency department • Fast track • Flow • High impact teams (HIT) • Just in time • Lean • Random variation • Rapid admission unit • Rapid cycle testing (RCT) • Redesign • Reengineering • Safety net • Smoothing • Staffing ratios • Utilization • Workforce shortages

1 The Challenge

Long waits, delays, cancellations, and resource overloads have become commonplace in healthcare. For many years, healthcare providers have simply added more resources to solve the problem—building more beds and adding more staffs. This approach has become increasingly impractical due to human resource shortages and limited finances. Now healthcare providers have been forced to look at different approaches to solving the problem. In addition, many of the traditional approaches have only served to hide the real underlying problem: significant inefficiencies in timing and flow of resources during the delivery of healthcare.

M. Williams (✉)
The Abaris Group, 700 Ygnacio Valley Road, Suite 270, Walnut Creek, CA 94596, USA
e-mail: mwilliams@abarisgroup.com; TheAbaris@aol.com

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_3, © Springer Science+Business Media New York 2013

The problems of healthcare crowding and waiting have become epidemic across the country. Nowhere is this more evident than with our country's emergency departments (ED). 45.5 % of all hospital admissions come from the ED (McCaig and Burt 2004). All too often when one thinks of an ED visit, one assumes that there will be long waits. This is frustrating to the patient, their families, and the hospital staff as well. ED saturation and subsequent ambulance diversion have also reached crisis proportions in most urban communities. Ambulance diversion, or the sending of an ambulance to another hospital because the ED chosen is "closed," has a substantial backward effect on the ambulance industry as well. The ambulance industry is one healthcare provider that has experienced significant success with optimizing their resources and productivity to assure performance using traditional industrial engineering concepts (Stout 1986). However, an October 2001 US Government study shows that:

> Ambulance diversions have impeded access to emergency services in metropolitan areas in at least 22 states since January 1, 2000. More than 75 million Americans live in areas affected by these ambulance diversions. (US GAO 2003)

These problems have arisen because of a dramatic increase in ED utilization and a perceived parallel decline in resources. The CDC reports that the number of ED visits for 2003 rose by 3.1 % to 112 million patients while the total number of EDs declined during the past 10-year period by 12.2 % (McCaig and Burt 2004). In a survey they conducted on behalf of the American Hospital Association, the Lewin Group documented that 62 % of the surveyed hospitals had reported being "at" or "over" capacity, with this proportion rising to 79 % for urban hospitals and 90 % for level I trauma centers and hospitals over 300 beds (Lewin Group 2002).

Hospital capacity is a major driver to the overall healthcare capacity and patient flow challenges across the country. In a recent report on ED crowding, the US General Accounting Office (GAO) noted the connection between the ED and the rest of the hospital:

> While no single factor stands out as the reason why crowding occurs, GAO found that the factor most commonly associated with crowding was the inability to transfer emergency patients to inpatient beds once a decision had been made to admit them as hospital patients rather than to treat and release them. When patients "board" in the emergency department due to the inability to transfer them elsewhere, the space, staff and other resources available to treat new emergency patients are diminished. (GAO 2003)

The Lewin ED study also noted that the lack of critical care beds was a factor for ED crowding. While staffing and other factors were also mentioned, the lack of critical care beds was cited as the most common cause in most of the regions studied.

Hospitals themselves are complex organizations, and thus waiting and delays are also common and often a convenient explanation of the problem. Delays in scheduling a non-emergent surgery can be weeks. High patient admission volumes in the morning at a hospital and patient discharges that occur in the afternoon assure that many patients will wait in a queue, whether that be at their home, in a physician's office, at the hospital's admitting department, or perhaps in the ED.

Hospitals had historically responded to capacity problems by adding more staffs and beds. However, frequently these changes only make the problem worse when the underlying processes and practices do not change and the new beds or staffs eventually get used with capacity problem ultimately returning. Limited physical capacity and staffing often result in the use of alternative but not the most desirable resources such as "boarding" patients in the ED. At best, these responses are just a band aid. Yet, in a survey conducted in 2004, approximately 51 % of all hospitals in the country were rebuilding or expanding their EDs, some of which are taking this step to simply accommodate the holding and boarding of patients (Healthcare Financial Management Association 2004).

Due to the lack of space or the funds to expand or add staff, more and more hospitals are being forced to look at improving their flow of patients by studying bottlenecks and limitations in their process that artificially add to the problem. Unfortunately, many hospitals initially attempt to focus on the symptoms of the problem, the ED or in some cases the emergency medical system (EMS) delivery system that brings the patient to the hospital. The challenge is that the ED and inpatient capacity and flow are inextricably linked, and resolving only the subsystems, such as the ED, will only have limited success. In fact, isolated achievements only serve to provide short-term successes but actually hurt the other departments by artificially growing the problems that they face. For example, ED initiatives often and appropriately prioritize getting ED patients that are being admitted to a hospital bed a priority, and some simply set high performance standards to achieve their goal. But often, the hospital does not have a bed to send that patient to whether it is due to poor in-house staffing or lack of contemporary in-house bed utilization and discharge practices. There is even some evidence that mandatory nurse ratios may increase patients' risk for mortality and morbidity (Aiken et al. 2002).

To achieve a total and sustainable success to the patient flow and capacity problems, healthcare providers must embrace the interdependencies of their individual departments and services and accept solutions that view the entire continuum of care rather than the mere silos within.

2 Key Drivers

The key drivers to hospital-wide capacity problems and their solutions are as follows:

Increasing Demand and Declining Capacity. It is clear from CDC data that demand for key services is rising and the overall number of hospitals is declining. While the total number of ED visits rose 3–5% per year during the 5-year period from 1998 to 2003, the US population growth was only an averaged 1.1 % per year for that same 5-year period (US Census, *YEAR*). As ED visits continue to rise, so do their associated hospital admissions. The number of uninsured Americans below 65 rose from 42 million to 48.1 million during 1998 to 2004 for the first 6 months of each year (AHRQ 2004). As the number of uninsured increases, so does their impact on EDs and subsequent hospital care as the uninsured and underinsured tend to use the ED as their healthcare safety net. According to the Urgent Matters report: *Walking a Tightrope*: *The State of the Safety Net in Ten U.S. Communities*, despite long waits for safety net populations in an ED, the ED was perceived to be more convenient and more accessible than for long waits for specialty care and multiple provider visits for testing and procedures (Regenstein et al. 2004).

Decreasing Revenues. The overall decline in Medicare and Medicaid reimbursement is not helping. Hospitals increasingly have to work harder with fewer resources. Budget pressures at state and federal government levels are resulting in decreases from these key funding sources.

Workforce Shortages. Hospitals are facing many challenges in recruiting and retaining key workforce positions. The nursing shortage is well documented and only expected to get worse. It has been reported that one in ten nursing positions remain unfilled (Sochalski 2002). There are many other largely unrecognized workforce shortages including radiology technicians, pharmacists, and even medical coders, all critical to the ability of a hospital to maintain and expand its capacity.

Rising Costs of Care. After nearly a decade of relatively stable costs of healthcare brought on by the mandates of managed care and consumerism, there has been much erosion and hospital costs are rising rapidly. The average adjusted expense of a hospital admission in 1997 was $1,031 and in 2002 it was $1,289, up 29 % for this 5-year period (AHA 2005).

Limits in Technology and Informatics. Tight revenue often drives limits to capital expenditures with information technology (IT) resource acquisitions near the bottom of the list (Robeznicks 2005). Key historical and forecasted data resources are missing in many hospitals. These resources are desperately needed to more precisely match resources to demand.

Limited Industrial Engineering Wherewithal. Coupled to the IT challenge and the lack of data is the inability to study, interpret, and respond to opportunities to improve the practices, policies, and procedures that limit demand. Without key data and supporting tools of industrial engineering, a hospital is left to react to patient events rather than to respond to forecasted needs unlike the practice in many other industries (Chase et al. 2001). Contemporary principles of "Lean" and "Just in Time" (JIT) introduced by the visions of Toyota and Federal Express are not embraced in a significant way by the healthcare industry.

3 Key Myths of Healthcare Delivery

One of the most significant factors driving hospital capacity constraints are key myths held by many on their perception of the drivers for capacity challenges. For example, it is a long-held belief by hospitals and their providers that ED visits and inpatient admissions are isolated events that are dependant on variables out of the hospital's control. The fact is that ED visits and subsequent hospital admissions have significant predictability and thus the ability to forecast demand and the necessary resources to a very precise level.

The CDC has also reinforced a long-held belief that the number of EDs in this country is declining, thus driving excessive throughput times and excessive ED diversion. A recent California study reported that, while the number of EDs has declined in that state, the actual number of treatment stations or "beds" has increased substantially (Melnick et al. 2002). The article goes on to reinforce that total ED bed capacity, not the number of EDs themselves, is the more appropriate metric for historical comparison of capacity.

Insufficient hospital beds are another myth held by many healthcare providers. A common hospital frustration is that they "do not have enough beds," but the fact is that in most hospitals, even when there is waiting in the ED, admitting area, or a private physician's office for an inpatient bed, the patient ultimately gets to a hospital bed. Very few patients are transferred to another hospital in order to access an inpatient bed. This clearly demonstrates that actual capacity itself is not the problem, but rather there is a misaligned capacity as compared to demand. Hospital demand for beds typically occurs early in the day, but the patients that are being discharged do not go home until later in the day. Figure 3.1 provides the typical hospital admissions by hour of the day and compares that to the same hospital's discharges by hour of the day demonstrating the mismatch of capacity to demand for hospital beds over a 24-h period.

Another myth is the largely held belief that there is a nursing shortage in the country. While there may be pockets of shortages within the nation, quite the contrary is true nationwide. According to an American Nursing Association study, there is currently an excess number of nurses when compared to demand and there will not be an actual shortage until the year 2010 (Peterson 2001). The fact of the matter is that many nurses do not want to work in a dysfunctional and seemingly unsafe healthcare environment, thus creating an artificial shortage.

Another common but mistaken myth is that ED visits and hospital admissions in general are isolated events. EDs commonly staff for the unknown. Individual hospital admissions from the ED are often interpreted from the in-house unit floor as a "surprise." The everyday bed management meetings that many hospitals have undertaken due to bed "shortages" (typically called bed control meetings) rarely value the predicted ED inpatient demand but rather inappropriately focusing on other hospital bed needs and only the present "boarders" in the ED.

These myths are perpetuated by a lack of appreciation of current data that already exists to make predictions, limited availability of forecasting tools (e.g., ED arrival times, hospital patient discharge times), and lack of knowledge about the nature and impact of artificial variation created by the healthcare delivery system itself and its impact on capacity and flow.

Knowing and valuing the factors that limit the rate of patient flow and increase waiting are essential steps to optimize healthcare capacity and flow delivery. If the wrong problems are solved (e.g., adding staffs and beds), as is the approach taken by many hospitals, then there will be much wasted resources and the problems will

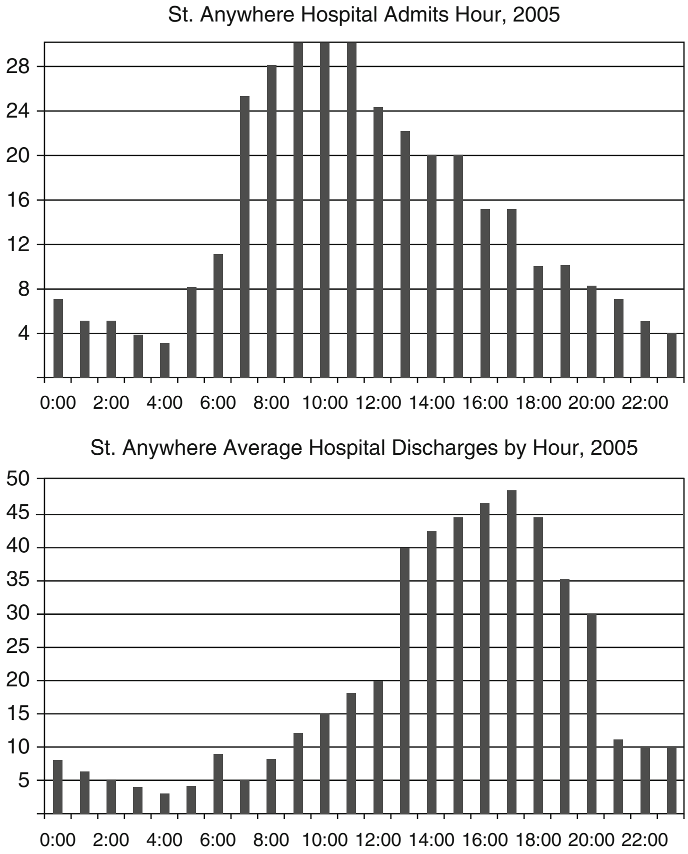


Fig. 3.1 Typical hospital admissions and discharges by hour of the day

only get worse. For many hospitals, the solutions are not with building a bigger "sandbox" but rather building a more effective "sandbox."

Most healthcare providers with excessive waiting and bottlenecks do have a commitment to their organizations but do not realize that they have the internal tools to solve the problem. The central source of problem resolution comes with a principle introduced by Dr. Donald Berwick, MD, President of the Institute of Health Improvement (IHI) wherein he published the first law of improvement: "Every system is perfectly designed to achieve the results that it achieves" (Berwick 1996). Hospitals are at capacity and EDs are overcrowded because they have been incorrectly designed that way. That is, they have process, flow, and sometimes physical design flaws. Thus, the answer to capacity and flow problems is likely to be with the fundamental rethinking and redesigning of their entire healthcare delivery system that created the capacity and flow problems in the first place.

The goal of any healthcare provider is getting the right patient at the right time, with the right provider and with the right information all timed with the right interventions. When these elements are synchronized and waste is eliminated, it simply takes less staff and less space to provide healthcare services, and thus this alone increases capacity and access for the next patient.

Using contemporary capacity and flow management principles will:

- Improve access
- Reducing waiting
- Lower costs
- Improve outcomes
- Improve staff satisfaction
- Improve the customer experience

Best practice improvement initiatives are now demonstrating that it is possible to reduce the stress of the healthcare system experiencing delays and waiting (e.g., ED diversion) and eliminate waiting and delays in access through the enhancement of flow of patients and their information through the care delivery system (see www.abarisgroup.org, www.ihi.org, and www.urgentmatters.org).

These changes are occurring one provider at a time and also system-wide, as is the case of the 18 hospitals in the Sacramento County (CA) area that committed to a profound regional capacity and patient flow change process that ultimately resulted in the region's 73 % drop in ED saturation and EMS diversion (Patel et al. 2006).

4 Physical Plant Considerations

To put it mildly, healthcare physical plants are not known for their accomplished design. One only needs to think about an ED waiting room to conjure up images of uncomfortable chairs, painfully out-of-date color schemes, and long delays while reading year-old magazines.

The physical plants for many healthcare sites are a long way from designs that are healing and efficient and promote patient flow. And yet it may seem trite to use the architectural principle that "form must follow function" but nothing could be closer to the truth. Thus for most architects, physical plant changes and new additions themselves are not likely to fix capacity and flow problems, but if physical space is incorrectly designed or more importantly designed to a flawed process or a hypothetical process, then the physical plant may in fact be the rate-limiting factor.

Healthcare delivery systems must be designed to support contemporary flow and capacity management functions. Poor design may have a substantial affect on patient capacity and flow. Key sources of physical plant bottlenecks include:

- Lack of long-range planning or a master plan, thus requiring a patchwork of architectural remodels or "solutions" that do not work well together
- Excessive redundancies of healthcare provider departments that drive duplicative spaces, excessive equipment, and excessive steps needed to move patients through the system

- Inadequate use of technology and thus technology or technology related to a bed (e.g., telemetry) becoming the rate-limiting factor
- Inadequate space management limiting the effective use of current space and the inability to appropriate use of underutilized space
- Multiple and duplicative recording and management systems, thus resulting in ineffective integration and fragmentation of patient information space
- Designs that are a perfect fit for one style or philosophy of care or for a particular manager but quickly out of date for the next
- Reluctance to automate from manual systems to computerized approaches
- Departments that are landlocked due to inadequate planning
- Narrow treatment bays inhibiting efficient operations or economies of scale
- Gross-to-net-area ratios that are inadequate due to poor planning or being forced to use a limited footprint
- Poor circulation patterns due to poor planning or a lack of expertise
- Poor proximities of essential functions resulting in long travel distances, inefficient flow, and higher staffing ratios
- Mechanical and electrical systems that have limited capacity to grow
- Seemingly random placement of vertical shafts, elevators, and stairways inhibiting necessary expansions
- Poor signage and flow design, thus assuring that patients and families will have difficulty finding their way
- Inadequate support space for staff and ancillary departments
- New technologies that are difficult to accommodate due to inflexible infrastructures and a lack of master planned utilities
- Lack of phased replacement plan in existence for the gradual upgrade of a hospital's infrastructure
- Inadequate safety and clinical care systems that are unable to advance for the changing environment of infections, bioterrorism, and the like
- Designs that are not welcoming, healing, or do not act in a positive supportive way when there is waiting

Traditional space planning and architectural design also tend to reinforce results that memorialize waiting, delays, and inefficiencies. For example, a typical redesign of an ED will start with how large the waiting room should be. A common calculation for ED waiting space is 3 seats per treatment bed or 15 net square feet per seat (ACEP 1993). For a 50,000 volume ED, that ratio will produce approximately 28 treatment beds, using a common guide of 1,800 visits per treatment bed. This will generate a waiting area space of approximately 420 square feet, which if converted to treatment beds would be sufficient to generate two more beds and in some cases eliminate the need for a waiting room. Thus, waiting rooms tend to memorialize waiting.

Removing the traditional architectural biases will require leaving behind traditional architectural design concepts and replacing those with approaches that truly think "out of the box" and perhaps "out of the universe." The essential design strategies that will promote contemporary capacity and flow management strategies are as follows:

Flexibility

- Locate growth departments along open edges of the facility or adjacent to soft departments such as offices, storage, or courtyards.
- Utilize structure systems that can be adapted.
- Organize infrastructure to allow new components to be plugged in.
- Develop a universal approach to examination and treatment rooms.
- Avoid locating several fast-growing departments adjacent to each other unless there are outlets to permit growth.
- Build in shell space.
- Design in appropriate horizontal buffer zone space to allow for expansion of technology and thus allow modification of electrical, mechanical, and IT systems to support such.
- Do not short-change support space for supply, staff, and logistical support such as education and IT services.
- Do not landlock areas that will need replacement or additions in the future.

Efficiency

- Optimize functional internal relationships based on the highest frequency of need and intensity of use.
- Balance care needs with support departments, but do not allow the support departments to drive the assumptions on space.
- Design circulation and infrastructure patterns, so they can be adapted as needs change.
- Plan for facility development priorities that create logical sequencing for the future.
- Incorporate "smart" building planning to allow adapting to technology advances.
- Emphasize space planning that allows for a variety of models of care.
- Plan for bringing more services to the patient (e.g., radiology, point-of-care laboratory testing).

Quality

- Create research environments that promote environments that promote healing and comfort (see www.planetree.org).
- Design around enhanced productivity and staff morale.
- Focus on designs that will have a positive effect on the customer and market share.
- Minimize traditional irritants such as noise, glare, and privacy.
- Target opportunities to deinstitutionalize the facility and improve access such as parting, waiting areas, and nutrition.
- Recognize patient imperatives of safety, comfort, and privacy.
- Allow for space personalization for patients and staffs.

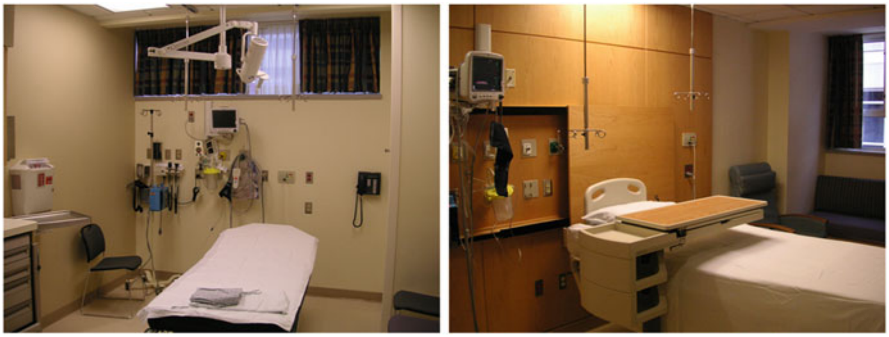


Fig. 3.2 Before and after photo of a patient treatment room moving from a technical and more austere design to a warm and healing environment. *Photos courtesy of Frank Zilm*, *AIA Zilm & Associates*, *Inc*. *Kansas City*, *MO*

Figure 3.2 provides a before and after picture of a traditional patient bed versus one that is designed for comfort and aesthetics in a "healing" environment. The key ingredient to enhance flow and capacity through design is to create flexibility, whether that is for care patterns, mix variations, new demographics, new technologies, or new modes of care or reimbursement. It is also helpful to develop physical configurations that are based on acuities and levels of care. For example, zones should be created that match urgent and outpatient care with diagnostics as well as specialized care needs, such as psychiatric, pediatrics, and geriatrics. Healthcare design should encompass physical planning to enhance the ability to achieve cost savings. With the advent of JIT and Lean manufacturing production strategies, excessive storage and hording will be a thing of the past (Williams 2004). This kind of physical accommodation will also be necessary to support team configurations, care approaches, and proximities that provide efficiencies and enhance flow. For example, if point-of-care laboratory testing is to be the future standard of practice for patient care, as it is becoming in some EDs, then there must be physical bedside space to accommodate this change.

Consideration should also be given to decentralizing space where appropriate and relocating nonessential services to improve treatment capacity. It is becoming a mantra in space-compromised hospitals to maintain "first floor space as patient care space" as a criterion for considering relocating administrative, PBX, and other nonessential first floor uses to improve access and expand patient care services. Remember also that privacy concerns continue to remain a priority in healthcare, and this concept should always be at the forefront of healthcare space planning. Finally, taking from a Disneyland concept, the customer or the guest experience goal should be "to make the best first impression and the best last impression" to the patient and their family and to design aesthetics and environmental quality into the planning process (Disney Institute 2001).

5 Key Contemporary Solutions for Capacity and Flow Management

Through much trial and error, healthcare providers have learned the basic steps which must be taken to successfully and substantially improve patient flow. When speaking of conceptual approaches, it is important to think of high-leverage opportunities rather than attempting the universe of opportunities. Presented below are seven high-leverage steps that can be taken to dramatically and profoundly change the way a healthcare provider conducts its business and thus improve patient flow and enhance capacity.

5.1 Develop Robust Products That Decompress ED and Inpatient Volumes

Some hospitals have taken proactive steps to put patients in care delivery models that speed the care process itself, stage the patient for expedited care, or reduce overall length of stay. Few hospitals, though, have all of the necessary models or product lines in place. Others have products that underperform and thus should be significantly reengineered to create a true performance-based and high-leverage throughput delivery system.

5.1.1 Fast Track

One such product is an ED Fast Track. This product is typically located in a dedicated area of the ED designed to treat lower acuity patients in a speedier manner. However, most current ED Fast Tracks are slow, not producing anywhere near the 60-min ideal throughput time that should be the goal for a Fast Track. For most EDs, 80 % of their volume is considered non-urgent and therefore that volume would lend itself to more of a primary care treatment and flow model that provides faster services supported by more efficient tools (e.g., checkbox clinical records, point-of-care testing). Most EDs should cycle approximately 40–50 % of its patient volume through this faster care model, thus dramatically reducing the total time on task, providing a protected and efficient care plan for those patients so that they do not get trumped by higher acuity patients, improving patient satisfaction and dramatically improving the bed capacity for the remaining higher acuity patients.

5.1.2 Clinical Decision Unit

A clinical decision unit (CDU) should be considered for hospital admissions that do not truly need a traditional in-house bed. CDUs are 8–12-bed units designed for patients that would traditionally be admitted for conditions that simply need more therapy or care but do not necessarily need an inpatient bed. Most hospitals do not have a CDU. A typical CDU admission would include patients who have a need for longer diagnostic testing (e.g., rule-out cardiac chest pain), therapy (e.g., asthmatic), or other conditions that lend themselves to limited time protocols. These patients typically get admitted to the hospital and thus take up a hospital bed for up to 2 days. The average admission time for a CDU patient is 14 h as compared to the 24–48 h that their admission would have taken if the patient were in a traditional hospital bed. For most EDs, a CDU substantially reduces ED admissions to the hospital by up to 30 %, and the CDU also has a bonus of dramatically improving inpatient capacity as well. The overall length of stay for these CDU patients is dramatically less than their in-house counterpart patients. For a rule-out chest pain patient that is admitted to the hospital, the length of stay will range from 23 to 48 h compared to only 10 to 14 h for CDU patients. For example a hypotensive patient needing fluid replacement might stay in the hospital for an average 24–36 h in-house but would likely only take 12–14 h in CDU (Graff 1998). These collective "saved" hours will dramatically open up inpatient beds due to the dropped overall utilization.

5.1.3 Rapid Admission Unit

A rapid admission unit (RAU) is a designated area for patients that are going to be admitted to the hospital but there is no available hospital bed to begin the admission work-up, orders, and paperwork. All hospitals should consider an RAU that provides peak weekday coverage for direct admissions and for staging inpatient admissions from the ED when there is no inpatient bed available. This model uses a 2-h throughput model for patients. The RAU is staffed only during peak weekday hours and thus not 24 h or 7 days per week. The RAU area could accommodate the admission process, initiate early orders including taking the admission orders from the private physicians, and evaluate the correct type of bed for the patient avoiding the common practice of further unnecessary patient moves during their inpatient stay.

5.1.4 Discharge Lounge

Hospitals should also consider a discharge lounge that provides a quality location for discharge patients who would otherwise be waiting to be discharged from their hospital bed. With the discharge lounge, these patients wait in the discharge lounge for prescriptions, transportation home, care education, or home healthcare scheduling. The patient that is going home but simply waiting for these logistical services is sent to an area of the hospital that is near the door where they will be picked up. Refreshments are served in a very nice area and perhaps a meal provided, and they receive any final patient education or medications needed. This unit can dramatically speed the day-of-discharge times freeing up critical needed beds for that day's admissions.

5.2 Match Staffing to Demand

Healthcare providers should take steps to precisely match staff to demand. For most providers, this is more of a guess than a precise process management tool. If the staffing has not been carefully studied or allocated to precise demand, a mismatch of service delivery will occur and there will be a resultant backlog of patient flow. This precise demand management review should be done for all core hospital functions including the ED, laboratory, radiology, housekeeping, central supply, and inpatient care units.

The steps for matching demand to patient flow are as follows:

5.2.1 Analyze the Staffing Data

Collect and use historical data by month, day of week, and hour of day to project needs and demand for services. Plotting the hospital admissions by various time periods is useful in identifying seasonal, weekly, or daily patterns and is the first step in understanding the demand/capacity ratio of services.

5.2.2 Adjust Staffing to Demand

Once the patterns of demand have been identified, the capacity of the system to handle the expected demand can be increased by arranging to have appropriate staff available during peak times. Staffing demand includes not only direct staff but also ancillary services such as registration, laboratory, and radiology services. A weight of 40–60 % of technical versus nursing staff is recommended to sufficiently balance workload and to assure that nurses are used primarily for their nursing responsibilities.

5.2.3 Prepare Contingency Plans

Even as patterns of peak demand are identified and staffing patterns adjusted, there will be times when unexpected demand occurs. The provider should establish backup systems such as on-call systems or other contingency plans for meeting unanticipated demand. Having procedures in place wherein the unit or the department can call on staff from other parts of the hospital to support them during unexpectedly high demand times can also be an effective method for reducing delays. It is also important to have formal and protected contingency plans for both nursing and ancillary staff and to establish these contingency plans for unpredictable delays.

5.2.4 Ensure That the Management Team Has Sufficient Resources, Tools, and Ability to Meet Objectives

Creating precise demand management strategies requires managers to be surrounded with the right tools. What is often missing for the managers to create precise staffing is precise data. The data do not exist or access to the right data at the right level of details is limited. Key leadership positions should utilize the robust information systems and other resources available to them to make such calculations and adjustments to staffing to meet the needs and trends.

5.3 Reduce Unnecessary Utilization

One of the most potent sources of delay in hospitals, especially in the ED, is a patient waiting for laboratory and radiological procedures and results. This is particularly a challenge in the ED environment, where primarily confirmatory tests are routinely ordered and are part of the accepted risk management process. Healthcare providers should identify diagnostic tests that contribute to neither the patient's diagnosis nor the patient's treatment regime but rather are primarily confirmatory in nature or comply with perceived risk management objectives.

Studies on utilization and productivity and variations between physicians should also be refined and completed on an ongoing basis. This can be discreetly accomplished and should be considered as a place to start. Utilization standards for these tests should be developed with the overall goal of reducing unnecessary utilization.

5.4 Synchronize Care Delivery

For most healthcare providers, a significant source of delays is found within the inpatient unit or the ED itself. For example, most EDs have slow entry times from the time the patient arrives to the time the patient gets to a bed. Many hospitals have slow discharge times from when the physician writes the order to the time the patient leaves the hospitals. This is because many of the services and activities that are needed to complete the patient transaction are out of synch with the process.

Treating patients swiftly requires coordinating all processes as well as in ancillary departments, such as laboratory and radiology. It is important to standardize as many tasks as possible in order to achieve the synchronized and efficient care delivery system desired. Important steps for this effort follow.

5.4.1 Focus on Getting the Patient to the Provider

Most patients want to see the physician, so processes that interfere with that should be reexamined. In the ED, the point when the physician enters the exam room is the point around which everything else should revolve. Coming to an agreement on the importance of this point is crucial in achieving synchronization.

5.4.2 Evaluate, Delete, or Retime Processes That Do Not Enhance Turnaround Time

An example of this effort would be to continue to study the activities and behaviors of caregivers and eliminate unnecessary steps and activities or reschedule these activities. This includes streamlining triage when there is a delay due to a lack of beds, reducing assessment exams, and reducing duplicate questions between the registration/admitting staff and the physician nurse team (e.g., “why are you here today?”). Moving processes closer to the patient is important.

5.4.3 Establish Protocols for Top Diagnoses

Hospitals and EDs should establish a number of protocols or care maps on care management. These protocols would provide a total set of agreed-upon steps to be taken in the diagnosis and treatment of particular types of patients. Protocols can greatly reduce delays by streamlining the transition of patients from one step in the treatment process to another. These protocols will also be effective in identifying steps in the treatment process that can be eliminated or provided by other professionals, rather than solely by nurses or physicians.

5.4.4 Based on Protocols, Initiate Action

Once a protocol has been agreed upon, a patient who arrives at a hospital bed or in the ED with a condition for which a protocol is in place can be moved immediately through the steps, eliminating the delays that often occur in ordering appropriate tests. For example, with a patient who has an appropriate extremity injury (e.g., meeting the Ottawa extremity rules), with a pathway X-ray guideline, can be moved directly from triage to radiology rather than waiting to be seen by a physician. Another example would be for known asthmatics to have their breathing treatments initiated by the nurses in a timely manner as a result of an established protocol.

5.5 Reconnect Services Within the Hospital

It is not unusual to have a philosophical disconnect between the ED and the departments that support them. This is not uncommon for hospitals, as each department traditionally operates with its own "silo" accountable for the individual performance and service delivery standards. Many of the flow issues facing most EDs for example are manifestations of processes occurring elsewhere in the hospital, particularly in the flow of inpatients through the service system.

In reality, each department is part of the larger system involving prehospital, hospital patient care units, other hospital departments, laboratory, radiology, other support services, community physicians, consultants (physician specialists and other professional disciplines), as well as patients, their families, and the communities in which they live. While a smooth-functioning unit depends on the services that others in the wider system of care provide for the unit, this level of functionality can be difficult to achieve since others may not see themselves as part of this wider vision of the unit's system. This "we" and "they" philosophy permeates most hospital cultures. Once the incentives are aligned, patient care managers held accountable, and the patient put into the center (e.g., not "your" patient or "my" patient but "our" patient) a significant breakdown of the silo mentality occurs. You then start seeing breakthrough behaviors as this culture reverberates throughout the hospital. Some of this culture change occurs with a move away from the "push" methodology where patients have to push to the next unit versus a "pull" program where the accepting unit actually pulls the patient to the unit perhaps even coming to the ED to take "our" patient and to avoid further delays. One such "pull" model is the Adopt-a-Boarder Program at Stony Brook Hospital in New York where unit nurses have agreed to accept patients in their hallways if there are no beds (see www.urgentmatters.org/enewsletter/voll issue4/P adopt boarder.asp). This breakthrough model was the brainchild of floor nurses trying to assist the ED with the flow of admitted patients.

5.6 Obtain the Active Engagement of Hospital Physicians in Flow Initiatives

No hospital will be completely successful in reengineering their ED and inpatient throughput without active medical staff involvement. Nurses and managers can only reengineer to a certain level of operations that they control. The key to medical staff collaboration on this topic is to look for the "win–wins." Being armed with good data is also a must.

Most physicians are aware that delays in hospital discharges are likely to be a reason for the shortage of inpatient beds on any given day. But what most physicians do not realize is that it is likely that less than 5 % of the medical staff account for 70 % of the late discharges. In any hospital, a small number of physicians do not make their rounds until after 4:00 p.m., preferring instead to clear their office of scheduled patients or perform elective procedures in the morning. Most medical staff members, when presented with the data, are shocked that those few physicians are a substantial reason for why other physicians cannot get a bed for their patients or why patients are boarding for long hours in the ED.

Other common medical staff steps to improve capacity and flow include:

- Establishing a hospitalist program
- Creating inpatient care maps for common admission diagnoses
- Hiring nurse practitioners and physician assistants to assist with the discharge process
- Clinically based case managers to facilitate time drivers
- Establishing and enforcing admission and discharge policies for the telemetry and ICU units
- Conducting length-of-stay studies by physician to look for outliers
- Conducting time-of-discharge studies by physician to look for outliers
- Developing a "bed czar" position to bird dog key bed bottlenecks and to assure appropriate bed utilization
- Evaluating day-of-discharge ancillary test needs and adjusting the schedule to assure that results are on the chart during early-morning discharge rounds.
- Conducting "hallway" market research studies of medical staff members on what can be done to improve length-of-stay and day-of-discharge timing.

5.7 Expedite the Unit as a Transition to Other Services

The most common complaint by ED practitioners is delays in the patient's admission, particularly in locating and moving a patient to a bed. The ED is merely a transitional treatment site, with the disposition of the patient to another treatment location or to discharge being the end point in the ED process. The same is true for an ICU patient that is waiting for a telemetry bed or the telemetry patient waiting for the medical/surgical bed. Delays occur not only with diagnosis and treatment of the patient but also in moving the patient from the unit to another point of service in the hospital.

5.7.1 Create a Capacity Control and Communication Center

Most hospitals need more robust real-time capacity management strategies. Changes are needed to assure that all admissions and discharges are coordinated and the capacity managed through a single command center that is supported by real-time bed tracking software. A data-driven capacity command center limits the existence of the so-called phantom bed process of the patient needing beds, but the unrecognized bed that has just been cleaned on one unit or an open bed being held all day for the surgery patient and therefore not listed as available will all be valued and appropriately used in real time. Capacity command center with the appropriate IT technology interfaces can also identify the “mission-critical” beds that are the chief source of today’s bottleneck (e.g., telemetry) and target resources to remove the bottleneck (e.g., STAT bed clean teams). This change should come in the form of a centralized capacity control center that coordinates all bed requests and all discharges and monitors the bed turn and placement process. This center may also manage the logistics of tertiary referrals.

5.7.2 Establish Discharge Times from the Patient Care Units Ahead of the Busy Admit Times from the ED

For most hospitals, the ED is the chief source of their admissions. Delays of inpatient beds result when discharge times on inpatient care units do not precede busy ED admit times. It is not unusual for hospitals to have the bulk of their hospital discharges occurring after 3:00 in the afternoon and many not until 5:00–7:00 pm. Patients waiting for admission are queued and must wait to be transferred to a department where patients are still occupying beds. Analyzing data on the peak admit and discharge times and creating robust medical and nursing staff initiatives and product lines for the ED and patient floors can help to eliminate this problem. A new concept called “slotting” or scheduling discharges for specific times throughout the day may also be helpful as an adjunct to this effort.

5.7.3 Forecast Inpatient Bed Needs

While the ED is often the chief source of admitted patients, it is rarely valued as an important contributor to the overall hospital’s function and, more importantly, the ED inpatient bed demand is rarely anticipated. A focus is needed on proactive bed control strategies to respond to predictable forecasted need. The unit staff of the anticipated destination for the patient admitted from the ED experiences the arriving patient as a new demand on its resources. This demand can be handled more smoothly if that unit can be given advance warning of the arriving patient. Staff at this arriving location can then prepare their system for the arrival of the patient. Establishment of a “bed-ahead” system is also an efficient way to transition patients when there is forecasted demand. With this system, the receiving unit anticipates demand and has an open bed available in advance of the request from the ED.

5.7.4 Develop Refined Bed Control and Surge Protocols

Hospitals struggle with daily bed crunches, but even if these were repaired, it is rare that hospitals, outside their disaster protocols, have conducted preplanned capacity to address temporary surges such is routinely the case during the annual flu season.

5.7.5 Establish Bed Control Briefings and Action Plans

Hospitals should establish bed control briefings as a true empowered capacity management tool. This includes clearly defined meeting expectations and appropriate and timely attendance, with individual defined preparation steps and meeting response steps. Appropriate and consistent meeting start/stop times, attendance by key staff (case management), and staff attendance that is prepared and exits the meeting with a specific plan consistent with the bed needs of today, valuing predictable ED bed needs, should also be goals. Logistical support for these meetings might include having an established form that calculates and dashboards bed needs and resources. A strongly reinforced characteristic and expectation of these meetings and the staffs that attend would be a "pull" system mentality where each department is reaching out to compromised departments and "pulling" the patient or the resource (e.g., meals) to the next step. Success of this effort may require executive leadership attendance at the initial meetings and accountability for meeting goals for future meetings.

5.7.6 Establish a Hospital Activity Barometer and Surge Action Plan

Hospitals should develop predefined roles for each department that measures current workload and functionality and also establishes preplanned activities should there be temporary surges. This barometer should assure drilled-down capacity-building strategies for each department within the hospital. This written action plan would have detailed steps to be taken by each department to proactively respond or react to key capacity variables based on the color-coded need at the time. This could even include fundamental changes such as dispensing with fundamental hospital-wide housekeeping functions (cleaning offices) to reallocate to resources to STAT Bed Clean Teams or perhaps canceling routine meetings and having executive staff transport patients.

5.7.7 Revitalize the Role of the House Supervisor

Most house supervisor roles were designed to assist with bed management and bed allocation process, but so many duties have been added to that position, and they are supported with so little technology that the bed allocation role itself often becomes the bottleneck. Hospitals should alter the house supervisor role to assure that capacity management is a priority and reallocate routine functions to other appropriate staff. For example, if it is determined that the house supervisor is spending significant time on staffing challenges during compromised bed days, those functions should be permanently or temporarily reallocated to other staffs. Routine bed requests that have predictable and protocol-driven responses could be delegated to the Bed Command Center with only conflicts and resource challenges brought to the attention of the house supervisor. The variability in house supervisor roles, skills, and delivery should also be studied and addressed. Revising the communication devices should also be considered (e.g., cell phones versus pagers).

5.7.8 Develop Improved Interfaces with Outside Hospital Resources

One of the sources of delays for many hospitals is getting access for discharged patients to nursing homes, rehabilitation centers, home health, and other outpatient resources. The potential exists here to “slot” or schedule a time for these nursing home admissions. There are even some hospitals that are leasing nursing home beds in advance to assure that the forecasted patient demands are met. These interfaces need to be evaluated, barriers removed, and access improved for the patient flow process to improve.

6 Resolving Capacity and Flow Problem Drivers

There are a variety of methodologies for assessing, developing a plan, and implementing a change process that impact a healthcare provider’s capacity and patient flow. The most effective methods typically center on the following key action steps.

6.1 Conducting a Diagnostic Study

Key to a healthcare provider’s success to improving flow and thus increasing capacity is to know where the constraints and bottlenecks are. The IHI uses the mantra: “How much of the time do we get it right?” in terms of moving patients through the system (Haraden et al. 2004). The IHI model asks two questions:

- Do you park more than 2 % of your admitted patients at some time during the day for at least 50 % of the time? These patients may be “boarding” in the ED, waiting in the admitting office, holding in post-anesthesia recovery, or even in a private doctor’s waiting room or even a nursing home waiting for an inpatient bed.
- Does your hospital have a midnight census of 90 % or greater of your bed capacity more than 50 % of the time? A high midnight census is likely to be symptomatic of a bottleneck for beds as there is limited capacity to admit new patients in the evening or the morning hours, a considerably high-bed-demand period.

Parking patients and high midnight census are clear indicators that the hospital is struggling with flow problems. Sometimes the solution can be as simple as "smoothing" the capacity and demand, reducing workflow variations or better managing the rest of the chain of resources inside and outside the hospital (e.g., home healthcare, nursing homes) to reduce peaks and take advantage of low-demand periods.

6.2 Measuring and Understanding Variation

Variation, while ever present in healthcare delivery systems, when left unchecked is tyranny. The key is to understand and manage the correct type of variation. Random or the so-called natural variation is the kind of variation that cannot be controlled. As an example, the types and severity of disease processes typically cannot be controlled unless the hospital is a specialty hospital. ED arrivals cannot be controlled unless the hospital is controlling a portion of those arriving by ambulance through ED diversion. Some forms of natural variation can be managed such as normal distribution of staff skill sets or care gaps that might be impacted by care maps or additional education.

Nonrandom variation or artificial "variation" can be controlled and in many cases must be eliminated for a healthcare delivery system to be optimized. This is a variation that is artificially introduced into the healthcare delivery system. Examples of artificial variation include the practice of scheduling elective surgeries to peak during the middle of the week but to dramatically decline on Friday afternoons. Practitioner skill sets outside the normal curve or methods or the delayed timing of physician discharge-day rounding on patients will add artificial variation to the patient flow process. Hospitals that do not have published discharge times on the day of discharge or that do not manage their published time also add artificial variation and thus introduce bottlenecks into the discharge process and ultimately to the entire healthcare delivery flow process. Another common source of artificial variation in a hospital is liberal admission and discharge practices amongst physicians to the telemetry or intensive care units or the lack of published or managed admission/discharge criteria for those same units, which permits significant variation and artificially limits other appropriate patient access to these beds. Again, the resources are there, but they are artificially being limited based on variations in practices, policies, or procedures.

6.3 Develop Interventions That Address the Key Problems

Understanding and measuring the constraints and bottlenecks within a healthcare delivery system is a critical first step and is key to the success of any effort to optimized flow and capacity. It is important that the providers solve the right problems with sufficient resources. However, it is not uncommon for a provider to try to solve a problem with the wrong intervention. It is also very common for a hospital to react to problems by merely adding beds or staffs, only to find that theses beds get filled fast or the staff used more without solving the real problem. The real villain for many of the inpatient units is the lack of written and supervised clinical entry or exit criteria, thus inviting overutilization of these critical resources. Merely adding beds or staffs invites continued overutilization. Another example in the ED is the frequent mentioning of slow laboratory or radiology test turnaround times. If the hospital is successful in changing its laboratory CBC turnaround time or speeding up the CT scan test results from radiology, it could find out that the problem was not within the laboratory or the radiology departments but rather with delays and bottlenecks surrounding the laboratory or the radiology process. For example, an ED physician may hold six patient care charts and write orders for all six patients before handing these six charts to the one unit secretary to "order" the test, thus artificially batching orders that can only be ordered through the computer one at a time. Another example might be test results that are sitting in the ED printer waiting to be picked up and inserted into a chart for the physician to read. Another common problem is patient access. The laboratory may respond quickly to the ED but often cannot get to the patient because radiology is with the patient, or in another case the CT scanner may be available but there is no one to transport the patient.

6.4 Using Accelerated Implementation Processes to Assess the Impact of Interventions and Then to Roll Out Successes to the Entire Enterprise

Many healthcare providers use traditional committee structures and protracted time frames to implement their interventions on flow and capacity. While some of these providers have some limited success, often successes are not sustained or they are so fact specific (e.g., based on today's volume or rate-limiting factors) that once the underlying assumptions change, the intervention does not have the impact once hoped for. In addition, traditional committees that can take a year or more to study the issues and implement their change processes are hampered by changing staff members, attending issues, and even having the underlying problem change.

Even as such, most of the processes of change are not just changing the processes and policies but getting the people to move with these changes. To truly engage hospital staff requires a bottom-up and not a top-down approach to problem identification, change implementation, and sustenance. Most staff members will report the many consultant reports "that have sat on the shelf" or the many times that "administration did not listen" to them or have not made a commitment to "fixing the real problems."

Using an accelerated implementation process fundamentally addresses these issues by creating a stronger staff-driven change process and then empowering the staff to make the changes. In addition, the staffs are given tools to implement small changes immediately and test these changes to make sure that they are successful and solve the underlying problem.

The keys for rapid capacity and flow improvement are accelerated implementation teams called high impact teams (HIT) using the rapid cycle testing (RCT) method of change implementation.

6.4.1 High Impact Teams

HIT are a hybrid version of other accelerated implementation teams (see www.teachmeteamwork.com/docs/ toplOteam.pdf). The core structure of HIT is different from traditional healthcare committees in every way. First, they are not committees but small collaboratives of line and middle management staff who are representative of discipline and the expertise of "real-world" problems and solutions for specific issues (e.g., bottlenecks in the ED triage area, backups in surgery, or late discharges of patients from the hospital to go home). Second, these teams are fully empowered to make change. They need not ask permission, seek authority, or go through a line of command to implement the small tests of change that they will be empowered to make. Typically these teams operate for a specific and abbreviated number of sessions (e.g., five to six meetings) perhaps over 5 days, 5 weeks, or, at the most, 5 months. Their job is to study, brainstorm bottlenecks and interventions, and roll out changes during the period of the five sessions. The reason for the limited number of sessions is that it eliminates much of the "fluff" of a traditional study and change process. With only five sessions, at most each team can only afford to do broad brainstorming of the problems for one or two sessions. Any more and the committee time erodes into precious intervention and implementation process time.

There are also a limited number of members on each team. Each team member is highly leveraged to represent their peers but also any closely associated peer group (e.g., RN for LVNs and unit secretaries for patient care technicians, general diagnostic radiology technologists for CT and other specialty radiology staffs). Why such limited attendance? In practical terms, limiting the numbers limits the number of late arrivals, risk for missed homework assignments, and excessive dialogue and repeat memories during each HIT session.

A typically ED HIT team that is looking to improve ED diagnosis and treatment process might include:

- ED nurses (2)
- ED physician
- Unit secretary
- Laboratory
- Radiology
- Case manager

For a typical inpatient team, the HIT makeup might include:

- Charge nurse from a representative unit
- RNs from representative units (2)
- Hospitalist
- Case manager
- Housekeeping
- Bed control nurse

Typically each team adopts and operates within a type frame of ground rules that are designed to accelerate the process, limit delays, and, more important, encourage breakthrough creativity. The ground rules have been assimilated by asking a number of HIT participants what would make a session successful. A sample set of ground rules for a HIT might look like the following:

- All sessions are just 90 min.
- Sessions start and end on time.
- All pagers and cell phones are turned off or set to vibrate mode. The team member's full attention at the sessions is essential.
- Encourage wild ideas.
- Respect everyone's opinion.
- All sessions are action oriented. There are no minutes, just action plans.
- All team members come prepared with all homework assignments completed.

All of the meetings are carefully scripted and supported with a coach to assure that the progress on the goals of the meetings is being achieved and the ground rules are honored.

An emphasis for the HIT is identifying a small number of interventions that will have the highest yield whether that is impacting the bottleneck, ease of implementation, or cost-effectiveness to achieving the goals of the initiative. These are often referred as the "low-hanging fruit." These four to six targets are designed to limit distraction and the so-called solver world hunger appetite that these teams sometimes have and also further assist with the potential for success as there are likely to be a number of HIT in a hospital-wide initiative. For example, if there are 6 HIT each delivering 6 interventions, this would equate to 36 interventions being rolled in a short period of time, a number that would test any sophisticated hospital's ability to absorb change and understand the results. The reality is that each HIT works at its own pace, and there is generally no crunch of 30 plus interventions occurring at the same time.

Note that there is very little executive management involvement at the HIT level. Key management is typically involved at a steering team level to set broad project goals, to monitor the progress, and to eliminate bottlenecks that might arise about the team's authority, responsibility, and empowerment. For example, an ED HIT may want to study and perhaps roll out a trial of the use of point-of-care laboratory testing in the ED's Fast Track. Point-of-care tests might include bedside analysis and reporting of results of urine pregnancy, blood sugar, hematocrits and hemoglobins, certain chemical tests, and, in the cases of more significant clinical concerns,

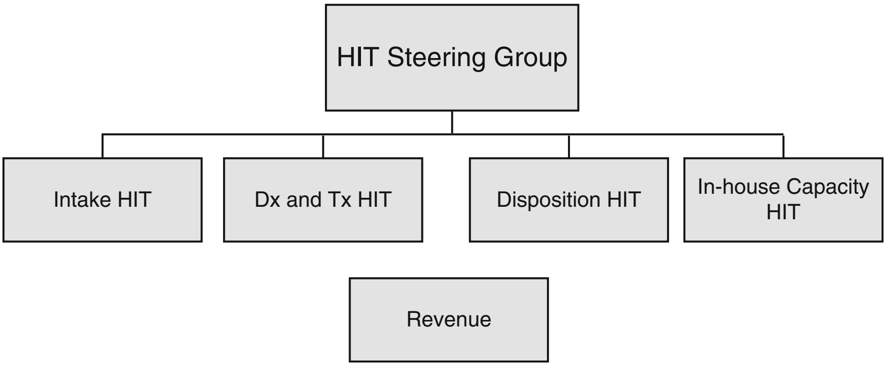


Fig. 3.3 Typical HIT and reporting structure

blood gases and cardiac enzymes. Let us assume, hypothetically only, that the response from the laboratory leadership is "no" to the use of ED bedside testing for a variety of perceived cost and quality control reasons. If that bottleneck cannot be resolved by the HIT members, the matter is referred to the steering team to remove the bottleneck because that team is empowered to study any option and use whatever tool is needed to reach the global goal established by the steering team. The HIT would be allowed to trial the point-of-care testing process to see if it works and if the cost and quality concerns are a reality or perhaps can be mitigated. Figure 3.3 provides a typical organization chart for a HIT. Table 3.1 provides sample ground rules for a HIT.

6.4.2 Rapid Cycle Testing

RCT is a contemporary industrial engineering concept designed to test changes on a small scale to assure clarity of the intervention's assumptions and intervention effectiveness to allow minor customization of the interventions to correct for found timing or intervention sizing issues that might make the intervention more effective. Another key reason for using RCTs is to allow tests on a small change to minimize risk to patient care, flow, or staff adoption. Another important reason for using the RCT process is to help scale the implementation process.

For example, an ED might want to eliminate the patient triage process when beds are available in the back of the ED. The reason might be that the HIT may have identified that triage itself adds 15–25 min of unnecessary delay to the care process, is a big patient dissatisfier, and does not result in safer bed placement or elimination of care processes. In most EDs, if triage were simply and abruptly eliminated, there would be chaos throughout the department and substantial safety concerns among the clinical staffs. Triage is also widely held to be a legally required step in the care process, and many bedside nurses expect it to be completed in many EDs in a comprehensive way. Some nurses might even use the terms "you're going to kill a patient if you eliminate triage."

Table 3.1 Typical HIT meeting ground rules

Typical HIT Ground Rules
Ground rules (approved by each HIT):
1. All meetings will begin and end on time
2. Meetings are limited to 90 min unless permitted by the committee members
3. Beepers and phones are to be placed on "vibrate" mode during the meetings. Only emergencies should be responded to
4. All team members will stay on track. A timekeeper will be used at each meeting, and agenda items will have assigned time limits for discussion
5. All team members will follow through with assignments and come prepared to the meetings
6. All team members will regularly attend meetings. In the unlikely event they cannot attend a designee should be sent in their place
7. No veto power or "sacred cows" during the brainstorming sessions
8. Thinking "wildly" is encouraged during all brainstorming and action plan sessions
9. One idea at a time
10. Defer judgment/respect all opinions
11. Build on the opinions of others
12. Stay focused
13. Titles stay outside the door

Source: The Abaris Group, Walnut Creek, CA

Triage or "to sort" is often mistakenly associated with Napoleon, but rather it was one of his key French surgeons, Dominique Jean Larrey (Richardson 2002), who invented the concept of a sorting tool to be used only when there were insufficient resources in the battlefield for the demand being presented. But triage, as deployed in most EDs, has been found to be a bottleneck itself when it is used when there are sufficient ED beds, for example early in the morning before the volume of ED patients begins to rise. The HIT might desire to trial a "no-triage" protocol concept for a week but would likely face a barrage of staff skepticism. Thus this might be trialed for a day, a shift, or even just several hours. Depending upon the perceived staff concerns, the time period for the trial can be customized, at the very least, to during the portion of a shift that HIT members are on duty, thus willing to commit to the test of change during that period. This might be, for example, for the first 3 h of the shift next Thursday that Dr. Smith and Nurse Jones, both HIT members, are on duty. It is helpful that each team has a target or an aim statement to be used and interventions to be deployed using Nolan's Plan, Do, Check, Act (PDCA) model (Langley et al. 1996) provides a sample of an aim statement.

RCTs are sequentially rolled out using small tests of change. Gradually, but over a relatively short period of time, the trial is adjusted and either abandoned or expanded based on the results of each trial provides an ED example of an RCT to triage a new radiology protocol. RCTs, as a tool, can be applied to a wide variety of issues. For example, an inpatient HIT—after considerable analysis, research of best practices, and brainstorming—may wish to trial an admission/discharge nurse concept on one of the floors to accelerate the time the patient is admitted or sent home, depending on the need at any given time. This is a published, best practice, concept. The concept might be trialed next Monday during the day shift, studying data acquired during the trial and then adjusting the role or the duties on Tuesday for a Wednesday trial. The triage on Wednesday is successful, so it is expanded to include one more med/surgery area on Friday and further expanded until it is fully rolled out the following week. Although this process was changed through small tests, it was fully deployed hospital-wide in less than 2 weeks.

The key to a successful RCT trial is to have the following PDCA questions and supporting data in mind:

- What are we trying to change?
- What changes will have the biggest impact?
- How will we know if the change made a difference?
- Did the change made a difference and if not why?

Having a clear understanding of the goal of the change is critical. Many healthcare providers, while building enthusiasm about a change process or a series of change processes, attempt to just implement changes without carefully thinking them out or having sufficient baseline data to know if the change will be effective or even if the change will affect the originally defined problem. It has often been witnessed that well-meaning health providers burn out in frustration due to:

- Tying to solve the wrong problem
- Changing processes that will not impact the targeted problem
- Insufficient use of the change process to impact the problem
- Not knowing if they have solved the problem due to insufficient data collection

Testing on a small scale also allows the collection of small data sets to define the baseline and to measure movement. Oftentimes, staff may not feel that they have access to baseline data, or data are perceived to be too difficult to get (e.g., ED time flow data from a complex patient-tracking system). The solution is the manual collection of mini samples. In industrial engineering terms, you "collect big data for big decisions and little data for little decisions." This means, a small sample should suffice for testing on a small scale.

In industrial engineering circles, a sample of 30 events, if properly collected without bias (e.g., 30 consecutive events), should be sufficient to measure baseline current "as-is" status and 30 events during the RCT should measure impact. For example, a HIT might wish to explore speeding up the process of when a patient leaves a hospital bed to when the bed is put into the computer for purposes of notifying housekeeping to clean the bed. Perhaps the trial intervention is to empower the charge nurse to also put the discharge order into the system due to a perceived bottleneck of a large number of discharge orders coming in at one during the peak discharge times. To obtain baseline data, the "patient departure time to time noted in the computer" is hand collected on 30 consecutive patients during Tuesday's peak discharge time and then the new charge nurse scope implemented on Wednesday, for 30 consecutive patients; those times are hand recorded and compared to the Tuesday baseline experience.

The concepts of HIT, RCTs, and other accelerated implementation processes not only create a toolkit for radically improving the flow of patients and the capacity of a healthcare provider but also revolutionize and energize every aspect of the decision process and every stakeholder in that process. HIT are also simple to implement and can be used with a wide variety of issues, including revenue management, building of designs, and also major emergency response planning.

7 Conclusions and Extensions

Healthcare providers are increasingly compromised with growing demand and limited resources. The resulting impacts are excessive waiting, prolonged patient flow, and customer dissatisfaction. A key ingredient to improving the healthcare delivery system is to better understand the dynamics, drivers, and myths that impact healthcare patient flow and capacity. In addition, healthcare providers should conduct independent and objective analyses of their particular bottlenecks and process and flow constraints to assure that steps are taken that will address the real problems. Providers should also understand that key industrial engineering tools will create high-leverage solutions, many of which do not require more staffs or beds but rather the reallocation of staffs and beds to demand. Traditional change processes are often slow and ineffective and thus frustrating to providers who are trying to make a difference. Key to implementing and sustaining success is the use of accelerated implementation models, such as the HIT and RCT models noted in this chapter.

References

Agency for Healthcare Research and Quality. (2004). *The Uninsured in America, 1996–2004*. Rockville, MD: Agency for Healthcare Research and Quality.

Aiken, L. H., Clark, S. P., Sloane, D. M., Sochalski, J., & Silber, J. H. (2002). Hospital nurse staffing and mortality, nurse burnout and job dissatisfaction. *JAMA, 288*(16), 1987–1993.

American College of Emergency Physicians (ACEP). (1993). *Emergency Department Design*. Dallas, TX: American College of Emergency Physicians (ACEP).

American Hospital Association. (2005). *AHA Hospital Statistics 2005 Edition*. Chicago, IL: American Hospital Association.

Berwick, D. M. (1996). A primer on leading the improvement of systems. *British Medical Journal, 312*(March 9), 619–622.

Chase, R., Aquilano, N., & Jacobs, F. R. (2001). *Operations Management for Competitive Advantage*. New York: McGraw-Hill Irwin.

Disney Institute. (2001). *Be Our Guest: Perfecting the Art of Customer Service*. New York, NY: Disney Institute.

Graff, L. G. (Ed.). (1998). *Observation units: implementation and management strategies*. Dallas TX: American College of Emergency Physicians.

Haraden, C., Resar, R., Henderson, D., et al. (2004). *Capacity management breakthrough strategies for improving patient flow, frontiers of health services, management*. Chicago, IL: American College of healthcare Executives. 20, No. 4.

Healthcare Financial Management Association. (2004). *Financing the future survey*. Westchester, IL: Healthcare Financial Management Association.

Langley, C., Nolan, K., Nolan, T., Norman, C., & Provost, L. (1996). *The Improvement Guide: A Practical Approach to Improving Organizational Performance*. San Francisco, CA: Jossey-Bass.

McCaig, L. F., & Burt, C. W. (2004). *National Hospital Ambulatory Medical Care Survey: 2002 Emergency department summary* (Advance data from vital and health statistics, Vol. 340). Hyattsville, MD: National Center for Health Statistics.

Melnick, G., Bamezai, A., Green, L., & Nawatje, E. (2002). *California emergency department capacity and demand*. Oakland, CA: California Healthcare Foundation.

Patel, P. B., Derlet, R. W., Vinson, D. R., Williams, M., & Wills, J. (2006). Ambulance diversion reduction: the Sacramento solution. *Journal of Emergency Medicine, 24*(2), 206–213.

Peterson, C. (2001). Nursing shortage: not a simple problem—no easy answers. *Online Journal of Issues in Nursing, 6*(1).

Regenstein, M., Nolan, L., Wilson, M., Mead, H., & Siegel, B. (2004). *Walking a Tightrope: The State of the safety net in ten U.S. communities*. Washington, DC: George Washington University Medical Center Department of Health Policy.

Richardson, R. (2002). *Surgeon to Napoleon's Imperial Guard*. London: Quitter Press Ltd.

Robeznicks, A. (2005). *A call to action*. Chicago, IL: Modern Healthcare.

Sochalski, J. (2002). Nursing Shortage redux: turning the corner on an enduring problem. *Health Affairs, 21*(11), 157.

Stout, J. (1986). Ambulance Systems Design. *JEMS, 11*, 85–96.

The Lewin Group. (2002). *Analysis of AHA ED and hospital capacity survey*. Chicago, IL: The Lewin Group.

U.S. General Accounting Office (GAO). (2003). *Hospital emergency departments: crowded conditions vary among hospitals and communities*. Report No. GAO-03-460. Washington, DC: U.S. GAO.

Williams, M. (2004). Materials management and logistics in the emergency department. In M. Rice (Ed.), *Emergency medicine clinics of North America* (Vol. 22, No.l). Philadelphia, PA: Saunders.

Part II
Crowding and the Consequences of Delay

Chapter 4
Emergency Department Crowding: The Nature of the Problem and Why It Matters

Kirk Jensen

Abstract The current state of the American healthcare system has focused national attention on the core issues of patient satisfaction, patient safety, and patient flow. Acute-care settings are often plagued with waits, delays, and dissatisfaction. Nowhere is this more observable and the impact more palpable than in hospital emergency departments (EDs). The profession, the public, and the press have highlighted this area as an important healthcare system concern. There is a disparity between the high level of interest in the core issues, an appreciation of the true nature of the problem, and the ability to effectively implement the solutions. Even in the hospitals and emergency departments where the required knowledge and competencies are available, the ability to effectively integrate them into a functioning and effective improvement program may not exist.

Emergency departments are complex operational micro-systems. This chapter outlines and defines key challenges, opportunities, and solutions surrounding emergency department crowding. It also provides a wide-ranging overview of the key drivers behind emergency department crowding and the opportunities for improvement, including important safety, service, and workforce implications. Clinical, volume, workforce, and system issues all play a role in solving crowding.

Keywords Admissions • Boarding • Behavioral health boarders • Capacity • Crowding • Emergency services • ICU utilization • Length of stay • Non-urgent • Patient safety • Pediatrics • Physician burnout • Physician productivity

K. Jensen (✉)
BestPractices, 10306 Eaton Place Suite 180, Fairfax, VA, USA
e-mail: kjensen@best-practices.com

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_4, © Springer Science+Business Media New York 2013

1 Introduction

As defined by the American College of Emergency Physicians (ACEP) Crowding Resources Task Force in 2002, crowding is "*a situation in which the identified need for emergency services outstrips available resources in the ED. This situation occurs in hospital EDs when there are more patients than staffed ED treatment beds and wait times exceed a reasonable period. Crowding typically involves patients being monitored in non-treatment areas (e.g., hallways) awaiting ED treatment beds or inpatient beds. Crowding may also involve an inability to appropriately triage patients, with large numbers of patients in the ED waiting area of any triage assessment category*" (Case et al. 2004). Studies often use the mean occupancy rate, or the "number of patients in the emergency department divided by the number of treatment spaces," to measure crowding in the ED (Fiore 2012).

ED crowding has plagued hospitals for some time, with the first statewide conference held in New York in 1987 to address the growing issue. Initial efforts to stem crowding in the ED attributed growing "non-urgent" use of the ED as the primary cause and also worked to reduce congestion through ambulance diversion from crowded hospitals to non-crowded facilities. More recently, however, studies of ED crowding have centered on "boarding," or "the practice of holding admitted patients in the emergency department when there is no proper place for them in the institution," while all but downplaying "non-urgent" use of the ED as a significant contributor to crowding (Asplin et al. 2008).

Emergency departments continue to experience crowding in conjunction with expanding roles of the ED, leading to a rising number of emergency visits and more unscheduled visits to the emergency room (Asplin et al. 2008). For instance, visit rates in the ED rose by more than 30 % from 1997 to 2007, and the emergency department is increasingly serving as a "safety net for underserved patients, particularly adult Medicaid beneficiaries" (Schuur and Venkatesh 2012). Moreover, ED crowding has garnered considerable attention outside of the medical community. According to a recent poll by ACEP, roughly seven out of ten Americans "believe emergency departments were approaching a crisis due to overcrowding" (Blum et al. 2005).

ED crowding is not confined to a particular region, or to urban rather than rural hospitals, but is instead a ubiquitous problem plaguing hospitals throughout the nation. According to the Agency for Healthcare Research and Quality, among 3,833 hospital EDs studied, roughly half "report operating at or above capacity" (USDHHS 2012). Indeed, recent studies demonstrate that throughout the nation, 91 % of ED directors "reported crowding to be a problem," with roughly 40 % experiencing it on a "daily basis" (Schneider et al. 2003). However, while ED crowding is a characteristic of both urban and rural hospitals, doctors in urban and rural hospitals differ over the primary concerns for patient safety risks. In urban settings, physicians reported crowding as "the greatest safety concern," while rural emergency physicians centered on "consultant availability" (Moskop et al. 2009).

Blum stated that "crowding affects everyone—young, old, rich, poor. It is happening in cities, suburbs and rural areas. It occurs at teaching and non-teaching hospitals" (Blum et al. 2005).

Simply put, crowding exists when there is insufficient capacity to meet "the demands of the next patient who needs emergency care" (Asplin et al. 2008). In studies of this phenomenon, crowding is often placed in a conceptual framework of input factors, throughput factors, and output factors to identify common characteristics and potential solutions to this problem. Nathan R. Hoot, PhD, and Dominik Aronsky, MD, PhD, conducted a review of research on ED crowding, published in the *Annals of Emergency Medicine*, that identified commonly "studied causes of ED crowding" (Hoot and Aronksy 2008, p. 130). According to Hoot and Aronsky, common input factors studied consist of "non-urgent visits," "frequent-flyer patients," and "influenza season," while common throughput factors focused on "inadequate staffing." Lastly, frequently studied output factors in studies of ED crowding were "inpatient boarding" and "hospital bed shortages" (Hoot and Aronksy 2008, p. 130).

2 What Impacts Emergency Department Length of Stay and Crowding

Several factors impact ED length of stay (LOS) and crowding, including walk-ins, admitted/discharged patients, boarders, and physician productivity. As previously mentioned, while "non-urgent visits" may have been a commonly studied input factor of ED crowding, more recent studies have largely downplayed its effect on crowding (Pitts et al. 2012). Specifically, non-urgent visits to the emergency room have been shown to have virtually no effect on crowding. Instead, these studies have focused on the practice of boarding patients and hospital bed shortages as the primary factors in ED crowding.

2.1 Boarders

As previously mentioned, the practice of boarding has been recognized as a major contributor to crowding in the emergency department. Boarding is defined as the practice in which "patients are held or 'boarded' in emergency departments waiting for inpatient beds in the hospital" (Blum et al. 2005). The Institute of Medicine has described boarding as one of the most significant contributors to crowding in the emergency department, declaring that hospitals must "end the practice of boarding patients in the ED and ambulance diversion, except in the most extreme cases, such as a community mass causality event" (Asplin and Magid 2007, p. 274). One report estimated that "if no inpatients were housed in the EDs studied, the number of crowded departments would decrease by more than 30 %" (Schneider et al. 2003). While reducing the number of boarded patients is certainly more difficult than simply housing them, it nevertheless demonstrates the significant impact this practice has on ED crowding.

2.2 ICU Utilization and Emergency Department Capacity

Researchers have also conducted studies to examine the effect of greater ICU capacity on LOS in the hospital. One noteworthy study, conducted by K. John McConnell et al., revealed an association between increased ICU capacity and ambulance diversion. Specifically, their report noted that an "increase in ICU beds from 47 beds to 67 beds" led to "statistically significant decreases in ambulance diversion" (McConnell et al. 2005, p. 476). McConnell et al., however, discovered that increased ICU capacity did not lead to a significant decrease in the ED LOS and, as a result, increasing the ICU capacity is unlikely to lead to a decreased level of boarding in the ED (McConnell et al. 2005). Similar conclusions have been made in regard to the number of beds in the ED and crowding.

2.3 Physician Productivity

Frequent interruptions resulting from an overcrowded emergency department have a negative impact on physician productivity. Unsurprisingly, error rates increase with "distractions and interruptions" in the nonmedical workplace, and as a result, such instances in the emergency department are likely a source of "medical errors attributed to the ED" (Schneider et al. 2003, p. 171).

2.4 ED Beds Versus Admissions

Much research has been devoted to reducing crowding in the emergency room through an increase in the number of beds in the emergency department. Studies demonstrate that simply increasing the number of beds in the ED does little, if anything, to reduce the level of crowding. Rahul Khare et al., in their study of the board times in the ED, effectively summarized the effects of increasing the number of beds on LOS, stating:

> In an analogous manner, one can image the ED to be a pipe and patients as water passing through the pipe. If we enlarge the diameter of the middle of the pipe but leave the end the same, we analogously have increased the number of ED beds without improving the departure rate (Khare et al. 2009, p. 581).

In sum, improving the departure rate of patients, rather than increasing the number of beds in the ED, leads to lower levels of ED crowding, thereby further underscoring the effects of boarding patients on crowding.

3 How Do We Measure Crowding?

Crowding in the emergency department has been difficult to measure. In a review of models to describe and predict "ED patient loads and crowding," Jennifer Wiler et al. (2011) weighed the pros and cons of models and formulas, including formula-based equations, regression models, and queuing theory-based models, among others. Overall, each method provided many benefits to predicting ED crowding times, while each suffered from notable limitations. Most important, however, is simply the ability to predict emergency department peak times, even if it is only a rough estimate, and using this valuable information to prepare for peak loads and to improve patient flow throughout the hospital.

3.1 Economic Impact and the Opportunity Cost of Crowding

Crowding can have a significant financial impact on hospitals. Several studies have examined or modeled the opportunity cost of boarding patients, prolonged ED LOS, diversions, and patient elopements. One has only to plug in the revenue (contribution margin) of the patients at one's own healthcare facility that fits each of these categories to get an approximation of the substantial revenue that can be generated or recovered by solving ED crowding (Pines et al. 2011; Falvo et al. 2007a, b; Bayley et al. 2005).

3.2 Behavioral Health and Crowding

Boarding is especially prevalent for psychiatry patients, including children. A survey conducted by ACEP revealed that among the 328 ED directors who responded, nearly 80 % reported "their hospital 'boards' psychiatric patients in the emergency department." Even more startling, however, were the number of hospitals without beds dedicated to psychiatry patients. Specifically, of the directors who responded, 60 % admitted that their emergency department did not have an area dedicated to psychiatry patients (EMS n.d.).

3.3 Pediatrics and Crowding

Children make up a significant amount of ED visits, and, as a result, pediatricians can play an important role in reducing crowing in the emergency department. In 2000, of an estimated 108 million ED visits, 30 million were for children 0–18 years of age (Committee on Pediatric Medicine 2004). With this in mind, pediatricians can take steps to help alleviate crowding in the emergency department. Among them, pediatricians can help educate parents, so they “may make better-informed decisions” if they are prepared and “working closely with local institutions and providers of emergency services to ensure coordination of effective primary and subspecialty follow-up care” (Committee on Pediatric Medicine 2004, p. 885).

3.4 Physician Burnout and Crowding

Research has revealed high levels of physician burnout in hospitals across the nation. In general, burnout is “a constellation of symptoms relating to behavior at the workplace,” including “emotional fatigue, depersonalization, lost enthusiasm, and a failed sense of personal accomplishment” (Schattner 2012, p. 2). Researchers at the Mayo Clinic polled 7,288 physicians on their “quality of life and job satisfaction,” and the results indicated that 46 % of respondents exhibited one burnout symptom or more (Schattner 2012, p. 1). Burnout is important for the general health and well-being of not only physicians but also patients entrusted in their care. Specifically, burnout can lead to reduced “professionalism and lessen the quality of care” (Schattner 2012, p. 1).

3.5 Patient Safety and Crowding

ED crowding has been shown to have a significant effect on patient safety, which decreases with increasing LOS in the hospital. Through a study of adult admissions in California hospitals in 2007, including 995,379 admissions at 187 different hospitals, Sun et al. demonstrated that “high ED crowding was associated with 5 % greater odds of inpatient death, 0.8 % longer hospital length of stay, and 1 % increased costs per admission” (Sun et al. 2012). Increased time in the emergency department has been connected to decreasing quality of care. According to registry data, hospitals that more strictly follow guidelines have lower mortality rates (Hollander and Pines 2007). For instance, “patients who waited more than eight hours for a bed received care inferior to that of patients who waited less than four hours for a bed” (Hollander and Pines 2007, p. 497). Consequently, crowding and increased time spent in the emergency department have been linked to increased patient mortality and decreased quality of care, respectively.

4 What Can Be Done to Solve Emergency Department Crowding

Consensus reviews on practical solutions to ED crowding emphasize the necessity of looking at the "hospital-wide nature of patient flow problems" rather than focusing solely on the emergency department (Asplin and Magid 2007, p. 274). In the 2008 ACEP Task Force Report on Boarding, ACEP made several recommendations centered on patient flow in hospitals as solutions to ED crowding. In particular, ACEP created an extensive list of internal emergency department actions and hospital-wide solutions to improve patient flow and thereby reduce crowding in the emergency department, including implementing bedside registration, limiting triage to "what is crucial and [bypassing] triage altogether when beds are available," using scribes for documentation, and a "fast track area" to remove patients from the "mainstream of patients," among others (Asplin et al. 2008, pp. 10–11). In regard to hospital-wide recommendations, ACEP advised promoting an "institutional awareness" of the danger of ED crowding, a shift to a "24/7 operational culture," coordination of scheduling for elective patients and surgical cases, and need to "address delays in moving emergency patients to the hospital caused by waiting for nursing reports" (Asplin et al. 2008, p. 12).

Moreover, in their review of potential solutions to emergency department boarding, Rabin et al. suggested that if current strategies do not work, "legislation may be required to effect meaningful change" (Rabin et al. 2012, p. 1757). In their review, Rabin et al. called on health policy leaders and CEOs to make a serious commitment to reducing crowding in the ED. Further, they point to effective legislation throughout the world in reducing boarding. For instance, Britain implemented a "Four-Hour Rule" in which "98 percent of emergency departments patients be seen, treated, and either discharged or placed in an inpatient bed within four hours" with hospital CEOs held responsible for meeting this demand, leading to "96 percent of British patients were either moved to inpatient beds or discharged in four hours," as of 2010 (Rabin et al. 2012, p. 1762).

5 Conclusion

There are many challenges facing those who work in emergency departments. Emergency departments, as outlined in this chapter, are for many reasons, encompassing the simple and complex, overcrowded and at times unable to safely and speedily meet the demands for their services. The problems of overcrowding and diversion will continue to be impacted by forces outside the control of any one hospital or ED. Armed with a thorough appreciation of the causes and consequences of ED crowding, understanding the drivers of ED crowding, and focusing on solutions within the control of the emergency department and the hospital are the first steps that healthcare teams can take on the journey of optimizing the quality, safety, and efficiency of the service.

References

Asplin, B. R., & Magid, D. G. (2007). If you want to fix crowding, start by fixing your hospital. *Annals of Emergency Medicine, 49*(3), 273–274.

Asplin, B., Blum, F., Broida, R., Bukata, W., Hill, M., Hoffenberg, S., et al. (2008). Emergency department crowding: High-impact solutions. ACEP boarding task force, 2–14.

Bayley, M. D., Schwartz, J. S., Shofer, F. S., Weiner, M., Sites, F. D., Traber, K. B., & Hollander, J. E. (2005). The financial burden of emergency department congestion and hospital crowding for chest pain patients awaiting admission. *Annals of Emergency Medicine, 45*(2), 110–117.

Blum, F., Keaton, B., Kellermann, A., Schafermeyer, S., Suter, R. et al. (2005). *Meeting the challenge of emergency department overcrowding/boarding*. Report from a roundtable discussion, 4–11, American College of Emergency Physicians, Irving, TX.

Case, R. B., Fite, D. L., Davis, S. M, Hozhaj, S., Jaquis, W. P., Seay, T., & Yeh, C. (2004). Emergency department crowding information paper. Emergency Medicine Practice Subcommittee on Crowdin, Amercian Council of Emergency Physicians, Irving, TX.

Committee on Pediatric Emergency Medicine. (2004). Overcrowding crisis in our nation's emergency departments: Is our safety net unraveling? *Pediatrics, 114*(3), 878–888.

EMS. (n.d.). Psychiatric patients, including children, routinely boarded in emergency departments. EMS press release. Retrieved http://www.ems1.com/ems-products/press-releases/406718-Psychiatric-Patients-Including-Children-Routinely-Boarded-In-Emergency-Departments.

Falvo, T., Grove, L., Stachura, R., & Zirkin, W. (2007a). The financial impact of ambulance diversions and patient elopements. *Academic Emergency Medicine, 14*(1), 58–62.

Falvo, T., Grove, L., Stachura, R., Vega, D., Stike, R., Schlenker, M., & Zirkin, W. (2007b). The opportunity loss of boarding admitted patients in the emergency department. *Academic Emergency Medicine, 14*(4), 332–337.

Fiore, K. (2012). ED crowding getting worse. *MedPage Today*. Retrieved http://www.medpagetoday.com/EmergencyMedicine/EmergencyMedicine/33426.

Hollander, J., & Pines, J. (2007). The emergency department crowding paradox: The longer you stay, the less care you get. *Annals of Emergency Medicine, 50*(5), 497–499.

Hoot, A., & Aronksy, D. (2008). Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine, 52*(4), 126–136.

Khare, R. K., Powell, E. S., Reinhardt, G., & Lucenti, M. (2009). Adding more beds to the emergency department or reducing admitted patient boarding times: Which has a more significant influence on emergency department congestion? *Annals of Emergency Medicine, 53*(5), 575–585.

McConnell, K. J., Richards, C. F., Daya, M., Bernell, S. L., Weathers, C. C., & Lowe, R. A. (2005). Effect of increased ICU capacity on emergency department length of stay and ambulance diversion. *Annals of Emergency Medicine, 45*(5), 471–478.

Moskop, J. C., Sklar, D. P., Geiderman, J. M., Schears, R. M., & Bookman, K. J. (2009). Emergency department crowding, part 1—concept, causes, and moral consequences. *Annals of Emergency Medicine, 53*(5), 605–611.

Pines, J. M., Batt, R. J., Hilton, J. A., & Terwiesch, C. (2011). The financial consequences of lost demand and reducing boarding in hospital emergency departments. *Annals of Emergency Medicine, 58*(4), 331–340.

Pitts, S., et al. (2012). National trends in emergency department occupancy, 2001 to 2008: Effect of inpatient admissions versus emergency department practice intensity. *Annals of Emergency Medicine, 60*(6), 679–686.

Rabin, E., Kocher, K., McClelland, M., Pines, J., Hwang, U., Rathlev, N., et al. (2012). Solutions to emergency department 'boarding' and crowding are underused and may need to be legislated. *Health Affairs, 31*(8), 1757–1766.

Schattner, E. (2012). The physician burnout epidemic: What it means for patients and reform. The Atlantic Monthly Group, 1–3

Schneider, S., Zwemer, F., Doniger, A., Dick, R., Czapranski, T., & Davis, E. (2001). Rochester, New York a decade of emergency department overcrowding. *Academic Emergency Medicine, 8*(11), 1044–1050.

Schneider, S. M., Gallery, M. E., Schafermeyer, R., & Zwemer, F. L. (2003). Emergency department crowding: a point in time. *Annals of Emergency Medicine, 42*(2), 167–172.

Schuur, J. D., & Venkatesh, A. K. (2012). The growing role of emergency departments in hospital admissions. *The New England Journal of Medicine, 367*(5), 391–393.

Sun, B. C., Hsia, R. Y., Weiss, R. E., Zingimond, D., Liang, L. J., Han, W., et al. (2012). Effect of emergency department crowding on outcomes of admitted patients. *Annals of Emergency Medicine, 61*(6), 1–7.

U.S. Department of Health and Human Services. (2012). Hospital emergency departments get tips from AHRQ on how to reduce crowding and better triage patients. Agency for healthcare research and quality. http://www.ahrq.gov/news/newsletters/research-activities/mar12/0312RA1.html.

Wiler, J. L., Griffey, M. D. & Olsen, T. (2011). Review of modeling approaches for emergency department patient flow and crowding research. *Academic Emergency Medicine, 18*(12), 1371–1379.

Chapter 5
The Consequences of Emergency Department Crowding and Delays for Patients

Megan McHugh

Abstract Emergency department crowding and the consequential delays in patient care have been long-standing problems, but as recently as 10 years ago, there was little evidence of their impact on patients. Since that time, there has been a substantial and convincing number of studies published indicating that crowding and delays compromise access to emergency care, quality of emergency care delivered, and patient outcomes, including survival. Further, there is evidence that these threats to patient well-being are more prevalent among vulnerable populations, specifically racial and ethnic minorities. This chapter presents results of recent studies that investigated the relationship between emergency department delays and crowding and patient outcomes. It concludes with a discussion about strategies to reduce crowding and delays.

Keywords Emergency care • Delays • Crowding • Patient outcomes

1 Introduction

Patients wait longer for care in the emergency department (ED) than they used to. The average wait time to see an ED provider in 2009 was 58.1 min, up from 46.5 min in 2003 (Hing and Bhuiya 2012). Even severely ill patients, for example those with acute myocardial infarction, wait longer for care (Wilper et al. 2008). Indeed, the percentage of ED patients who are seen by a physician within the recommended time is declining (Horwitz and Bradley 2009), and less than one-third of hospitals achieve recommended wait times for 90 % or more of their patients (Horwitz et al. 2010). As an example, as many as one-third of EDs do not meet 90-min treatment targets for heart attack patients ("One-third of EDs may fail

M. McHugh (✉)
Northwestern University, Chicago, IL, USA
e-mail: megan-mchugh@northwestern.edu

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_5, © Springer Science+Business Media New York 2013

to meet 90-minute target for heart attack patients" 2007). Prolonged wait times have clear implications for quality since many patients come to the ED with time-sensitive conditions. Although prolonged wait times for emergency care may be caused by a number of factors, for example, a shortage of available specialists and administrative inefficiencies, they are frequently attributed to crowded conditions. The causes of ED crowding are multifactorial and include an increase in demand for ED services and higher severity of patients, a commensurate reduction in the number of EDs in the USA, and insufficient hospital bed capacity, which leads to patient boarding (where admitted ED patients wait in the ED for an available inpatient bed) and lower ED capacity for incoming patients (Hoot and Aronsky 2008). Several studies have shown that crowding is prevalent across the USA, particularly in large, urban hospitals (Burt et al. 2006; General Accountability Office 2003, 2009; IOM 2006).

ED crowding leads to treatment delays for both adult and pediatric patients. For example, crowding is associated with delays in antibiotic therapy (Fee et al. 2007; Kennebeck et al. 2011), pain management (Pines and Hollander 2008; Shenoi et al. 2011), and medications for stroke and heart attack (i.e., thrombolysis) (Schull et al. 2004). In a national survey, ED physicians identified ED crowding (and the resulting delays in care) as their top patient safety concern (Sklar et al. 2010). Concerns about delays in EDs have led organizations such as the Institute of Medicine (IOM), Joint Commission, and Institute for Healthcare Improvement to issue recommendations to hospitals to address the problem of crowding. Additionally, crowded conditions and delays in emergency care have even been picked up by the lay press, with sensational headlines, such as "A crowded emergency room can kill you" (*The Washington Post*, December 12, 2012) and "The Diverted Ambulance: How ER Crowding Kills." (*Time Magazine*, June 27, 2011).

This chapter summarizes what is known about the consequences of delays and crowding for patients. In March 2013, a MEDLINE search of English-language articles was conducted using terms for emergency department (e.g., emergency services, emergency, ED); delay and crowding (e.g., length of stay, time, flow, capacity); and patient outcomes (e.g., quality of care, outcome, survival). The review focused on studies that quantified the impact of delays and crowding. Additionally, the search revealed disparities in delays and crowding by race and ethnicity, and those disparities are described. Evidence to date on strategies for addressing crowding and delays is also discussed.

2 Decreased Access to Care

When there are delays to treatment in the ED, patients' ability to access care is compromised. Many patients have a defined, limited period that they are willing to wait (Shaikh et al. 2012). Faced with long wait times or a crowded waiting room, some incoming ED patients may simply leave without seeing a provider. Nationally, approximately 2 % of all people who arrive to an ED seeking care end up leaving without being seen (Niska et al. 2010). Although there may be several factors influencing a patient's decision to leave without being seen, patient surveys consistently show that long wait time is a primary reason for departure (Clarey and Cooke 2011; Shaikh et al. 2012; Wilson et al. 2012). These findings are corroborated by several quantitative analyses showing that ED occupancy rates exceeding 100 % and long ED length of stay are associated with higher rates of patients leaving before being seen (Fernandes et al. 1997; Weiss et al. 2005). According to Monzon et al. (2005) who investigated the incidence and causes of patient elopement, "long wait times result in a system of 'rationing by queuing,' in which the scarce resources of the emergency department are distributed on the basis of how long people are willing to wait to see a physician" (p. 1089).

The elopement of patients from the ED is concerning because of the risk to the patients and the hospitals. Patients who leave without being seen tend to be a vulnerable group of patients. They are more likely to be non-white, covered by Medicaid, and lack health insurance, stable housing, a valid telephone number, and a family physician (Mohsin et al. 2007; Monzon et al. 2005; Pham et al. 2009; Rowe et al. 2006). The majority of patients who leave without being seen are triaged with low acuity scores (i.e., their conditions are less urgent) and seek medical care within a week (Fernandes et al. 1994). However, a sizable minority are in need of urgent medical attention and are ultimately hospitalized or experience adverse outcomes or death associated with delayed care (Baker et al. 1991; Rowe et al. 2006). For example, one study followed 498 patients who left an ED without being seen and found that 60 % received medical care within a week and 14 were hospitalized. Among the 40 % who did not subsequently seek medical attention, one-quarter had been triaged as urgent in the ED and one patient died six days after leaving the ED (Rowe et al. 2006).

ED delays also have implications for patients arriving by ambulance and the availability of ambulance services. When there are delays in ED care or the ED is operating beyond full capacity, it often results in ambulance offload delays, meaning that it takes longer for emergency medical service personnel to transfer a patient to an ED stretcher and for the ED staff to assume responsibility for the patient (Cooney et al. 2011). Offload times can vary from a few minutes to several hours during which the ambulance is out of service to respond to 9-1-1 calls (Eckstein and Chan 2004). To illustrate, in Los Angeles, ambulance crews are expected to transfer a patient within 15 min and to notify dispatchers if they will be unavailable for longer than 15 min. During a 12-month period (April 2001 to March 2002), there were 21,240 incidents in which ambulances were out of service due to lack of availability of an open ED gurney, and the median waiting time per incident was 27 min. This delayed transfer time compromises the capacity of the EMS system to respond to new calls. To date, there has been little research on the effect of offload delays on the transported patient or those awaiting care in the community.

In some communities, when EDs are at full capacity and unable to safely accommodate more high-acuity patients, they may go on diversion, meaning that ambulances are rerouted away from the closest ED to an alternative ED. ED crowding is a common cause of ambulance diversion, and ambulance diversion is a commonly used surrogate marker for ED crowding (Pham et al. 2006). Studies of the frequency of ambulance diversion indicate that they are a common occurrence. In 2003, an estimated 501,000 ambulances were diverted in the USA, equivalent to about one every minute (Burt et al. 2006). A point-in-time study conducted at 7 PM on Monday, March 21, 2011, revealed that 11 % of EDs in the USA were on diversion. However, there is very little current data available on whether diversion hours have increased or decreased over time.

Ambulance diversion has been considered as a safety measure, as it protects incoming patients from long waits upon arrival at a crowded ED and assures that current ED patients do not have to "compete" for care with incoming patients. However, as one would expect, transporting patients to a more distant hospital results in longer transport times, though estimates show the increase in transport time to be 5 min or less (Pham et al. 2006). To date, there is no information available on the impact of ambulance diversion on the patients who are already present in the ED.

3 Compromised Care Quality

The Centers for Medicare and Medicaid Services, National Quality Forum, and Joint Commission have identified several time-sensitive process of care measures that have become accepted indicators of care quality. Examples include:

- Heart attack patients given fibrinolytic medication within 30 min of arrival
- Heart attack patients given percutaneous coronary interventions (PCI) within 90 min of arrival

Heart attacks often occur when blood clots cause blockages in blood vessels, depriving the heart of sufficient oxygen. Fibrinolytic drugs help dissolve blood clots and improve blood flow to the heart. PCI procedures open blocked vessels and help prevent further damage to the heart. They can also increase a patient's chance of surviving a heart attack. The earlier the PCI is performed, the more effective it is in improving patient outcomes. Both measures are publicly reported on the CMS Hospital Compare website, allowing consumers an opportunity to compare the percentage of time their local hospital meets these guidelines with state and national averages. Data from the website show that hospitals meet the fibrinolytic medication measure approximately 60 % of the time; they meet the PCI measure approximately 95 % of the time.

While it is clear that delays in life-saving treatment for heart attacks and other conditions do occur, what are the consequences for patients? To date, research has generally examined the impact of crowding and delays on path management and the incidence of medical errors and adverse events.

3.1 Pain Management

Approximately 32 % of ED patients present with moderate or severe pain, and crowded conditions in the ED often means that patients spend more time in pain than they otherwise would. Studies of various patient conditions consistently show that pain management is compromised when the ED is crowded. Operating above ED capacity was associated with longer times to pain management, undertreatment of pain management, and administration of inappropriate pain medications for older adults with hip fracture (Hwang et al. 2006). Two studies of general, adult ED patients with conditions warranting pain care found that there was a direct correlation between measures of ED crowding and receipt of pain care, time to pain assessment, time to analgesic medication ordering, and time to analgesic medication administration (Hwang et al. 2008; Pines and Hollander 2008). The findings hold true for adults with abdominal pain (Mills et al. 2009) and even pediatric patients. When a children's hospital was operating at the highest levels of crowding, children with long-bone fractures were less likely to receive timely pain medication and less likely to receive effective pain medication (Sills et al. 2011a, b). One reason why pain management may be compromised is that providers are simply too busy to appropriately assess and treat patients with painful conditions (Pines and Hollander 2008). Indeed, pain assessments are often not documented when EDs are operating above full capacity (Hwang et al. 2006).

3.2 Preventable Medical Errors and Adverse Events

EDs are high-risk, high-stress environments. Providers treat a broad case mix of patients without the benefit of a medical history and often deal with frequent interruptions (IOM 2006). Perhaps it is not surprising that the ED is the source of a considerable number of medical errors that result in adverse events. Although the causes of medical errors are multifactorial, several studies indicate that the odds of a patient experiencing a medical error or adverse event increase when EDs are crowded.

3.2.1 ED Occupancy Rates and Preventable Medical Errors and/or Adverse Events

One study of patients presenting conditions of heart attack, asthma exacerbation, or dislocation requiring procedural sedation at four Massachusetts EDs revealed that preventable medical errors were more than twice as likely to occur when the EDs were crowed (measured by occupancy rates and a work index) (Epstein et al. 2012). Examples of preventable medical errors included the following:

- Initial electrocardiogram (ECG) showed ST-elevation myocardial infarction (or STEMI, a serious type of heart attack during which one of the major arteries is blocked), but no mention in physician notes (cardiac treatment delayed several hours).
- Hyperglycemia prompted multiple insulin doses and initiation of insulin drip (patient developed hypoglycemia, required IV D50 solution).
- Patient given both beta-blocker and calcium channel blocker (resulted in severe hypotension, required IV fluids).

The association between crowding and preventable medical errors was nonlinear, with most errors occurring at the highest levels of crowding. Another study of asthma patients at a children's hospital found that children were 9–14 % less likely to receive an asthma assessment score or the appropriate medication when the ED was crowed (measured by occupancy and number waiting to see a physician) (Sills et al. 2011a, b). A third study examining the consequences of high ED census on emergency psychiatric patients at one facility found that patient aggression and agitation increased when the ED census increased, resulting in more frequent use of safety interventions, including seclusion, restraint, and medication (El-Mallakh et al. 2012). Although this study did not focus on preventable medical errors or adverse events, seclusion, restraints, and medication have short-term and long-term detrimental implications for the patient and the physician–patient relationship, and regulatory agencies and advocacy groups are pushing for a reduction in their use (Knox and Holloman 2012).

A fourth study produced mixed results. Non-acute coronary syndrome chest pain patients experienced more adverse outcomes (e.g., death, delayed myocardial infarction, development of congestive heart failure) when waiting room census was high and the total patient care hours (sum of the hours for all patients presently in the ED) were high, but not when the ED was at high occupancy or had a large number of admitted patients (Pines et al. 2009a, b).

3.2.2 Patient Boarding and Preventable Medical Errors and/or Adverse Events

Interestingly, evidence on the consequences of boarding on prevalence of adverse events is mixed. When patients spend extended periods of time in the ED awaiting an inpatient bed to become available, they are said to be "boarding." Boarding is a frequently used indicator of delay and crowding. In one study, researchers concluded that 28 % of boarders had an undesirable event, such as a missed relevant home medication or missed ED treatment, and 3 % had a preventable adverse event, for example, suboptimal blood pressure control or arrhythmia (Liu et al. 2009). However, this study did not include a comparison group. Another study of patients with chest pain, pneumonia, and cellulitis admitted from the ED at two urban hospitals revealed that boarding was not associated with increased risk of medication errors or adverse events (Liu et al. 2011). However, another study found that non-STEMI patients spending more than 8 h in the ED were less likely to receive guideline-recommended therapies and more likely to have recurrent myocardial infarction (Diercks et al. 2007).

3.2.3 Other Measures of Crowding and Preventable Errors and/or Adverse Events

Several retrospective studies from single institutions show numerous adverse events associated with delayed care for ED patients:

- A study of older adults in one Canadian hospital found that the longer the older patients spent in the ED, the more likely they were to experience an adverse event. Indeed, after adjusting for patient factors, the odds of experiencing an adverse event increased by 3 % for every hour the patient spent in the ED (Ackroyd-Stolarz et al. 2011).
- A study of all ED patients in a community hospital found that ED crowding (measured through a work index score) was associated with giving medications at incorrect doses, frequencies, durations, or routes and giving contraindicated medications (Kulstad et al. 2010).
- ED patients with longer times to surgery consult for a small bowel obstruction had a greater chance of surgical resection, which puts patients at greater risk for postoperative morbidity and mortality (Hwang et al. 2011a, b).
- In one Taiwanese hospital, ED crowding was associated with greater risk of blood culture contamination, leading to unnecessary administration of antimicrobial agents and studies and sometimes unnecessary hospitalizations (Lee et al. 2012).

4 Increased Risk of Mortality

The most commonly investigated outcome of ED delays and crowding is mortality. These studies have considered various measures of crowding and delays on mortality across different patient populations. Unfortunately, the results of the studies have been inconsistent, making it difficult to interpret the true impact of crowding and delays on patient mortality.

4.1 Diversion and Mortality

Several studies investigated the relationship between ambulance diversion and mortality. A number of early studies investigating the impact of ambulance diversion on patient outcomes among diverted patients found that ambulance diversion is not associated with increased risk of mortality. These studies included trauma patients and general emergency medical service patients (Begley et al. 2004). The lack of a tie between ambulance diversion and mortality was attributed to regulations that prevented critically ill patients from being diverted (Pham et al. 2006).

More recent studies have investigated the link between hospital diversion hours and outcomes for admitted patients (not necessarily transported by ambulance). These newer studies used large, regional data sets and have found a direct link between ambulance diversion hours and mortality. For example, an investigation of diversion hours across 187 non-federal acute care hospitals in California revealed that patients admitted to the hospital on days within the top quartile for ambulance diversion hours experienced 5 % greater odds of inpatient death, controlling for patient diagnosis and comorbidities. Two studies looked at ambulance diversion hours and mortality among heart attack patients. One study using data from California found that heart attack (acute myocardial infarction) death rates are 3 % higher if the closest ED is on 12 or more hours of diversion on the day of the heart attack. The authors speculated that treatment delays may have contributed to the greater risk of death, but patients in the study were typically accepted by another ED within a mile. Another possible explanation is that diverted patients were less likely to end up at EDs with readily available and potentially lifesaving catheterization labs (Shen and Hsia 2011). Similarly, a study of New York City hospitals found that at least 58 deaths per year were attributed to diversion in the city’s five boroughs (Yankovic et al. 2010).

In sum, there is currently little evidence indicating that ambulance diversion leads to greater mortality for those patients diverted. However, if a hospital is on diversion for an extended period of time, patients admitted with time-sensitive conditions on that day may have a greater risk of mortality.

4.2 Boarding and Mortality

There is remarkable consistency among findings from studies that have investigated the relationship between patient boarding and mortality. They conclude that patient boarding is associated with mortality. A study of critically ill patients at one Greek hospital found that patients who spent more than 6 h in the ED after the decision to admit was made had a 5.7 times greater risk of dying in the hospital than patients who experienced shorter delays (Intas et al. 2012). A similar study of critical care patients from a US hospital also found that the potential for hospital mortality increased the longer the patient was boarded in the ED (Clark and Normile 2007). Finally, in a study of a consortium of 120 hospital intensive care units, boarding in the ED for more than 6 h was associated with a 17.4 % in-hospital mortality rate, compared to 12.9 % for those who boarded less than 6 h (Chalfin et al. 2007).

Two additional studies investigated the relationship between boarding time and mortality among all patients admitted to a single ED. One study conducted at a US hospital found that mortality increased the longer a patient was boarded. Mortality was 2.5 % in patients who boarded less than 2 h and 4.5 % for patients who boarded 12 h or more (Singer et al. 2011). Similarly, a study of all admitted patients at one Irish hospital revealed that 30-day mortality rose the longer a patient was boarded. For example, patients who boarded 1 h (10th percentile for boarding time) had an 8.7 % chance of mortality, compared to 14.8 % among patients who boarded 14 h (75th percentile for boarding time) (Plunkett et al. 2011).

4.3 ED Occupancy and Mortality

Another common measure of ED crowding is occupancy. Four studies indicate that ED occupancy is related to mortality, across various populations of patients. Two of these studies were conducted in Korea. First, 28-day mortality rates were higher for community-acquired pneumonia patients during the highest tertile for occupancy (109 % or greater) (Jo et al. 2012). Also, a study of 34 EDs in Korea with long ED lengths of stay (an average of 6 h or more) found that 30-day mortality rates were highest for pediatric patients when ED volume reached its highest quartile (Cha et al. 2011).

The other two studies were conducted in Australia. The first investigated all patients presenting to a tertiary hospital. Presentation during the highest quartile for occupancy was associated with increased in-hospital mortality at 10 days. Researchers estimated that high occupancy rates accounted for approximately 13 deaths per year at that facility (Richardson 2006). A second study, which included data from three hospitals, found a direct relationship between scores on an overcrowding hazard scale (based on hospital and ED occupancy, after adjusting for age, diagnosis, urgency, etc.) and deaths on days 2, 7, and 30 (Sprivulis et al. 2006).

4.4 Length of Stay and Mortality

While boarding time and ED occupancy appear to have a significant influence on patient mortality, all but one study found no statistical relationship between the total time a patient spent in the ED (ED length of stay) and patient mortality. These studies investigated ED length of stay and mortality across a variety of patients, including trauma patients (Di Bartolomeo et al. 2007; Servia et al. 2012), intensive care unit patients (Carter et al. 2010; Saukkonen et al. 2006), critical care patients (Clark and Normile 2007), heart attack patients (Diercks et al. 2007), and all admitted patients (Flabouris et al. 2013).

The exception was a study of trauma patients admitted to the trauma service at one hospital. Researchers found that hospital mortality increased for each additional hour a patient spent in the ED, with 8.3 % of patients staying in the ED between 4 and 5 h ultimately dying (Mowery et al. 2011). Notably, this study excluded patients who spent more than 5 h in the ED because of significantly lower acuity and mortality. This exclusion criterion may explain why the study findings are different than other research efforts investigating the same issue.

4.5 Delayed Care and Mortality

Finally, several researchers have investigated the impact of delays in care on patient mortality. It is difficult to draw general conclusions about these studies as a whole because they focused on different diagnoses and their results are mixed. Four found that treatment delays were related to mortality (or that prompt treatment was associated to survival); two did not.

- Traumatic brain injury patients in Korea who underwent craniotomy or drainage of hematoma within 4 h of arrival were twice as likely to survive than those who waited more than 4 h (Kim 2011).
- Delivery of fibrinolytic therapy within 30 min for ST-segment elevation myocardial infarction patients was associated with lower in-hospital mortality (2.9 %) than for patients who waited for 31–45 min (4.1 %) or more than 45 min (6.2 %) (McNamara et al. 2007).
- Chest pain patients with delayed ECG acquisition had a threefold increase in the risk for death (Diercks et al. 2006).
- Shorter time to antibiotic administration for patients with severe sepsis or septic shock is associated with reduced mortality but only if appropriate antibiotics are administered (Gaieski et al. 2010).
- There was no association between time to first antibiotic dose and mortality for patients admitted with pneumonia (Quattromani et al. 2011).
- There was no difference in mortality when trauma team activation or trauma team consult occurred more than 30 min after patient arrival at the ED and when occurred less than 30 min after patient arrival (Ryb et al. 2012).

5 Equity in Delays

The literature review revealed an interesting finding concerning the equity of emergency care—not all patients are equally likely to experience delays. In fact, ED length of stay is longer for minority patients than for white patients (Hwang et al. 2011a, b). These differences have persisted over time, occur in both adult and pediatric populations, and occur across a number of conditions

(Herring et al. 2009; James et al. 2005; Wu et al. 2009). To some degree, this reflects delays at high-minority serving facilities (Pines et al. 2009a, b). Nationally, the number of existing EDs has declined, and safety net hospitals (those that deliver a significant amount of care to the uninsured, Medicaid, and other vulnerable populations) are at the greatest risk of closing, disproportionately disenfranchising minorities, immigrants, and the poor (Shen et al. 2009). The shrinking supply of EDs leads to crowding at other facilities (Hsia et al. 2011a, b). The left-without-being-seen rate is higher at hospitals that serve low-income patients (Hsia et al. 2011a, b). Nationally, admitted African American patients wait longer for inpatient beds than admitted white patients, which may be a result of their accessing more crowded hospitals (Pines et al. 2009a, b).

However, studies also show disparities in wait times within hospitals by race and ethnicity. For example, Hispanic children wait 10 % longer than non-Hispanic white children when treated at the same hospital, even after adjusting for triage status (i.e., immediacy with which patients should be seen) and payment source (Park et al. 2009). Also, blacks have longer wait times than non-blacks within the same facility (Pines et al. 2009a, b). In one large sample representative of the US EDs, blacks and Hispanics had a 10 % lower chance of being triaged within the appropriate 15-min window compared to whites within the same hospital (Nottidge et al. 2009).

6 Strategies to Address ED Delays and Crowding

Over the past 10 years, there have been a tremendous number of efforts to improve the quality of emergency care by addressing the problem of ED crowding. Examples include a Robert Wood Johnson Foundation-funded effort, Urgent Matters, which finds, develops, and disseminates strategies aimed at improving patient flow and reducing ED crowding. Urgent Matters led a learning collaborative during which they provided technical assistance to hospitals to address delays and crowding. Similarly, the Institute for Healthcare Improvement, Agency for Healthcare Research and Quality, American Hospital Association, and American College of Emergency Physicians have all made resources available to hospital leaders and/or held seminars and webinars to provide guidance on reducing ED crowding. For example, hospital and ED managers can go online to learn about specific interventions through AHRQ's Innovation Exchange (http://www.innovations.ahrq.gov/) and Urgent Matters' Toolkit (http://urgentmatters.org/toolkit/985888?pg=all).

The challenge with these multiple initiatives is that there has been a proliferation of resources for hospitals and long lists of potential interventions, yet little information is available that compares the effectiveness of the interventions. For example, in 2008, Hoot and Aronsky published a systematic review of solutions for ED crowding. They grouped the interventions into three categories (increased

Table 5.1 Effectiveness of solutions to address ED crowding, according to the American College of Emergency Physicians

Solution	Description
High-impact solutions	
Moving admitted ED patients to inpatient areas, such as hallways and conference rooms	Often referred to as the Full Capacity Protocol, this approach eliminates boarding in the ED and instead has each hospital unit caring for a small number of additional patients awaiting a bed, thereby spreading the burden of admitted patients across the hospital, freeing the ED of the sole burden
Coordinate the discharge of hospital patients before noon	Timely discharge can improve the flow of patients in the ED by reducing boarding. However, discharging patients by noon requires changes in culture and process in the inpatient units
Coordinate the scheduling of elective and surgical patients	Elective surgical cases are typically heaviest earlier in the week, making it difficult to admit patients from the ED. Spreading surgical cases throughout the week increases the likelihood that an inpatient bed will be available for admitted ED patients
Additional solutions	
Bedside registration	Instead of registering patients near the waiting room, patients are immediately placed in a bed and registrars go to the patient
Fast track units	Non-urgent patients are triaged to a separate area for care, which gives staff the ability to quickly handle low-acuity patients
Observation units	Patients are directed to a specialized outpatient unit where they can be observed for 8–24 h before an admission decision is made
Physician triage	Placing a physician in triage allows the patient immediate access to a diagnosis, and, if appropriate, the physician can quickly treat and discharge the patient
Cancelling elective surgeries	When admitted ED patients are awaiting a hospital bed, cancelling elective surgeries would help to free beds

Source: American College of Emergency Physicians. Emergency department crowding: high-impact solutions. April (2008)

resources, demand management, and operations research) but did not attempt to identify which categories of strategies were most effective.

In 2008, the American College of Emergency Physicians (ACEP) released a report that attempted to rank interventions based on their effectiveness (American College of Emergency Physicians 2008). Results are summarized in Table 5.1. High-impact solutions were considered to be those that would have “significant” impact on reducing boarding and improving patient flow. Additional solutions are those that may be effective in reducing delays and crowding, but the implementation costs may be prohibitive. Although hospital and ED managers may find such a ranking system to be useful, ACEP did not provide details on how the rankings were developed, and the conclusions are not always consistent with existing literature (Hoot and Aronsky 2008).

More recently, a consensus conference sponsored by Academic Emergency Medicine was held to review interventions that have been implemented to reduce crowding and summarize the evidence of their effectiveness on the delivery of emergency care and identify strategies that may help reduce crowding or improve the quality of care provided during episodes of crowding (Pines and McCarthy 2011). Participants concluded that despite the large number of operational interventions that have been implemented to reduce crowding and delays, there have been very few rigorous evaluations, so their value remains unclear. Because it is unclear which interventions are most effective, it is challenging for hospital and ED managers and staff to identify where to invest their time and energy.

One of the recurring themes of the consensus conference was the importance of incorporating engineering and operations research perspectives into efforts to address crowding and delays. For example, hospitals should design systems and work processes that better match supply and demand, potentially through the use of simulation (Soremekun et al. 2011). While operations management, the science of business operations, has long been used to improve service in other industries, for example restaurants, hotels, and airlines, the adoption of operations management in health care has been a more recent trend.

Adoption of strategies to reduce delays and crowding has been variable. A team of researchers surveyed academic EDs about the adoption of ACEP-identified interventions (Liu et al. 2013). Their results showed that the most common interventions were fast track (79 % of EDs), bedside registration (55 %), and observation units (53 %). Strategies that were least likely to be adopted were cancelling elective surgeries (14 %), physician triage (12 %), and coordinating the elective surgical case schedule (11 %).

It is worth noting that recent discussions about solutions to ED crowding largely center on strategies that hospitals can undertake to address the problem. It reflects current research suggesting that ED crowding occurs primarily when sick ED patients are admitted to the hospital but cannot be placed in an inpatient bed due to high hospital occupancy rates (Forster et al 2003; Rathlev et al. 2007; Moskop et al. 2009). In the 1980s and early 1990s, it was widely assumed that “inappropriate” use of the ED (e.g., prescription refills, ankle sprains, headaches) was largely driving crowding; however, research has shown that the number of patients with minor illnesses and injuries have a negligible effect on wait times for more acutely ill ED patients (Moskop et al. 2009; Schull et al. 2007).

Notably, high demand for ED services is also widely cited as a cause of ED crowding. Indeed the number of patients seeking care from EDs rose 23 % between 1997 and 2007 (Niska et al. 2010). Reasons for the rise in demand are not well understood (Boyle et al. 2012), but the rise is often attributed to an inability to access care elsewhere (Hoot and Aronsky 2008). However, the rise in ED visits has corresponded with a rise in primary care activity. Further, countries with robust primary care and after-hour options also experience major ED crowding (Pines et al. 2011). Still, developing desirable alternatives to ED services, for example, collocating primary care services within or adjacent to EDs, may help to alleviate crowding, but the evidence is weak (Boyle et al 2012).

7 Discussion

Ten years ago, reflecting on the issue of crowding in EDs, leaders of the emergency medicine community wrote, “a large gap remains in our understanding of whether crowding adversely affects quality, and if so, what are the nature and frequency of these quality problems” (Magid et al. 2004). Since that time, the number of published articles linking delays and crowding to outcomes has grown tremendously. Today we can say with greater certainty that ED crowding leads to delays and that ED crowding and delays compromise care quality and put patients at greater risk of an adverse event, including death. Crowding and delays in the ED compromise care across all six domains of quality identified in the landmark IOM report, Crossing the Quality Chasm (Table 5.2) (IOM 2001).

Although the conclusion that crowding and delays adversely impact patients is largely uncontroversial, the quality of the evidence linking delays in care to patient outcomes can and should be improved. The studies cited in this chapter were largely single-institution observational cohort studies, though several reflected pooled data, typically from EDs in a specific geographic area. Researchers should continue to take advantage of national data sources on ED visits including Medicare claims data, the Agency for Healthcare Research and Quality’s Healthcare Cost and Utilization Project databases, and the National Center for Health Statistics’ National Hospital Ambulatory Medical Care Survey.

One limitation of this literature review is the potential for publication bias. Studies showing significant, positive, results are more likely to be published than studies showing nonsignificant or “negative” results. As a result, we may expect to see more studies linking ED crowding and delays to poor patient outcomes. Although this is a concern, our review still found a number (though a minority) of studies with nonsignificant findings.

Still, the number of studies investigating delays in emergency care is likely to increase in the coming years. Because ED crowding and delays have been associated with adverse outcomes, the Centers for Medicare and Medicaid Services (CMS) has shown an increasing interest in tracking crowding and delays (McClelland et al. 2012). CMS added several time-based performance measures to their Hospital Outpatient Quality Reporting Program, which means that hospitals must report data on the measures in order to receive the full annual update to their

Table 5.2 Impact of crowding and delays on the IOM's six dimensions of health care quality

Quality domain	Definition	Examples of the impact of crowding and delays
Safe	Avoiding injuries to patients from the care that is intended to help them	• Increased mortality • More preventable medical errors and adverse events
Effective	Providing services based on scientific knowledge to all who could benefit	• Greater likelihood of patients leaving before being seen • Undertreatment of pain management; delivery of inappropriate pain medications
Patient centered	Providing care that is respectful of and responsive to individual patient preferences, needs, and values and ensuring that patient values guide all clinical decisions	• Lower patient satisfaction • Boarding of admitted patients in the ED
Timely	Reducing waits and sometimes harmful delays for both those who receive and those who give care	• Longer waits to treatment • Longer transport times if EDs are on diversion
Efficient	Avoiding waste, including waste of equipment, supplies, ideas, and energy	• Longer lengths of stay • Delayed transfer of ambulance patients means ambulances are out of service for longer periods
Equitable	Providing care that does not vary in quality because of personal characteristics	• Patients at safety net hospitals are more likely to experience crowding and delays • Even within hospitals, minorities face longer delays than nonminorities

Medicare payment rate. Further, several measures of timely ED care have recently been released on the Hospital Compare website:

- Median time patients spent in the ED before they were admitted to the hospital as an inpatient
- Median time patients spent in the ED after the doctor decided to admit them as an inpatient before leaving the ED for an inpatient room
- Average time patients spent in the ED before being sent home
- Average time patients spent in the ED before they were seen by a healthcare professional
- Average time patients who came to the ED with broken bones had to wait before receiving pain medication
- Percentage of patients who left the ED before being seen
- Percentage of patients who came to the ED with stroke symptoms who received brain scan results within 45 min of arrival

The addition of these measures on the CMS website, allowing consumers to make comparisons across local hospitals, may further motivate hospital leaders to adopt strategies to reduce ED crowding and delays.

References

Ackroyd-Stolarz, S., Read Guernsey, J., Mackinnon, N. J., & Kovacs, G. (2011). The association between a prolonged stay in the emergency department and adverse events in older patients admitted to hospital: A retrospective cohort study. *BMJ Quality and Safety, 20*(7), 564–569.

American College of Emergency Physicians. (2008). *Emergency department crowding: High-impact solutions*. Dallas, TX: ACEP.

Baker, D. W., Stevens, C. D., & Brook, R. H. (1991). Patients who leave a public hospital emergency department without being seen by a physician. Causes and consequences. *Journal of the American Medical Association, 266*(8), 1085–1090.

Begley, C. E., Chang, Y., Wood, R. C., & Weltge, A. (2004). Emergency department diversion and trauma mortality: Evidence from Houston, Texas. *The Journal of Trauma, 57*(6), 1260–1265.

Boyle, A. B., Beniuk, K., Higginson, I., & Atkinson, P. (2012). Emergency department crowding: Time for interventions and policy evaluations. *Emergency Medicine International, 2012*, 838610.

Burt, C., McCaig, L., & Valverde, R. (2006). Analysis of ambulance transports and diversions among US emergency departments. *Annals of Emergency Medicine, 47*(4), 317–326.

Carter, A. W., Pilcher, D., Bailey, M., Cameron, P., Duke, G. J., & Cooper, J. (2010). Is ED length of stay before ICU admission related to patient mortality? *Emergency Medicine Australasia, 22* (2), 145–150.

Cha, W. C., Shin, S. D., Cho, J. S., Song, K. J., Singer, A. J., & Kwak, Y. H. (2011). The association between crowding and mortality in admitted pediatric patients from mixed adult-pediatric emergency departments in Korea. *Pediatric Emergency Care, 27*(12), 1136–1141.

Chalfin, D. B., Trzeciak, S., Likourezos, A., Baumann, B. M., & Dellinger, R. P. (2007). Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine, 35*(6), 1477–1483. doi:10.1097/01.CCM.0000266585.74905.5A.

Clarey, A. J., & Cooke, M. W. (2011). Patients who leave emergency departments without being seen: Literature review and English data analysis. *Emergency Medicine Journal*. doi:10.1136/emermed-2011-200537.

Clark, K., & Normile, L. B. (2007). Influence of time-to-interventions for emergency department critical care patients on hospital mortality. *Journal of Emergency Nursing, 33*(1), 6–13.

Cooney, D. R., Millin, M. G., Carter, A., Lawner, B. J., Nable, J. V., & Wallus, H. J. (2011). Ambulance diversion and emergency department offload delay: Resource document for the National Association of EMS Physicians position statement. *Prehospital Emergency Care, 15* (4), 555–561. doi:10.3109/10903127.2011.608871.

Di Bartolomeo, S., Valent, F., Rosolen, V., Sanson, G., Nardi, G., Cancellieri, F., & Barbone, F. (2007). Are pre-hospital time and emergency department disposition time useful process indicators for trauma care in Italy? *Injury, 38*(3), 305–311.

Diercks, D. B., Kirk, J. D., Lindsell, C. J., Pollack, C. V., Jr., Hoekstra, J. W., Gibler, W. B., & Hollander, J. E. (2006). Door-to-ECG time in patients with chest pain presenting to the ED. *The American Journal of Emergency Medicine, 24*(1), 1–7.

Diercks, D. B., Roe, M. T., Chen, A. Y., Peacock, W. F., Kirk, J. D., Pollack, C. V., Jr., et al. (2007). Prolonged emergency department stays of non-ST-segment-elevation myocardial infarction patients are associated with worse adherence to the American College of Cardiology/American Heart Association guidelines for management and increased adverse events. *Annals of Emergency Medicine, 50*(5), 489–496.

Eckstein, M., & Chan, L. S. (2004). The effect of emergency department crowding on paramedic ambulance availability. *Annals of Emergency Medicine, 43*(1), 100–105.

El-Mallakh, R. S., Whiteley, A., Wozniak, T., Ashby, M., Brown, S., Colbert-Trowel, D., et al. (2012). Waiting room crowding and agitation in a dedicated psychiatric emergency service. *Annals of Clinical Psychiatry, 24*(2), 140–142.

Epstein, S. K., Huckins, D. S., Liu, S. W., Pallin, D. J., Sullivan, A. F., Lipton, R. I., & Camargo, C. A., Jr. (2012). Emergency department crowding and risk of preventable medical errors. *Internal and Emergency Medicine, 7*(2), 173–180.

Fee, C., Weber, E. J., Maak, C. A., & Bacchetti, P. (2007). Effect of emergency department crowding on time to antibiotics in patients admitted with community-acquired pneumonia. *Annals of Emergency Medicine, 50*(5), 501–509, 509.e1.

Fernandes, C. M., Daya, M. R., Barry, S., & Palmer, N. (1994). Emergency department patients who leave without seeing a physician: The Toronto hospital experience. *Annals of Emergency Medicine, 24*(6), 1092–1096. doi: S0196064494002416 [pii].

Fernandes, C. M., Price, A., & Christenson, J. M. (1997). Does reduced length of stay decrease the number of emergency department patients who leave without seeing a physician? *The Journal of Emergency Medicine, 15*(3), 397–399. doi: S0736467997000309 [pii].

Flabouris, A., Jeyadoss, J., Field, J., & Soulsby, T. (2013). Association between emergency department length of stay and outcome of patients admitted either to a ward, intensive care or high dependency unit. *Emergency Medicine Australasia, 25*(1), 46–54.

Forster, A. J., Stiell, G., Wells, A., Lee, C., & van Walraven, C. (2003). The effect of hospital occupancy on emergency department length of stay and patient disposition. *Academic Emergency Medicine, 10*(2), 127–133.

Gaieski, D. F., Mikkelsen, M. E., Band, R. A., Pines, J. M., Massone, R., Furia, F. F., et al. (2010). Impact of time to antibiotics on survival in patients with severe sepsis or septic shock in whom early goal-directed therapy was initiated in the emergency department. *Critical Care Medicine, 38*(4), 1045–1053.

General Accountability Office. (2009). *Hospital emergency departments. Crowding continues to occur, and some patients wait longer than recommended time frames*. Washington, DC: General Accountability Office.

Herring, A., Wilper, A., Himmelstein, D. U., Woolhandler, S., Espinola, J. A., Brown, D. F., & Camargo, C. A., Jr. (2009). Increasing length of stay among adult visits to U.S. emergency departments, 2001–2005. *Academic Emergency Medicine, 16*(7), 609–616.

Hing, E., & Bhuiya, F. (2012). Wait time for treatment in hospital emergency departments: 2009. In NCH Statistics (Ed.), *NCHS data brief*. Atlanta: US Department of Health and Human Services.

Hoot, N. R., & Aronsky, D. (2008). Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine, 52*(2), 126–136.

Horwitz, L. I., & Bradley, E. H. (2009). Percentage of US emergency department patients seen within the recommended triage time: 1997 to 2006. *Archives of Internal Medicine, 169*(20), 1857–1865. doi:10.1001/archinternmed.2009.336.

Horwitz, L. I., Green, J., & Bradley, E. H. (2010). US emergency department performance on wait time and length of visit. *Annals of Emergency Medicine, 55*(2), 133–141.

Hsia, R. Y., Asch, S. M., Weiss, R. E., Zingmond, D., Liang, L. J., Han, W., et al. (2011a). Hospital determinants of emergency department left without being seen rates. *Annals of Emergency Medicine, 58*(1), 24.e3–32.e3. doi:10.1016/j.annemergmed.2011.01.009.

Hsia, R. Y., Kellermann, A. L., & Shen, Y. C. (2011b). Factors associated with closures of emergency departments in the United States. *Journal of the American Medical Association, 305*(19), 1978–1985. doi:10.1001/jama.2011.620.

Hwang, U., Aufses, A. H., Jr., & Bickell, N. A. (2011a). Factors associated with delays to emergency care for bowel obstruction. *The American Journal of Surgery, 202*(1), 1–7.

Hwang, U., Richardson, L., Livote, E., Harris, B., Spencer, N., & Sean Morrison, R. (2008). Emergency department crowding and decreased quality of pain care. *Academic Emergency Medicine, 15*(12), 1248–1255.

Hwang, U., Richardson, L. D., Sonuyi, T. O., & Morrison, R. S. (2006). The effect of emergency department crowding on the management of pain in older adults with hip fracture. *Journal of the American Geriatrics Society, 54*(2), 270–275. doi:10.1111/j.1532-5415.2005.00587.x.

Hwang, U., Weber, E. J., Richardson, L. D., Sweet, V., Todd, K., Abraham, G., & Ankel, F. (2011b). A research agenda to assure equity during periods of emergency department crowding. *Academic Emergency Medicine, 18*(12), 1318–1323. doi:10.1111/j.1553-2712.2011.01233.x.

Intas, G., Stergiannis, P., Chalari, E., Tsoumakas, K., & Fildissis, G. (2012). The impact of ED boarding time, severity of illness, and discharge destination on outcomes of critically ill ED patients. *Advanced Emergency Nursing Journal, 34*(2), 164–169.

IOM. (2001). *Crossing the quality chasm*. Washington, DC: National Academy Press.

IOM. (2006). *Hospital-based emergency care: At the breaking point*. Washington, DC: National Academy Press.

James, C. A., Bourgeois, F. T., & Shannon, M. W. (2005). Association of race/ethnicity with emergency department wait times. *Pediatrics, 115*(3), e310–e315.

Jo, S., Kim, K., Lee, J. H., Rhee, J. E., Kim, Y. J., Suh, G. J., & Jin, Y. H. (2012). Emergency department crowding is associated with 28-day mortality in community-acquired pneumonia patients. *The Journal of Infection, 64*(3), 268–275.

Kennebeck, S. S., Timm, N. L., Kurowski, E. M., Byczkowski, T. L., & Reeves, S. D. (2011). The association of emergency department crowding and time to antibiotics in febrile neonates. *Academic Emergency Medicine, 18*(12), 1380–1385.

Kim, Y. J. (2011). The impact of time from ED arrival to surgery on mortality and hospital length of stay in patients with traumatic brain injury. *Journal of Emergency Nursing, 37*(4), 328–333.

Knox, D. K., & Holloman, G. H., Jr. (2012). Use and avoidance of seclusion and restraint: Consensus statement of the American association for emergency psychiatry project Beta seclusion and restraint workgroup. *Western Journal of Emergency Medicine, 13*(1), 35–40. doi:10.5811/westjem.2011.9.6867.

Kulstad, E. B., Sikka, R., Sweis, R. T., Kelley, K. M., & Rzechula, K. H. (2010). ED overcrowding is associated with an increased frequency of medication errors. *The American Journal of Emergency Medicine, 28*(3), 304–309.

Lee, C. C., Lee, N. Y., Chuang, M. C., Chen, P. L., Chang, C. M., & Ko, W. C. (2012). The impact of overcrowding on the bacterial contamination of blood cultures in the ED. *The American Journal of Emergency Medicine, 30*(6), 839–845.

Liu, S. W., Chang, Y., Weissman, J. S., Griffey, R. T., Thomas, J., Nergui, S., et al. (2011). An empirical assessment of boarding and quality of care: Delays in care among chest pain, pneumonia, and cellulitis patients. *Academic Emergency Medicine, 18*(12), 1339–1348.

Liu, S. W., Hamedani, A. G., Brown, D. F., Asplin, B., & Camargo, C. A., Jr. (2013). Established and novel initiatives to reduce crowding in emergency departments. *Western Journal of Emergency Medicine, 14*(2), 85–89. doi:10.5811/westjem.2012.11.12171.

Liu, S. W., Thomas, S. H., Gordon, J. A., Hamedani, A. G., & Weissman, J. S. (2009). A pilot study examining undesirable events among emergency department-boarded patients awaiting inpatient beds. *Annals of Emergency Medicine, 54*(3), 381–385.

Magid, D. J., Asplin, B. R., & Wears, R. L. (2004). The quality gap: Searching for the consequences of emergency department crowding. *Annals of Emergency Medicine, 44*(6), 586–588.

McClelland, M. S., Jones, K., Siegel, B., & Pines, J. M. (2012). A field test of time-based emergency department quality measures. *Annals of Emergency Medicine, 59*(1), 1.e2–10.e2.

McNamara, R. L., Herrin, J., Wang, Y., Curtis, J. P., Bradley, E. H., Magid, D. J., et al. (2007). Impact of delay in door-to-needle time on mortality in patients with ST-segment elevation myocardial infarction. *The American Journal of Cardiology, 100*(8), 1227–1232.

Mills, A. M., Shofer, F. S., Chen, E. H., Hollander, J. E., & Pines, J. M. (2009). The association between emergency department crowding and analgesia administration in acute abdominal pain patients. *Academic Emergency Medicine, 16*(7), 603–608.

Mohsin, M., Forero, R., Ieraci, S., Bauman, A. E., Young, L., & Santiano, N. (2007). A population follow-up study of patients who left an emergency department without being seen by a medical officer. *Emergency Medicine Journal, 24*(3), 175–179. doi: 24/3/175 [pii].

Monzon, J., Friedman, S. M., Clarke, C., & Arenovich, T. (2005). Patients who leave the emergency department without being seen by a physician: A control-matched study. *Canadian Journal of Emergency Medical Care, 7*(2), 107–113. doi: 5F7AC634DAFE4F7BBB08F817B0EA33B4 [pii].

Moskop, J. C., Sklar, D. P., Geiderman, J. M., Schears, R. M., & Bookman, K. J. (2009). Emergency department crowding, part 1 – Concept, causes, and moral consequences. *Annals of Emergency Medicine, 53*(5), 605–611.

Mowery, N. T., Dougherty, S. D., Hildreth, A. N., Holmes, J. H., 4th, Chang, M. C., Martin, R. S., Hoth, J. J., Meredith, J. W., & Miller, P. R. (2011). Emergency department length of stay is an independent predictor of hospital mortality in trauma activation patients. *The Journal of Trauma, 70*(6), 1317–1325.

Niska, R. W., Bhuiya, F., & Xu, J. (2010). *National hospital ambulatory medical care survey: 2007 emergency department summary. National health statistics reports, no. 7*. Hyattsville, MD: National Center for Health Statistics.

Nottidge, M. E., Ding, R., Zeger, S. L., Kelen, G. D., Steinwachs, D. M., & McCarthy, M. L. (2009). Racial and ethnic disparities in emergency department triage (Abstract). *Academic Emergency Medicine, 16*(4), S273.

Office, G. A. (2003). *Hospital emergency departments: Crowded conditions vary among hospitals and communities*. Washington, DC: GAO.

One-third of EDs may fail to meet 90-minute target for heart attack patients. (2007). *ED Management*, *19*(1): 1–3.

Park, C. Y., Lee, M. A., & Epstein, A. J. (2009). Variation in emergency department wait times for children by race/ethnicity and payment source. *Health Services Research, 44*(6), 2022–2039.

Pham, J. C., Ho, G. K., Hill, P. M., McCarthy, M. L., & Pronovost, P. J. (2009). National study of patient, visit, and hospital characteristics associated with leaving an emergency department without being seen: Predicting LWBS. *Academic Emergency Medicine, 16*(10), 949–955. doi:10.1111/j.1553-2712.2009.00515.x.

Pham, J. C., Patel, R., Millin, M. G., Kirsch, T. D., & Chanmugam, A. (2006). The effects of ambulance diversion: A comprehensive review. *Academic Emergency Medicine, 13*(11), 1220–1227.

Pines, J. M., Hilton, J. A., Weber, E. J., et al. (2011). International perspectives on emergency department crowding. *Academic Emergency Medicine, 18*(12), 1358–1370.

Pines, J., & Hollander, J. (2008). Emergency department crowding is associated with poor care for patients with severe pain. *Annals of Emergency Medicine, 51*, 1–5.

Pines, J. M., & McCarthy, M. L. (2011). Executive summary: Interventions to improve quality in the crowded emergency department. *Academic Emergency Medicine, 18*(12), 1229–1233. doi:10.1111/j.1553-2712.2011.01228.x.

Pines, J. M., Pollack, C. V., Jr., Diercks, D. B., Chang, A. M., Shofer, F. S., & Hollander, J. E. (2009a). The association between emergency department crowding and adverse cardiovascular outcomes in patients with chest pain. *Academic Emergency Medicine, 16*(7), 617–625.

Pines, J. M., Russell Localio, A., & Hollander, J. E. (2009b). Racial disparities in emergency department length of stay for admitted patients in the United States. *Academic Emergency Medicine, 16*(5), 403–410.

Plunkett, P. K., Byrne, D. G., Breslin, T., Bennett, K., & Silke, B. (2011). Increasing wait times predict increasing mortality for emergency medical admissions. *European Journal of Emergency Medicine, 18*(4), 192–196.

Quattromani, E., Powell, E. S., Khare, R. K., Cheema, N., Sauser, K., Periyanayagam, U., Pirotte, M. J., Feinglass, J., & Courtney, M. D. (2011). Hospital-reported data on the pneumonia quality measure "Time to First Antibiotic Dose" are not associated with inpatient mortality: Results of a nationwide cross-sectional analysis. *Academic Emergency Medicine, 18*(5), 496–503.

Rathlev, N. K., Chessare, J., Olshaker, D., et al. (2007). Time series analysis of variables associated with daily mean emergency department length of stay. *Annals of Emergency Medicine, 49*(3), 265–271.

Richardson, D. B. (2006). Increase in patient mortality at 10 days associated with emergency department overcrowding. *The Medical Journal of Australia, 184*(5), 213–216.

Rowe, B. H., Channan, P., Bullard, M., Blitz, S., Saunders, L. D., Rosychuk, R. J., Lari, H., Craig, W. R., & Holroyd, B. R. (2006). Characteristics of patients who leave emergency departments without being seen. *Academic Emergency Medicine, 13*(8), 848–852. doi:10.1197/j.aem.2006.01.028.

Ryb, G. E., Cooper, C., & Waak, S. M. (2012). Delayed trauma team activation: Patient characteristics and outcomes. *The Journal of Trauma and Acute Care Surgery, 73*(3), 695–698.

Saukkonen, K. A., Varpula, M., Rasanen, P., Roine, R. P., Voipio-Pulkki, L. M., & Pettila, V. (2006). The effect of emergency department delay on outcome in critically ill medical patients: Evaluation using hospital mortality and quality of life at 6 months. *Journal of Internal Medicine, 260*(6), 586–591.

Schull, M. J., Kiss, A., & Szalai, J. P. (2007). The effect of low-complexity patients on emergency department waiting times. *Annals of Emergency Medicine, 49*(3), 257–264.

Schull, M. J., Vermeulen, M., Slaughter, G., Morrison, L., & Daly, P. (2004). Emergency department crowding and thrombolysis delays in acute myocardial infarction. *Annals of Emergency Medicine, 44*(6), 577–585.

Servia, L., Badia, M., Baeza, I., Montserrat, N., Justes, M., Cabre, X., Valdres, P., & Trujillano, J. (2012). Time spent in the emergency department and mortality rates in severely injured patients admitted to the intensive care unit: An observational study. *Journal of Critical Care, 27*(1), 58–65.

Shaikh, S. B., Jerrard, D. A., Witting, M. D., Winters, M. E., & Brodeur, M. N. (2012). How long are patients willing to wait in the emergency department before leaving without being seen? *Western Journal of Emergency Medicine, 13*(6), 463–467. doi: 10.5811/westjem.2012.3.6895, wjem-13-463 [pii].

Shen, Y. C., & Hsia, R. Y. (2011). Association between ambulance diversion and survival among patients with acute myocardial infarction. *Journal of the American Medical Association, 305*(23), 2440–2447.

Shen, Y. C., Hsia, R. Y., & Kuzma, K. (2009). Understanding the risk factors of trauma center closures: Do financial pressure and community characteristics matter? *Medical Care, 47*(9), 968–978. doi: 10.1097/MLR.0b013e31819c941500005650-200909000-00006 [pii].

Shenoi, R., Ma, L., Syblik, D., & Yusuf, S. (2011). Emergency department crowding and analgesic delay in pediatric sickle cell pain crises. *Pediatric Emergency Care, 27*(10), 911–917.

Sills, M. R., Fairclough, D., Ranade, D., & Kahn, M. G. (2011a). Emergency department crowding is associated with decreased quality of care for children with acute asthma. *Annals of Emergency Medicine, 57*(3), 191–200.

Sills, M. R., Fairclough, D. L., Ranade, D., Mitchell, M. S., & Kahn, M. G. (2011b). Emergency department crowding is associated with decreased quality of analgesia delivery for children with pain related to acute, isolated, long-bone fractures. *Academic Emergency Medicine, 18*(12), 1330–1338.

Singer, A. J., Thode, H. C., Jr., Viccellio, P., & Pines, J. M. (2011). The association between length of emergency department boarding and mortality. *Academic Emergency Medicine, 18*(12), 1324–1329.

Sklar, D. P., Crandall, C. S., Zola, T., & Cunningham, R. (2010). Emergency physician perceptions of patient safety risks. *Annals of Emergency Medicine, 55*(4), 336–340.

Soremekun, O. A., Terwiesch, C., & Pines, J. M. (2011). Emergency medicine: An operations management view. *Academic Emergency Medicine, 18*(12), 1262–1268. doi:10.1111/j.1553-2712.2011.01226.x.

Sprivulis, P. C., Da Silva, J. A., Jacobs, I. G., Frazer, A. R., & Jelinek, G. A. (2006). The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments. *The Medical Journal of Australia, 184*(5), 208–212.

Weiss, S. J., Ernst, A. A., Derlet, R., King, R., Bair, A., & Nick, T. G. (2005). Relationship between the National ED Overcrowding Scale and the number of patients who leave without being seen in an academic ED. *The American Journal of Emergency Medicine, 23*(3), 288–294.

Wilper, A. P., Woolhandler, S., Lasser, K. E., McCormick, D., Cutrona, S. L., Bor, D. H., & Himmelstein, D. U. (2008). Waits to see an emergency department physician: U.S. trends and predictors, 1997–2004. *Health Affairs, 27*(2), w84–w95. doi:10.1377/hlthaff.27.2.w84.

Wilson, B. J., Zimmerman, D., Applebaum, K. G., Kovalski, N., & Stein, C. (2012). Patients who leave before being seen in an urgent care setting. *European Journal of Emergency Medicine, 11*, 11.

Wu, B. U., Banks, P. A., & Conwell, D. L. (2009). Disparities in emergency department wait times for acute gastrointestinal illnesses: Results from the National Hospital Ambulatory Medical Care Survey, 1997–2006. *The American Journal of Gastroenterology, 104*(7), 1668–1673.

Yankovic, N., Glied, S., Green, L. V., & Grams, M. (2010). The impact of ambulance diversion on heart attack deaths. *Inquiry, 47*(1), 81–91.

Chapter 6
Access to Surgery and Medical Consequences of Delays

Boris Sobolev, Adrian Levy, and Lisa Kuramoto

Abstract In this chapter, we present the results of recent wait list studies that quantified the risk of delaying patients awaiting elective cholecystectomy and for patients accepted for coronary artery bypass surgery. Wait lists are a common tool for managing access to elective surgery. When treatment is delayed, the condition of a patient on a surgical wait list may deteriorate and require urgent medical attention. In this case, emergency admission for the awaited procedure may be regarded as an adverse effect of waiting. However, little evidence is available on the health effects of delaying surgery for various conditions. Other than preoperative mortality, adverse events experienced by patients while on a wait list have not been systematically examined. Without these data, appropriate access time for surgery is usually determined on the basis of expert opinion. Our results have implications for developing waiting-time limits for elective surgical procedure.

Keywords Elective surgical procedures • Access to care • Wait lists • Health effects

B. Sobolev (✉)
School of Population and Public Health, The University of British Columbia,
Vancouver, BC, Canada

Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal
Health Research Institute, Vancouver, BC, Canada
e-mail: sobolev@interchange.ubc.ca

A. Levy
Department of Community Health and Epidemiology, Dalhousie University,
Halifax, NS, Canada

L. Kuramoto
Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal
Health Research Institute, Vancouver, BC, Canada

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_6, © Springer Science+Business Media New York 2013

Wait lists are a common tool for managing access to elective—medically necessary, but nonemergency—surgery in publicly funded health systems (Naylor 1991). While queuing according to urgency of intervention, or priority wait-listing, is perceived as a method for facilitating access to treatment within clinically appropriate times (MacCormick et al. 2003), waiting can adversely affect those delayed, causing worsening of symptoms or death (Ray et al. 2001).

When treatment is delayed, the condition of a patient on a surgical waiting list may deteriorate and require urgent medical attention, including emergency surgery. In this case, emergency admission for the awaited procedure may be regarded as an adverse effect of waiting.

Examples of clinical conditions which may require emergency surgery to be performed on patients who are on wait lists include inguinal hernia, spinal cord conditions, abdominal aortic aneurysm, and occlusive coronary artery disease (CAD). Also, routine operating room activity may be seriously disrupted by unexpected nonelective admissions of patients on wait lists (Buhaug 2002).

Alternatively, the patient's condition may deteriorate to such an extent that surgery is no longer possible. In population-based studies, death before coronary artery bypass graft (CABG) surgery, for instance, has been reported to occur in 0.4–1.3 % of patients for whom it is felt surgery can be safely delayed (Bernstein et al. 1997, Légaré et al. 2005, Morgan et al. 1998, Rexius et al. 2004). In addition, even if the surgery is still possible, a longer recovery may be necessary, or other complications may ensue.

Little evidence is available on the health effects of delaying surgery for various conditions (Derrett et al. 1999, McGurran and Noseworthy 2002, Turnbull et al. 2000). Other than mortality, adverse events experienced by patients while on a wait list have not been systematically examined (Morgan et al. 1998, Sobolev et al. 2013). Without such data, appropriate access times for surgery are usually determined on the basis of expert opinion (Naylor et al. 1990b).

In this chapter, we first provide an overview of the Canadian health care system and then present two examples that quantified the risk of delayed treatment for patients awaiting elective cholecystectomy and for patients accepted for CABG surgery. At the end of the chapter, we describe statistical methods for studying the risks of adverse events associated with wait lists. Our results have implications for developing acceptable limits for waiting times for elective surgical procedures.

1 The Canadian Health Care System

Canada's health care coverage is universally available and publicly provided; it is funded through provincial and federal taxes and insurance premiums (Reinhardt 1998). The legal basis of the Canadian health care system, the Canada Health Act, provides coverage for all medically necessary hospital and physician services. This means that Canadians seeking care go to a physician or hospital of their choice and present their health insurance cards.

Provincial governments are responsible for providing health insurance to residents. Individuals do not pay directly for services and no dollar limits or deductibles apply. The Canada Health Act thereby ensures that health care services are made available on the basis of need rather than on an individual's ability to pay. However, while legislation creates a framework, the Canadian health care system is actually a complex arrangement of funding mechanisms worked out between the federal and provincial governments (Levy and Gagnon 2002).

Most physicians and surgeons are paid on a fee-for-service basis, with the upper salary limits in some provinces. Patients are referred to specialists or surgeons by primary care physicians, who are considered the gatekeepers for access to specialized services.

Most Canadian acute care hospitals are operated as private non-profit organizations run by community boards of trustees, voluntary organizations, or municipalities. These hospitals receive global operating budgets established by provincial ministries of health and mostly determined annually by historical expenditures with some adjustments. Hospitals must look after their day-to-day allocation of resources within the operating budget (Klatt 2000).

Wait lists are used extensively as part of hospital or regional responses to limited budgets. Naylor introduced the concept of wait lists as a management tool (Naylor 1991) and a form of rationing (Naylor et al. 1993). In publicly funded health care, wait lists are commonly used to manage access to elective procedures, but the practice raises concerns about the delaying of necessary treatment (Naylor et al. 1995, Noseworthy et al. 2003).

2 Access to Elective Surgery

After a patient is referred, the surgeon assesses the patient and the severity of illness. The decision to operate is taken after surgery is indicated and the patient is deemed a suitable candidate. Patients are placed on the surgeon's wait list if they cannot be operated on immediately.

For non-life-threatening conditions, patients are enrolled on a first-come, first-served basis. For potentially life-threatening conditions, they are registered on a priority wait list. Patients are ranked by how urgently they need treatment, and a priority class is assigned to all patients to determine relative positions on the list. Patients with a higher priority will be selected for service ahead of those with a lower priority, regardless of when they are placed on the list. Patients in the same priority class are ranked in the order of arrival.

Patients are removed from the list if they reconsider the decision for surgery, if they accept surgery from another surgeon, if they decline admission, if they move out of the province, if they are deferred or suspended on medical grounds, if they are suspended for administrative reasons, if they die while awaiting surgery, if the physician decides to try a medical treatment instead of waiting for surgery, if their conditions preclude scheduling of surgery indefinitely, if their conditions improve and make the surgery unnecessary, if the operation is no longer possible, if the operation no longer offers the likelihood of improvement, or when surgery is done.

Access to surgical care in the hospital is usually managed through scheduling demands for service. Scheduling identifies patients available for the next service period and reserves hospital resources to ensure appropriate care before and after an operation (Blake and Carter 1997).

To plan the utilization of the surgical suite resources, the hospital releases, on a periodic basis, blocks of operating room time to each surgical service, which then places patients on the operating room schedule (Magerlein and Martin 1978). Some time slots are set aside for emergency cases. Any time which was not booked is made available to other services.

We use the term scheduling cycle for the sequence of events in surgical scheduling between two releases of operating time blocks. Within services, patients are selected from wait lists and scheduled for operation based on urgency, best use of allocated operating time, and the availability of hospital resources. However, an emergency case is sent to the operating room upon arrival, potentially causing cancelation of scheduled elective operations. On the other hand, if operating room time becomes available unexpectedly, patients may be added to the current schedule if they can come in at short notice. The service access is defined as immediate if patients are admitted within the scheduling cycle that had started at the time they were accepted for service. The access is said to be delayed if patients are admitted within a scheduling cycle that starts after their acceptance for service.

Before being added to the operating room schedule, all patients are assessed by an anesthesiologist as to suitability for surgery. If a patient's condition is not fit for surgery, scheduling of the operation may be postponed. Scheduling a patient for surgery may be also delayed for the following reasons: the patient decides to postpone surgery; a hospital ward, intensive care unit bed, or operating room is unavailable at the time scheduled; or the doctor decides to send the patient for additional preoperative investigation. The availability of other hospital resources is considered in selecting patients for scheduling the operation (Hamilton and Breslawski 1994).

Patients might be reinstated on the list following medical deferral, administrative suspension, self-deferral, or failure to attend (Armstrong 2000a).

3 Access to Coronary Artery Surgery

A specific example of scheduling should clarify the process by which a patient arrives at surgery. In the Canadian province of British Columbia (BC), priority wait lists are commonly used to manage access to elective procedures according to urgency of treatment (Noseworthy et al. 2003). In particular, patients with CAD are prioritized according to angina symptoms, coronary anatomy, and left ventricular function impairment in order to facilitate access to surgical revascularization within clinically appropriate times (Levy et al. 2005).

Initially, a patient presenting with symptoms of CAD is referred to a cardiologist who assesses the need for revascularization. The cardiologist evaluates the results of coronary angiography and decides on treatment (Grech 2003). If coronary angioplasty is not indicated, patients are referred to a cardiac surgeon who assesses their need and suitability for CABG.

When urgent assessment is required, patients are transferred to a hospital cardiac ward directly from the catheterization laboratory. If suitable for surgery, such patients remain in hospital until the operation.

Alternatively, patients are scheduled for an outpatient consultation with the cardiac surgeon at a later date. Following the consultation in which surgery is identified as necessary, surgeons register on their wait lists patients who require CABG and decide to undergo the operation in one of the four BC tertiary care hospitals at which the operation is performed, and at which the specific surgeon has admitting rights. A priority category is assigned to each patient according to the urgency of treatment.

The suggested time to surgery is 3 days for patients with left main coronary artery stenosis greater than 70 % (urgent group); 6 weeks for patients with persistent unstable angina, impaired left ventricular function, and significant obstruction defined as left-main stenosis, triple-vessel disease or double-vessel disease with significant proximal left anterior descending stenosis (semiurgent group); and 12 weeks for patients with intractable chronic angina, normal left ventricular function, and single-vessel disease or double-vessel disease with no lesion in the proximal left anterior descending artery (nonurgent group) (Levy et al. 2005).

At each hospital, the patient's access to surgery is managed through scheduling of operating room time. Patients are selected for scheduling both from hospital cardiac wards and from the surgical wait lists based on allocated operating room time-slots and priority.

Before being added to the operating room schedule, each patient is reassessed by an anesthesiologist as to suitability for surgery. The operation may be postponed for any of the reasons noted above, including if the anesthesiologist requests additional preoperative investigations, or when an emergency case comes in and scheduled operations are canceled. On the other hand, already scheduled patients may undergo surgery ahead of their scheduled dates if an operating room time slot becomes available.

A diagram showing the patient's path from presentation with symptoms of CAD to CABG can be found elsewhere (Sobolev and Kuramoto 2008).

4 Two Studies on Adverse Events While Waiting for Surgery

In order to understand the time-related nature of adverse events associated with wait lists, it will help to look at the results of recent studies of wait lists for elective surgical procedures. The full investigative methods are published elsewhere

(Sobolev et al. 2003, 2013, 2006a, b), so here we provide the most relevant elements for wait list outcomes.

4.1 Unplanned Emergency Admission While Awaiting Cholecystectomy

In the first example, that of patients with biliary colic caused by cholelithiasis, it can be seen that extended treatment delays may increase the probability that the patient will be admitted for cholecystectomy on an emergency basis. Emergency admission may be associated with more frequent or more severe attacks of biliary colic or other biliary complications such as acute cholecystitis, obstructive jaundice, cholangitis, or pancreatitis (Friedman 1993).

In order to assess the relationship between time spent on a wait list and the risk of emergency admission for this patient cohort, this study reviewed the timing and type of operations performed on patients on cholecystectomy wait lists maintained by the Department of Surgery, Queens University, Kingston, O.N., Canada (Sobolev et al. 2003). In this setting, eight general surgeons performed cholecystectomy, and there was no system for ranking the urgency of the patient. Each surgeon's office managed its wait list independently.

Surgeons on call made the decision to operate on patients who presented to the emergency department by evaluating (a) the clinical presentation for symptoms of increased pain or fever and signs of persisting or worsening abdominal tenderness, guarding or rebound or (b) the ultrasonographic finding of a thick-walled gallbladder with pericholecystic fluid or a positive finding of hepatobiliary iminodiacetic acid on radionuclide scan, or both (a) and (b).

Data on the timing and type of surgery were retrieved from the electronic hospital information system from fiscal years 1997 to 2000. The primary outcome investigated was emergency admission for cholecystectomy due to the worsening of symptoms while awaiting elective surgery. A wait list time was calculated for each patient based on the number of weeks from the last consultation visit to elective or emergency surgery. This approach assumes the last visit before surgery was the date when the decision to operate was made (DeCoster et al. 1999).

Elective patients spent a total of 5,712 person-weeks waiting to be admitted. The average weekly number of elective operations was 12.4 [95 % confidence interval (CI) 11.6–13.3] per 100 patients on the list.

The rate differed across enrolment periods, from 10.3 (9.1–11.5) in fiscal year 1997/1998 to 15.1 (13.4–16.9) in 1998/1999 to 13.2 (11.5–15.0) in 1999/2000. The median length of stay on the list was 6 weeks. However, there was considerable variation in individual waiting times. At the present time, there is no recommended waiting time for cholecystectomy.

The probability of undergoing elective surgery increased rapidly from 25 % within 3 weeks of the last clinic visit, to 50 % at 6 weeks and 75 % at 10 weeks, and

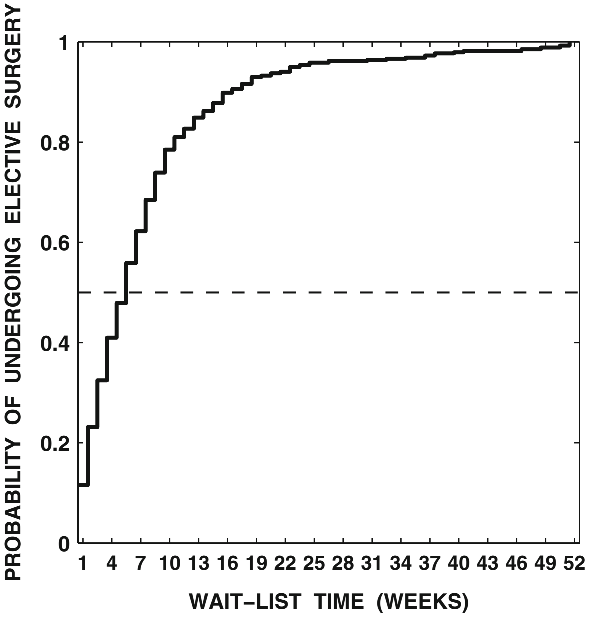


Fig. 6.1 Estimated probabilities of elective surgery

then gradually reached a plateau. Although 90 % of patients underwent surgery by 17 weeks, the remaining 10 % waited another 1–35 weeks (total, 18–52 weeks) for their operation (Fig. 6.1).

Surgeons with a low volume of cholecystectomies (less than 20 per year) operated on the majority of patients with extended delays. In general, low-volume surgeons had a primary interest in surgical oncology. This may explain the order in which their patients accessed cholecystectomy during the waiting period: the cholecystectomies were seen to be less medically necessary than oncological procedures.

Overall, 6.7 % of the patients waiting for elective cholecystectomy underwent surgery through unplanned emergency admissions. The proportion varied significantly across the categories of patient and service characteristics. Women patients, young (less than 25 years) and old (more than 75 years) patients, and patients operated on by lower volume surgeons, were admitted as emergency cases more often.

The average weekly emergency admission rate of patients on the wait lists was 0.9 (95 % CI 0.7–1.2) per 100 patients. However, the weekly emergency admission rate increased from 0.8 to 5.7 per 100 patients from the interval of the first 4 weeks to the interval of 40–52 weeks (see Fig. 6.2).

When adjusted for sex, age decade, period, and surgeon volume, the emergency admission rate was more than 1.5 times higher after 20 weeks, 2 times higher after 28 weeks, and 7 times higher after 40 weeks relative to the first 4 weeks of wait list time.

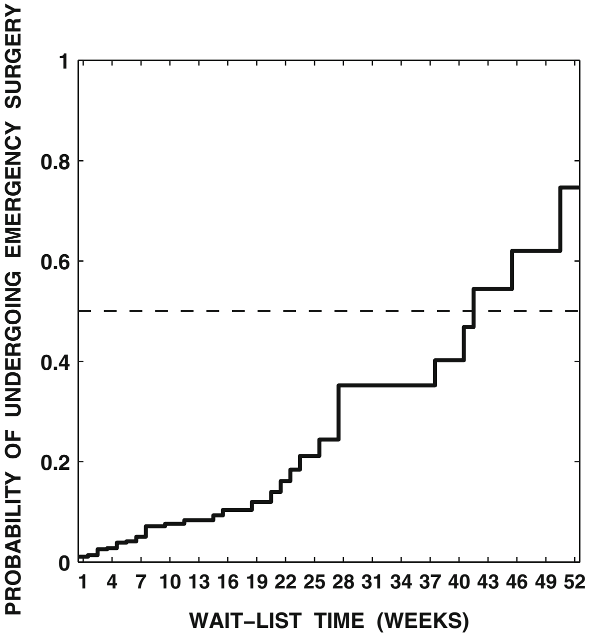


Fig. 6.2 Estimated probabilities of emergency surgery

Table 6.1 Probabilities of elective and emergency admission by different wait list intervals, conditional on remaining on the list until the start of each interval

Interval (weeks)	Elective	Emergency
< 4	0.32	0.02
4–7	0.44	0.02
8–11	0.49	0.03
12–19	0.53	0.03
20–27	0.43	0.11
28–39	0.33	0.19
40–52	0.60	0.40

Patients waiting 20 weeks or more were more likely to undergo emergency admission than those waiting shorter times after adjustment for age, sex, period, and surgeon volume. Table 6.1 shows the conditional probability of emergency admission during the seven wait list intervals compared with the corresponding figures for elective admission. Although the probability of emergency admission during the first 19 weeks was low, after 20 weeks the probability started increasing and approached 40 % in the interval of 40–52 weeks. Of 46 patients who waited for more than 20 weeks, 28 % were admitted as emergency cases, compared with 5 % of those who waited less than 20 weeks (715 patients).

The average weekly rates were 2.4 (95 % CI 1.3–4.0) and 0.7 (95 % CI 0.5–1.0) per 100 patients in these two groups, respectively, with the adjusted rate ratio being 2.7 (95 % CI 2.0–3.7).

4.2 *Adverse Events While Waiting for Cardiac Surgery*

In the second example, we describe adverse events among patients registered for CABG in BC (Sobolev et al. 2013). We studied records of patients in whom surgical revascularization was indicated at the time of consultation with a cardiac surgeon. The primary outcomes were the occurrence of death from all causes and unplanned emergency surgery while awaiting planned surgery. Data were extracted from a prospective database of all patients who were accepted for isolated first-time CABG in BC between 1992 and 2005.

At 52 weeks of follow-up, 83 % of patients had undergone planned surgery, 1 % died while awaiting surgery, 3 % had unplanned emergency surgery, 6 % of patients remained on the lists, and 7 % dropped out during follow-up for various reasons: continued to receive medical treatment (2 %), declined surgery (2 %), transferred to another surgeon or hospital ($<$ 1 %), or removed from the list due to other reasons (3 %). While the majority of the urgent patients had received surgery by 52 weeks, over one-tenth of nonurgent patients and less than 5 % of semiurgent patients were still on the list at 52 weeks.

In total, 0.9 % (95 % CI 0.7–1.0) of patients died before surgery. At 0.5 % (95 % CI 0.0–1.1), the urgent group had the smallest proportion of deaths on the wait list, whereas, 0.7 % (95 % CI 0.5–0.9) and 1.4 % (95 % CI 0.9–1.9) died before the operation in semiurgent and nonurgent groups, respectively.

Overall, the rate of death from all causes was 0.6 (95 % CI 0.5–0.7) per 1,000 patient-weeks. The rate varied from 0.9 (95 % CI 0.0–1.7) in the urgent group to 0.5 (95 % CI 0.4–0.6) in the semiurgent group and 0.6 (95 % CI 0.4–0.8) in the nonurgent group. After adjustment for sex, age decade, comorbidities at registration, calendar period of registration, and time between catheterization and registration, the death rate in the nonurgent group was similar to that of the semiurgent group [odds ratio (OR) = 1. 07, 95 % CI 0.69–1.65] (Table 6.2).

Table 6.2 Weekly rate of all-cause preoperative death in relation to urgency group, for patients registered for bypass surgery in 1992–2005, as measured by odds ratio derived from discrete-time survival regression models

	No. of	Total		
Urgency group	deaths	wait[a]	Death rate[b] (95 % CI)	OR[c] (95 % CI)
Urgent	4	4,676	0.9 (0.0–1.7)	–
Semiurgent	63	123,138	0.5 (0.4–0.6)	1.00
Nonurgent	32	53,232	0.6 (0.4–0.8)	1.07 (0.69–1.65)
All patients[d]	104	184,820	0.6 (0.5–0.7)	–

CI confidence interval, *OR* odds ratio[a]Waitingtime measured in patient-weeks

[b]Weekly rate was calculated as the number of all-cause deaths divided by the sum of waiting times (per 1,000 patient-weeks)

[c]Adjusted for sex, age decade, comorbidities at registration, calendar period of registration, and time between catheterization and registration

[d]Includes additional patients with urgency not provided

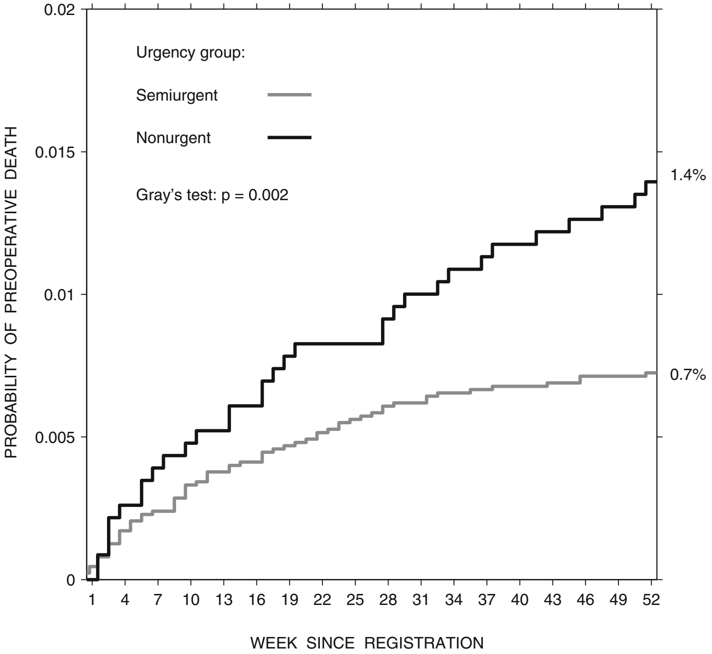


Fig. 6.3 Estimated cumulative incidence of preoperative death by wait-list week since registration in semiurgent and nonurgent groups

One measure for summarizing the risk of death in a competing risk setting is the probability of death by a certain time (Pepe and Mori 1993). Figure 6.3 shows the cumulative incidence of death while waiting for planned surgery in the semiurgent and nonurgent groups. The nonurgent group had a greater probability of death before planned surgery than the semiurgent group for almost all wait list weeks (Gray's test = 9. 4, df = 1, p = 0. 002). Compared to the semiurgent group, the odds of death before planned surgery was 1.9 times greater in the nonurgent group (OR = 1. 92, 95 % CI 1.25–2.95), after adjustment for sex, age decade, comorbidities at registration, calendar period of registration and time between catheterization and registration.

The difference between the cumulative incidence functions became larger with time on the list, approaching respective proportions of deaths in each group by 52 weeks. Considering that death rates were similar in these two groups, the higher proportion dying among nonurgent patients suggests that the longer waiting times in this group contribute to a higher chance of death before planned surgery.

In total, 3.2 % (95 % CI 2.9–3.5) of patients had unplanned emergency surgery. At 6.5 % (95 % CI 4.7–8.3), the urgent group had the highest proportion of unplanned emergency surgery, whereas 3.0 % (95 % CI 2.7–3.4) and 2.8 % (95 % CI 2.1–3.5) had emergency surgery before planned surgery in semiurgent and nonurgent groups, respectively.

Table 6.3 Weekly rate of unplanned emergency surgery in relation to urgency group, for patients registered for bypass surgery in 1992–2005, as measured by odds ratio derived from discrete-time survival regression models

	No. of	Total	Emergency surgery	OR[c]
Urgency group	emergency surgeries	wait[a]	rate[b] (95 % CI)	(95 % CI)
Urgent	48	4,676	10.3 (7.4–13.2)	4.9 (3.4–7.2)
Semiurgent	264	123,138	2.1 (1.9–2.4)	1.0
Nonurgent	65	53,232	1.2 (0.9–1.5)	0.7 (0.5–1.0)
All patients[d]	382	184,820	2.1 (1.9–2.3)	–

CI confidence interval, *OR* odds ratio[a]Waitingtime measured in patient-weeks

[b]Weekly rate was calculated as the number of unplanned emergency surgeries divided by the sum of waiting times (per 1,000 patient-weeks)

[c]Adjusted for sex, age group, coronary anatomy, comorbidities at registration, calendar period at registration, institution at registration, institution at catheterization, mode of admission at catheterization, urgency at admission for catheterization, and time between catheterization and registration

[d]Includes additional patients with urgency not provided

The rate of unplanned emergency surgery was 2.1 (95 % CI 1.9–2.3) per 1,000 patient-weeks. The rate decreased from 10.3 (95 % CI 7.4–13.2) in the urgent group to 2.1 (95 % CI 1.9–2.4) in the semiurgent group and to 1.2 (95 % CI 0.9–1.5) in the nonurgent group. After adjustment for sex, age group, coronary anatomy, comorbidities at registration, calendar period at registration, institution at registration, institution at catheterization, mode of admission at catheterization, urgency at admission for catheterization, and time between catheterization and registration, the unplanned emergency surgery rate in the urgent group was about five times higher than the semiurgent group (OR = 4. 93, 95 % CI 3.38–7.18) and the emergency surgery rate in the nonurgent group was about 30 % lower than the semiurgent group (OR = 0. 72, 95 % CI 0.54–0.97) (Table 6.3).

Figure 6.4 shows the probability of unplanned emergency surgery before planned surgery by a certain time in the urgent, semiurgent, and nonurgent groups. The urgent group had a greater probability of emergency surgery before planned surgery than the semiurgent and nonurgent groups (Gray's test = 29. 2, df = 2, $p < 0.\ 001$). The cumulative incidence functions were not different between the semiurgent and nonurgent groups (Gray's test = 0. 28, df = 1, $p = 0.\ 60$). Compared to the semiurgent group, the odds of emergency surgery before planned surgery were similar in the nonurgent group (OR = 0. 87, 95 % CI 0.63–1.20), after adjustment for sex, age group, coronary anatomy, comorbidities at registration, calendar period of registration, institution at registration, institution at catheterization, mode of admission at catheterization, urgency at admission for catheterization, and time between catheterization and registration.

Considering that the unplanned emergency rate was lower in the nonurgent group than in the semiurgent group, similar proportions of emergency surgery before planned surgery in these groups suggest that longer waiting times in the nonurgent group equalized the risk of emergency surgery before planned surgery.

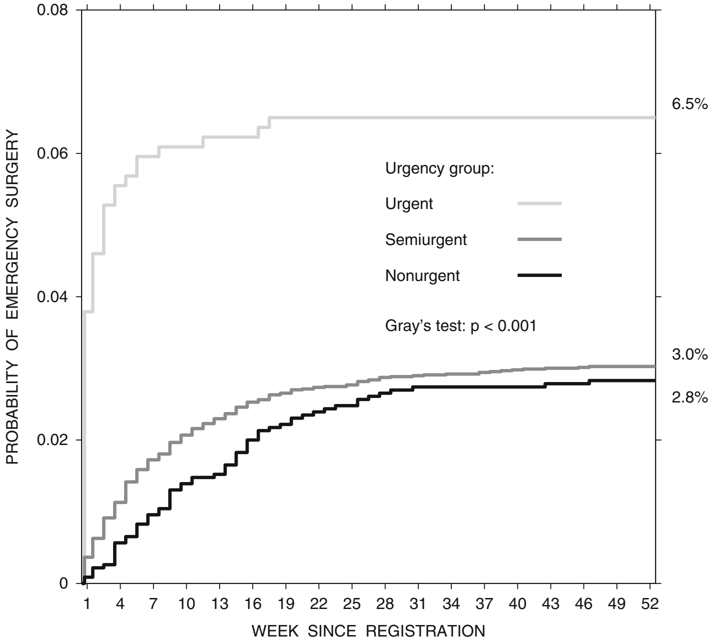


Fig. 6.4 Estimated cumulative incidence of unplanned emergency surgery by wait-list week since registration in urgent, semiurgent, and nonurgent groups

Another measure suggested for summarizing the risk of adverse events over time in the competing-risk setting is the probability of an adverse event conditional on not having experienced the competing event by a certain time (Lin 1997, Pepe and Mori 1993). Using this approach, we sought to improve the estimates of the risk of adverse events associated with delayed CABG for patients requiring and suitable for surgical revascularization. We therefore estimated the time-dependent probability of death and unplanned emergency surgery, given that planned CABG was not performed by certain times.

The extent of disease was a major factor influencing time to surgery. The differences in the cumulative incidence of planned surgery were significant over time between groups with higher incidence in the urgent group (Gray's test $= 494.5, \mathrm{df} = 2, p < 0.0001$) (Fig. 6.5). The average planned surgery rate was 140.1 per 1,000 patient-weeks in the urgent group, compared to 65.5 per 1,000 patient-weeks in the semiurgent group and 30.6 per 1,000 patient-weeks in the nonurgent group.

To compare proportions of patients dying by a certain time among those who had not received planned surgery by that time, we calculated the conditional probability of death from all causes in each group (Fig. 6.6). The conditional probability of death from all causes was greater in the semiurgent group than in the nonurgent group (Pepe's two-sample test $= 2.9, p = 0.002$).

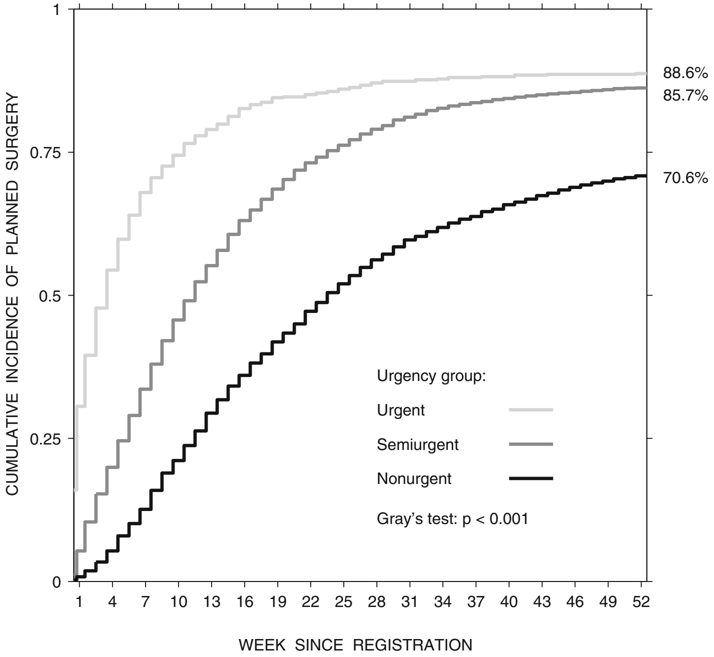


Fig. 6.5 Estimated cumulative incidence of planned surgery by week since registration in urgent, semiurgent, and nonurgent groups

Among patients whose delay to CABG exceeded 8, 16, 32, and 52 weeks, the probability of death from all causes was 0.4 %, 1.2 %, 5.5 %, and 13.0 % in the semiurgent group, and 0.5 %, 1.0 %, 3.2 %, and 7.9 % in the nonurgent group.

The conditional probability of unplanned emergency surgery was highest in the urgent group and lowest in the nonurgent group (Fig. 6.7). The conditional probability of emergency surgery was greater in the semiurgent group than in the nonurgent group (Pepe's two-sample test = 9. 8, $p < 0.\ 001$). Among patients whose delay to CABG exceeded 8, 16, 32, and 52 weeks, the probability of unplanned emergency surgery was 21.0 %, 38.7 %, 63.9 %, and 75.9 % in the urgent group, 3.9 %, 7.3 %, 20.7 %, and 37.8 % in the semiurgent group, and 1.3 %, 3.3 %, 8.3 %, and 14.1 % in the nonurgent group, respectively.

In each group, we estimated the time-dependent conditional probability that a patient, who may die, undergo unplanned emergency surgery, or undergo planned surgery, dies if not operated by certain times. Among patients delayed without treatment for 52 weeks, an estimated 13.0 % died in the semiurgent group and 7.9 % died in the nonurgent group from all causes. Similarly, an estimated 75.9 %, 37.8 %, and 14.1 % had unplanned emergency surgery in the urgent, semiurgent, and nonurgent groups, respectively, among patients delayed without treatment for 52 weeks.

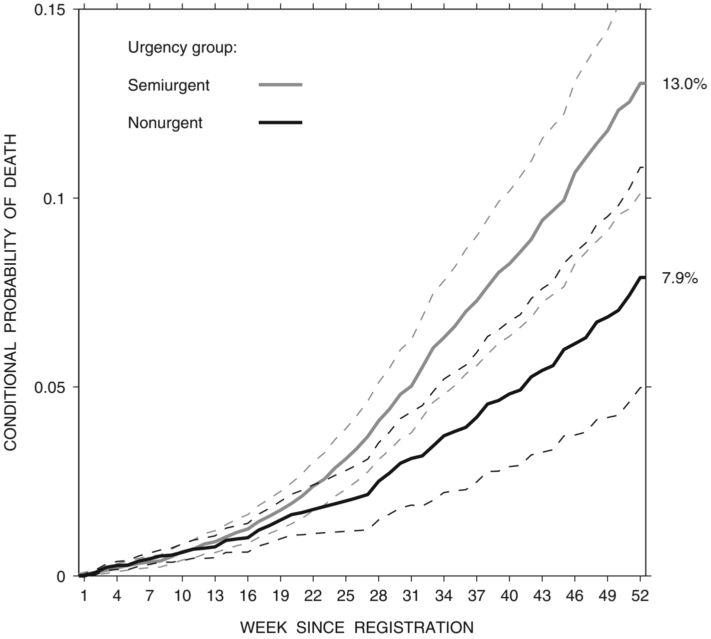


Fig. 6.6 Estimated conditional probability of all-cause death and 95 % confidence intervals by week since registration in semiurgent and nonurgent groups

5 Discussion

Implicit in treatment delays is a risk that the health status of patients may worsen while they are awaiting treatment. Adverse events experienced by patients while on a wait list have not been systematically examined in the literature. Without these data, appropriate access times for surgery are usually determined on the basis of expert opinion. Mortality is the main adverse event that has been examined. Other outcomes that should be considered include unplanned emergency admission, upgrade in severity, and cancelation of operation due to patient deterioration.

The research presented here is some of the first available that can be used to identify appropriate access times for patients waiting for elective operations. Our data included the health effects of delaying elective cholecystectomy and delaying patients registered on a wait list for CABG surgery. For adverse events while waiting for cholecystectomy, the main findings are that (Sobolev et al. 2003):

- The greater the length of time to treatment, the more likely it is that the patient will have to be admitted as an emergency case. The emergency admission rate in the population we studied increased 1.5 times at 20 weeks and continued to rise thereafter.

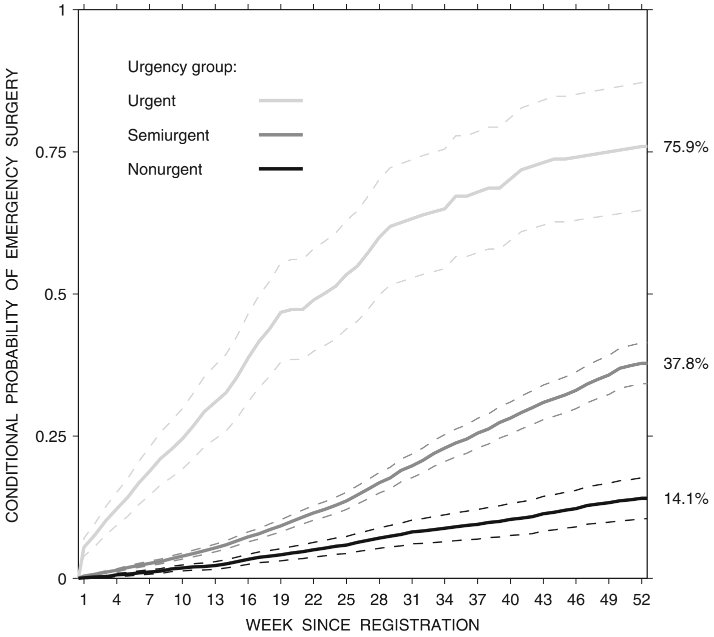


Fig. 6.7 Estimated conditional probability of unplanned emergency surgery and 95 % confidence intervals by week since registration in urgent, semiurgent, and nonurgent groups

- Nearly 7 % of patients waiting for elective cholecystectomy underwent surgery through unplanned emergency admissions.
- Longer waiting times were associated with the surgeons who only did low volumes of cholecystectomies, less than 20 per year.

These results have implications for developing target access times for elective surgery. The findings suggest that patients with biliary colic awaiting elective cholecystectomy for longer than 20 weeks have a substantially increased risk for development of acute symptoms that require an emergency operation. Therefore, an initial recommendation of 20 weeks might be considered as the maximum recommended waiting time for cholecystectomy. The 7 % rate of unplanned surgeries is bound to have a large impact on operating room schedules and resources.

For adverse events while waiting for CABG, the main findings are that:

- The risk of death from all causes was associated with longer wait list times.
- If patients remain unoperated by 1 year, 13 % of semiurgent patients and 8 % of nonurgent patients die from all causes.
- Longer delays contributed to a higher proportion of CABG candidates dying before surgery from all causes in the nonurgent (1.4 %) compared to the semiurgent group (0.7 %) despite similar rates of preoperative death observed in both groups.

- These longer delays in the nonurgent group also equalized the proportion of CABG candidates with unplanned emergency surgery in the semiurgent group (3.0 %) and nonurgent group (2.8 %).
- The use of statistical methods for competing risks improved estimates of the probability of death and the probability of unplanned emergency surgery associated with delaying surgery.

The contribution of the study on CABG is the estimated conditional probabilities of death while on a wait list derived from the population-based prospective database. These summary probabilities are not usually reported in prospective studies of mortality on wait lists. Section "Appendix" describes the rationale for this approach and lists previous studies that examined the risks of death while waiting for CABG. The findings suggest CABG operations should take place within the recommended 6 and 12 weeks for semiurgent and nonurgent cases as the protracted delay for surgical revascularization when it is indicated carries a significant risk of death even in patients judged to be at low risk.

In both examples, cholecystectomy and CABG, we observed that specific populations—women and the elderly—have higher risks of adverse events while awaiting elective surgery. If confirmed in other studies, this may have implications for the acceptable wait list times in these groups.

5.1 Limitations

In the cholecystectomy example, the retrospective nature of the data may be considered an important limitation of the study. Prospective studies examine how long patients accepted for treatment wait for surgery, whereas retrospective studies examine how long the patients who were admitted were required to wait after enrolment (Armstrong 2000b). If every wait ended in admission, the two study designs would generate equivalent results. However, patients accepted for treatment may expect to be removed from the wait list for reasons other than admission (Sobolev et al. 2000).

If patients removed from the list without surgery are not accounted for, the estimated probabilities of undergoing treatment may be biased toward a higher rate. Also, the analysis lacked data on comorbid medical conditions. In general, a large number of comorbid conditions may prevent aggressive treatment. Therefore, given its possible association with delay in treatment, comorbidity is a potentially confounding factor for the observed relationship between time on the wait list and emergency admission.

In the CABG example, data were extracted from a prospective database, so there was less potential for bias in the results. Other limitations were the potential misclassification of dates and the priority assignment. Some assurance that procedure dates were recorded accurately comes from the finding that the operation date for 99.3 % of records were between admission and separation dates (or within a few days), as reflected in discharge abstracts in hospital separations. Retrieved from the database, the priority category is a composite variable based on clinical information. The observation that higher priority patients were more likely to undergo CABG through the direct admission indicates that the degree of misclassification of priority was likely small.

The analyses for the CABG example were conducted with adjustment for patient and access management characteristics. The existing literature suggests that elderly patients are more likely to undergo revascularization as an urgent procedure (Christenson et al. 1999), that smaller diameter of the coronary vessels may account for the higher risk of adverse cardiovascular events among women (O'Connor et al. 1996), that coexisting conditions may delay open heart surgery (Naylor et al. 1990a), that institutional constraints and individual care providers may affect clinical outcomes (DeLong et al. 2001), that patients with a lower socioeconomic status may wait longer for cardiac surgery (Pell et al. 2000), and that changes in practice or the availability of supplementary funds may reduce the waiting time until surgery (Levy et al. 2005). To identify comorbidities at the time of registration, we used diagnoses reported in the DAD within 1 year prior to registration. The reference category was defined as no coexisting conditions. The first comparison category was defined as patients with any of the following conditions at presentation: congestive heart failure, diabetes mellitus, chronic obstructive pulmonary disease, cancer, or rheumatoid arthritis (Naylor et al. 1992). The second comparison category was defined as patients presenting with other coexisting chronic conditions, as defined elsewhere (Romano et al. 1993). Other confounders include hospital booking catheterization to address variation in standards and calendar year of surgery decision as a proxy of changes in practice and available funding. We also included the time between catheterization and surgery, the mode of admission for catheterization, urgency at admission for catheterization, which may differ substantially among hospitals affecting estimates of the total of delays in undergoing the operation (Légaré et al. 2010). The time between catheterization and registration was computed as the number of calendar weeks. The catheterization dates were obtained from the CIHI DAD and defined as the most recent diagnostic (Canadian Classification of Procedure (CCP) codes 4892–4898, 4996, 4997) or therapeutic (CCP codes 4802, 4803, 4809) catheterization performed within 1 year preceding and including the date of booking. We used the date of most recent catheterization procedures (diagnostic or therapeutic) because the results of this procedure are most likely linked to decision to operate (King et al. 2004).

Acknowledgements This research was supported by the Canada Research Chair Program (B.S.), the Michael Smith Foundation for Health Research Scholar Program (A.L.), and Canadian Foundation for Innovation (B.S., A.L.). The following cardiac surgeons are members of the BCCR Surgical Research Committee: Drs. James Abel, Richard Brownlee, Larry Burr, Anson Cheung, James Dutton, Guy Fradet, Virginia Gudas, Robert Hayden, Eric Jamieson, Michael Janusz, Shahzad Karim, Tim Latham, Jacques LeBlanc, Sam Lichtenstein, Hilton Ling, John Ofiesh, Michael Perchinsky, Peter Skarsgard, and Frank Tyers.

Appendix

In the following sections, we provide a brief overview of the methods used in our analyses. The detailed introduction to this methodology could be found in (Sobolev and Kuramoto 2008).

A.1 Marginal and Conditional Probability Functions

Competing risks naturally arise in wait list settings. A competing event is "any event whose occurrence either precludes the occurrence of another event under examination or fundamentally alters the probability of occurrence of this other event" (Gooley et al. 1999). A subject on the wait list is considered at risk for an adverse event from registration time until a censoring event, surgery, or an adverse event before surgery. Other events will be classified as competing risk events if their occurrence precludes the subsequent development of the primary event, or censoring events if their occurrence precludes observation, but not development, of the primary event. For example, relative to death before surgery, undergoing the planned operation is a competing risk; loss to follow-up is a censoring event.

In quantifying the risk of adverse events on wait lists, the Kaplan–Meier method is commonly used to estimate the cumulative probability of an event by a certain time after registration for the operation (Jackson et al. 1999, Koomen et al. 2001, Ray et al. 2001). It has been established, however, that the complement of the Kaplan–Meier estimator overestimates the incidence of the event in the competing risks setting (Gooley et al. 1999).

As patients on wait lists are subject to the competing events of surgery, death, or removal for other reasons, the method produces probability estimates that are only valid in a hypothetical situation when all competing risks are removed prior to the event without altering the risk of the adverse event of interest (Gaynor et al. 1993). This approach implicitly assumes that time to surgery and time to the adverse event are independent. Without this assumption, the Kaplan–Meier estimator is not valid and should not be used (Alberti et al. 2003). Furthermore, the independence of wait list outcomes cannot be verified from data, and the assumption may not be realistic, as a low probability of the adverse event may indicate either a low risk of this event or a high surgery rate.

Other investigators have reported the incidence of preoperative death per time unit of waiting for CABG (Bernstein et al. 1997, Cox et al. 1996, Koomen et al. 2001, Morgan et al. 1998, Ray et al. 2001, Rexius et al. 2004, Seddon et al. 1999). Although accurately describing the instantaneous hazard, death rates cannot be converted into the probabilities of death without an unrealistic and unverifiable assumption that time to surgery and time to death are independent (Gooley et al. 1999). Plomp and colleagues have reported on the variation in time to deaths among those who died before surgery (Plomp et al. 1999), but the proportion of CABG candidates dying over follow-up could not be derived from their figures.

A.2 Regression Models

Regression methods for pseudo-values of CIF for death are used to estimate the effect of urgency group, while adjusting for patient and access management (Klein and Andersen 2005). Pseudo-values of CIF for death are computed in the presence of surgery and other competing events. In the absence of censoring, for each patient, pseudo-values of CIF for death correspond to a series of binary variables equal to 0 before and 1 at or after death. Using generalized estimation equations to adjust for subject-level correlation between pseudo-values, the CIF is modeled at all distinct, observed event times. The working weight matrix is fixed to be the estimated product-moment correlation matrix between pseudo-values of the CIF. The effect of urgency was measured by ORs, adjusted for patient and access management characteristics.

References

Alberti, C., Metivier, F., Landais, P., Thervet, E., Legendre, C., & Chevret, S. (2003). Improving estimates of event incidence over time in populations exposed to other events—Application to three large databases. *Journal of Clinical Epidemiology, 56*, 536–545.

Armstrong, P. W. (2000a). First steps in analysing nhs waiting times: Avoiding the 'stationary and closed population' fallacy. *Statistics in Medicine, 19*, 2037–2051.

Armstrong, P. W. (2000b). Unrepresentative, invalid and misleading: Are waiting times for elective admission wrongly calculated? *Journal of Epidemiology and Biostatistics, 5*, 117–123.

Bernstein, S. J., Rigter, H., Brorsson, B., Hilborne, L. H., Leape, L. L., Meijler, A. P., et al. (1997). Waiting for coronary revascularization: A comparison between New York state, the Netherlands and Sweden. *Health Policy, 42*, 15–27.

Blake, J. T., & Carter, M. W. (1997). Surgical process scheduling: A structured review. *Journal of Social and Health Systems, 5*, 17–30.

Buhaug, H. (2002). Long waiting lists in hospitals. *British Medical Journal, 324*, 252–253.

Christenson, J. T., Simonet, F., & Schmuziger, M. (1999). The influence of age on the outcome of primary coronary artery bypass grafting. *Journal of Cardiovascular Surgery (Torino), 40*(3), 333–338.

Cox, J. L., Petrie, J. F., Pollak, P. T., & Johnstone, D. E. (1996). Managed delay for coronary artery bypass graft surgery: The experience at one Canadian center. *Journal of the American College of Cardiology, 27*, 1365–1373.

DeCoster, C., Carriere, K. C., Peterson, S., Walld, R., & MacWilliam, L. (1999). Waiting times for surgical procedures. *Medical Care, 37*, JS187–JS205.

DeLong, E. R., Nelson, C. L., Wong, J. B., Pryor, D. B., Peterson, E. D., Lee, K. L., et al. (2001). Using observational data to estimate prognosis: An example using a coronary artery disease registry. *Statistics in Medicine, 20*(16), 2505–2532.

Derrett, S., Paul, C., & Morris, J. M. (1999). Waiting for elective surgery: Effects on health-related quality of life. *International Journal for Quality in Health Care, 11*, 47–57.

Friedman, G. D. (1993). Natural history of asymptomatic and symptomatic gallstones. *American Journal of Surgery, 165*, 399–404.

Gaynor, J. J., Feuer, E. J., Tan, C. C., Wu, D. H., Little, C. R., Straus, D. J., et al. (1993). On the use of cause-specific failure and conditional failure probabilities. *Journal of the American Statistical Association, 88*, 400–409.

Gooley, T. A., Leisenring, W., Crowley, J., & Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: New representations of old estimators. *Statistics in Medicine, 18*, 695–706.

Grech, E. D. (2003). Pathophysiology and investigation of coronary artery disease. *British Medical Journal, 326*, 1027–1030.

Hamilton, D. M., & Breslawski, S. (1994). Operating room scheduling. Factors to consider. *AORN Journal, 59*, 665–680.

Jackson, N. W., Doogue, M. P., & Elliott, J. M. (1999). Priority points and cardiac events while waiting for coronary bypass surgery. *Heart, 81*, 367–373.

King, K. M., Ghali, W. A., Faris, P. D., Curtis, M. J., Galbraith, P. D., Graham, M. M., et al. (2004). Sex differences in outcomes after cardiac catheterization—Effect modification by treatment strategy and time. *Journal of the American Medical Association, 291*(10), 1220–1225.

Klatt, I. (2000). Understanding the Canadian health care system. *Journal of Financial Service Professionals, 54*, 42–51.

Klein, J. P., & Andersen, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics, 61*, 223–229.

Koomen, E. M., Hutten, B. A., Kelder, J. C., Redekop, W. K., Tijssen, J. G., & Kingma, J. H. (2001). Morbidity and mortality in patients waiting for coronary artery bypass surgery. *European Journal of Cardiothoracic Surgery, 19*, 260–265.

Légaré, J. F., Li, D., & Buth, K. J. (2010). How established wait time benchmarks significantly underestimate total wait times for cardiac surgery. *The Canadian Journal of Cardiology, 26*(1), e17–e21.

Légaré, J. F., MacLean, A., Buth, K. J., & Sullivan, J. A. (2005). Assessing the risk of waiting for coronary artery bypass graft surgery among patients with stenosis of the left main coronary artery. *Canadian Medical Association Journal, 173*, 371–375.

Levy, A. R., & Gagnon, Y. M. (2002). Canadian formulary decisions: Does pharmacoeconomics matter? *Pharmaceutical News, 9*, 47–55.

Levy, A., Sobolev, B., Hayden, R., Kiely, M., Fitzegrald, M., & Schechter, M. (2005). Time on wait lists for coronary bypass surgery in British Columbia, Canada, 1991–2000. *BMC Health Services Research, 5*, 22.

Lin, D. Y. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. *Statistics in Medicine, 16*, 901–910.

MacCormick, A. D., Collecutt, W. G., & Parry, B. R. (2003). Prioritizing patients for elective surgery: A systematic review. *ANZ Journal of Surgery, 73*, 633–642.

Magerlein, J. M., & Martin, J. B. (1978). Surgical demand scheduling: A review. *Health Services Research, 13*, 418–433.

McGurran, J., & Noseworthy, T. (2002). Improving the management of waiting lists for elective healthcare services: Public perspectives on proposed solutions. *Hospital Quarterly, 5*, 28–32.

Morgan, C. D., Sykora, K., & Naylor, C. D. (1998). Analysis of deaths while waiting for cardiac surgery among 29,293 consecutive patients in Ontario, Canada. *Heart, 79*, 345–349.

Naylor, C. D. (1991). A different view of queues in Ontario. *Health Affairs, 10*, 110–128.

Naylor, C. D., Baigrie, R. S., Goldman, B. S., & Basinski, A. (1990a). Assessment of priority for coronary revascularisation procedures. *Lancet, 335*, 1070–1073.

Naylor, C. D., Basinski, A., Baigrie, R. S., Goldman, B. S., & Lomas, J. (1990b). Placing patients in the queue for coronary revascularization: Evidence for practice variations from an expert panel process. *American Journal of Public Health, 80*, 1246–1252.

Naylor, C. D., Levinton, C. M., & Baigrie, R. S. (1992). Adapting to waiting lists for coronary revascularization. Do Canadian specialists agree on which patients come first? *Chest*, 101(3), 715–722.

Naylor, C. D., Levinton, C. M., Wheeler, S., & Hunter, L. (1993). Queueing for coronary surgery during severe supply-demand mismatch in a Canadian referral centre: A case study of implicit rationing. *Social Science and Medicine, 37*, 61–67.

Naylor, C. D., Sykora, K., Jaglal, S. B., & Jefferson, S. (1995). Waiting for coronary artery bypass surgery: Population-based study of 8517 consecutive patients in Ontario, Canada. The steering committee of the adult cardiac care network of Ontario [see comments]. *Lancet, 346*, 1605–1609.

Noseworthy, T. W., McGurran, J. J., & Hadorn, D. C. (2003). Waiting for scheduled services in Canada: Development of priority-setting scoring systems. *Journal of Evaluation in Clinical Practice, 9*, 23–31.

O'Connor, N. J., Morton, J. R., Birkmeyer, J. D., Olmstead, E. M., & O'Connor, G. T. (1996). Effect of coronary artery diameter in patients undergoing coronary bypass surgery. Northern New England cardiovascular disease study group. *Circulation, 93*(4), 652–655.

Pell, J. P., Pell, A. C., Norrie, J., Ford, I., & Cobbe, S. M. (2000). Effect of socioeconomic deprivation on waiting time for cardiac surgery: Retrospective cohort study. *British Medical Journal, 320*(7226), 15–18.

Pepe, M. S., & Mori, M. (1993). Kaplan-Meier, marginal or conditional-probability curves in summarizing competing risks failure time data. *Statistics in Medicine, 12*, 737–751.

Plomp, J., Redekop, W. K., Dekker, F. W., van Geldorp, T. R., Haalebos, M. M., Jambroes, et al. (1999). Death on the waiting list for cardiac surgery in the Netherlands in 1994 and 1995. *Heart, 81*, 593–597.

Ray, A. A., Buth, K. J., Sullivan, J. A., Johnstone, D. E., & Hirsch, G. M. (2001). Waiting for cardiac surgery: Results of a risk-stratified queuing process. *Circulation, 104*, I92–I98.

Reinhardt, U. E. (1998). Quality in consumer-driven health systems. *International Journal for Quality in Health Care, 10*, 385–394.

Rexius, H., Brandrup-Wognsen, G., Oden, A., & Jeppsson, A. (2004). Mortality on the waiting list for coronary artery bypass grafting: Incidence and risk factors. *Annals of Thoracic Surgery, 77*, 769–774.

Romano, P. S., Roos, L. L., & Jollis, J. G. (1993). Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: Differing perspectives. *Journal of Clinical Epidemiology, 46*, 1075–1079.

Seddon, M. E., French, J. K., Amos, D. J., Ramanathan, K., McLaughlin, S. C., & White, H. D. (1999). Waiting times and prioritization for coronary artery bypass surgery in New Zealand. *Heart, 81*, 586–592.

Sobolev, B., Brown, P., & Zelt, D. (2000). Variation in time spent on the waiting list for elective vascular surgery: A case study. *Clinical and Investigative Medicine, 23*, 227–238.

Sobolev, B., & Kuramoto, L. (2008). *Analysis of waiting-time data in health services research*. Berlin: Springer.

Sobolev, B., Mercer, D., Brown, P., FitzGerald, M., Jalink, D., & Shaw, R. (2003). Risk of emergency admission while awaiting elective cholecystectomy. *Canadian Medical Association Journal, 169*, 662–665.

Sobolev, B. G., Fradet, G., Kuramoto, L., & Rogula, B. (2013). The occurrence of adverse events in relation to time after registration for coronary artery bypass surgery: A population-based observational study. *Journal of Cardiothoracic Surgery, 8*, 74.

Sobolev, B. G., Levy, A. R., Kuramoto, L., Hayden, R., & FitzGerald, J. M. (2006a). Do longer delays for coronary artery bypass surgery contribute to preoperative mortality in less urgent patients? *Medical Care, 44*(7), 680–686.

Sobolev, B. G., Levy, A. R., Kuramoto, L., Hayden, R., Brophy, J. M., & FitzGerald, J. M. (2006b). The risk of death associated with delayed coronary artery bypass surgery. *BMC Health Services Research, 6*, 85.

Turnbull, R. G., Taylor, D. C., Hsiang, Y. N., Salvian, A. J., Nanji, S., OHanley G., et al. (2000). Assessment of patient waiting times for vascular surgery. *Canadian Journal of Surgery, 43*, 105–111.

Part III
Demand, Prioritization and Appointments

Chapter 7
Breakthrough Demand–Capacity Management Strategies to Improve Hospital Flow, Safety, and Satisfaction

Linda Kosnik

Abstract Health care facilities are experiencing overcrowding and hospital-wide waits and delays. Potential bottlenecks must be identified and alleviated by matching demand to capacity. The matching of demand to capacity for the services of a hospital is a complex function of multiple variables and queues across the health care system. Achieving successful demand to capacity management requires a large-scale cultural change to support inter and intra department collaboration, which is paramount to the efficient functioning and flow of any health care organization. This chapter outlines the techniques and tools necessary for creating a successful demand to capacity system which is the results of identifying and implementing interventions to critical stressors in the environment. The intent of such approaches is not only to improve flow but also to support health care systems in their goals to become more reliable, safe, and satisfying for patients and providers.

Keywords Flow • Demand to capacity matching • Job satisfaction • Recruitment and retention • Customer satisfaction • Emergency department overcrowding

1 Introduction

Health care facilities are experiencing overcrowding and hospital-wide waits and delays. Potential bottlenecks must be identified and alleviated by matching demand to capacity. The matching of demand to capacity for the services of a hospital is a complex function of multiple variables and queues. Health care units, services, and even professions have evolved in silos focused on meeting their own individual resource, expertise, customer, technology, and demand needs.

L. Kosnik (✉)
Overlook Hospital, 99 Beauvoir Avenue, Summit, NJ 07902, USA
e-mail: Linda.Kosnik@ahsys.org; l.kosnik@verizon.net

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_7, © Springer Science+Business Media New York 2013

These silos are perfectly designed to support units to operate as individual business units striving to maintain autonomy. Therefore, collaborative, inter/intradepartmental matching of demand to capacity is a complex challenge, but one paramount to efficiency and flow across any health care organization.

The future of health care is dependent on the identification of change initiatives that create a synergy between the strategic goals of the health care organization and the professional goals of the care provider to afford safe, effective, individual-oriented patient care. Such change initiatives must be more user-friendly and focused on creating buy-in through staff empowerment, workload management, and job satisfaction.

This chapter outlines the techniques and tools necessary for creating a successful demand to capacity system which is the result of identifying and implementing interventions to critical stressors in the environment. The intent of such approaches is not only to improve flow but also to support health care systems in their goals to become more reliable and safer, and more satisfying for patients and providers. Key system success factors are discussed including real-time communication, inter/intradepartmental and interdisciplinary collaboration, staff empowerment, standardization of best practices, and institutional memory. A dynamic, customizable color-coded demand to capacity management model is discussed, which uses the tools and techniques from crew resource management (CRM) and microsystem thinking.

The deployment of a demand to capacity management system is discussed with the web of work and examples cutting across the entire hospital system.

2 Background

Pressures from regulatory and government agencies, consumer advocate groups, and insurance companies have forced hospitals to focus on cost containment (Holtom and O'Neill 2004). Many consider this cost containment, although necessary, to create a significantly negative impact on employee satisfaction, driving a similar effect on patient care and outcomes. In addition, health care organizations are under pressure to bring on new technologies, processes and procedures to keep, if not increase, market share. This creates a conundrum of expectations for health care organizations, which must now strive to improve efficiency and cost containment while providing effective, safe patient care with maximal stakeholder satisfaction. As a result, health care organizations have implemented system-wide change initiatives under the guise of such terms as restructuring, reengineering, TQM, and CQI.

The work of efficiency and cost containment frequently looks to personnel management as it makes up what is perceived to be the "lion's share" of the controllable expenses. Therefore, such initiatives as "restructuring" have the most impact on the work of nursing as integrated delivery systems are developed, reductions in length of stay are prioritized, and multifunctional workers are used as a solution for staffing issues (Tonges et al. 1998).

The USA is currently experiencing the most significant health care provider shortage in its history. It is expected that this shortage will intensify as needs of health care personnel expand, the current work force continues to age, nurses continue to suffer the highest level of workplace stress of any profession (Laschinger 2004), and universities and colleges struggle to meet rising demand with limited academic resources. With hospitals reporting an average annual nursing turnover rate of 21 % and a cost of replacement to be 150 % of a nurse's annual salary, recruitment and retention have become two of the greatest fiscal challenges facing the health care workplace (Holtom and O'Neill 2004).

2.1 Job Satisfaction, Retention and Workload

The question of what decreases burnout and increases retention and job satisfaction has been widely studied. Empowerment and perceptions of organizational commitment are two factors that have been consistently positively correlated with job satisfaction (Kuokkanen et al. 2003). The concept of organizational commitment is core to retention in that it refers to employee job satisfaction as it applies to one's attachment, trust, and involvement in an organization (Kuokkanen et al. 2003). Empowerment can be divided into two categories: structural and psychological. Structural empowerment is perceived as access to support, supplies, opportunity, and information necessary to do one's job. Psychological empowerment is viewed as autonomy and the meaningfulness of the work. Research suggests that individual behaviors are actually a response to factors in the workplace, making the structural or environmental variables the most critical (Hatcher and Laschinger 1996). The more an individual perceives that he/she is empowered to control and drive his/her work experience, the greater the job satisfaction and the opportunity for the organization to retain that employee. Workers are empowered when they perceive that their environment provides access to power needed to get the job done. This perceived power is related to the individual's ability to access and mobilize support, information, and supplies. Therefore, there is opportunity to increase the sense of empowerment and autonomy by identifying factors that contribute to feelings of powerlessness and designing work environments and workloads that mitigate these factors.

The level of stress perceived to be experienced by a health care provider also directly correlates with job satisfaction.

In health care, stress management is a complex challenge. It is usually not one event that generates a stress response in an individual, unit, service or facility, but more likely multiple events. Different environments have different stressors and the sources of stress may even differ from individual to individual on the same unit. A shortage of resources and supports correlated with increased stress (French et al. 2000). Job satisfaction can provide the balance for stress, particularly in high stress areas such as Emergency Departments (ED) and Operating Rooms (OR). To a degree, stress can be positive and provide motivation or a sense of excitement, challenging an individual to function on a higher level, increasing efficiency, decision making, and effectiveness. This is the stress that drives an individual to function outside their comfort zone creating greater job satisfaction through professional opportunity and growth. On the other hand, stress resulting from frustration and powerlessness to mitigate or implement coping mechanisms in the environment is negative stress. This type of stress is frequently associated with psychological and physiological manifestations as it escalates. Stress is dependent on the individual and the impact perceived from situations and events on one's physical or psychological well-being (French et al. 2000). The factor most frequently correlated with workplace stress is workload (Tonges et al. 1998). Overload occurs when the demand exceeds the individual's ability to access resources or capacity (French et al. 2000). Developing supportive management, increasing opportunities for positive patient interactions, and creating a widespread sense of autonomy and empowerment have the potential to mitigate or diffuse stress (French et al. 2000).

2.2 Workload Management

Changing reimbursement and other economic factors in health care have led to higher patient acuity and hospital restructuring. These two factors have negatively impacted on workload, creating higher patient-to-nurse ratios, which compromises patient care and results in increased patient occurrences, complications, and errors (Aiken et al. 2002).

As physicians spend less and less time in hospitals, the nurse's role has come to include not only surveillance, but health instruction, disease prevention and overall care of the patients and their families. As the role has become increasingly complex, driven by advances in technology, documentation requirements, decreases in LOS, and increased acuity, nurses have found that they have less and less time to actually work with patients. In addition, these job changes have created a greater workload for the nurses who now require a broader spectrum of skills and experiences to be successful. As the sphere of nursing duties expands and the workload increases, nurses have begun to experience time constraints to providing optimal patient care. Workplace stress and overload occur when the job expectations are high and the ability to make decisions and problem-solve are low (Bojtor 2003). Conflicts over time and expanding job expectations and duties have resulted in provider stress and, consequently, compromised patient care. These compromises drive patient and staff dissatisfaction. Eighty-six percent of the nursing population believes that the nursing shortage has left little time for unexpected events or holistic patient care (Bojtor 2003). This is worrisome since caring for patients is the reason many individuals went into health care professions, making this alienation a significant threat to recruitment and retention.

French (French et al. 2000) found that workload was the most significant work environment stressor, with higher levels of stress being associated with lower levels of job satisfaction. Workload issues include inadequate staffing levels, demanding patients, multiple tasks not completed by the previous shifts, insufficient time to complete necessary tasks, and concerns about the quality of care (French et al. 2000). These findings suggest that improvement in the work environment and the way the work is done may lessen the stress nurses experience and increase retention.

Managing resources and improving patient flow through demand to capacity management are not only strategies to manage workload but have become the focus of international concern for improving patient care, quality, and safety. According to the Institute of Medicine (IOM) report, "Crossing the Quality Chasm," US health care should be safe, effective, patient centered, timely, efficient, and equitable (IOM 2001). The Joint Commission's new standard LD.3.15, which went into effect for accreditation review in January 2005, focuses on the importance of identifying and mitigating impediments to efficient patient flow throughout the hospital. It suggests that improved management of processes and the matching of capacity to demand can support the appropriate use of limited resources and, thereby, reduce the risk of negative outcomes to patients from the delays in delivery of care (JCAHO 2004).

The other catalysts for the focus of performance improvement initiatives on resource management and patient flow have been the nursing shortage and what was initially labeled as Emergency Department (ED) Overcrowding. The American College of Emergency Physicians (ACEP) defines (ED) overcrowding as "a situation in which the identified need for emergency services outstrips available resources in the ED. This situation occurs in hospital EDs when there are more patients than staffed ED treatment beds and wait times exceed a reasonable period" (JCAHO 2004). Much of the problem is a result of EDs seeing an ever increasing volume of patients who do not meet the severity of illness that constitutes an emergency. This results from individuals using the services of the ED in lieu of other health care services, and is driven by convenience, lack of insurance or a primary care physician, or because the perception of what constitutes an emergency varies among individuals. Unfortunately, these non-emergent patients seek the services of the ED at peak periods, placing additional burden on the ED at a time when resources may already be overloaded (Siddharthan et al. 1996).

The outcomes of overcrowding include consumers experiencing problems with accessing care, deterioration in the community "safety net" and compromised patient and staff safety. The impact on safety can be measured by treatment delays, higher error rates, increases in mortality and negative clinical outcomes, patients leaving without treatment, and higher readmission rates (JCAHO 2004).

Changes in reimbursement, competitive pricing, managed care, mergers and tightening of government spending have resulted in cost cutting that has shifted patient activity to the front end of a hospital stay where shorter stays can be better managed. As a result, the areas of the hospital that deliver this care, e.g., emergency departments, operating rooms and ICUs, began to experience long waits and delays for services. In response, hospitals that are already functioning in the red have been forced to focus on more efficient and effective operational performance. Many facilities have been driven to maximize resources and function with >90 % occupancy and >95 % productivity. Staffing coverage is often budgeted based on the average daily census (ADC), leaving few units with the resources to flex and maintain identified patient-to-nurse ratios when faced with unpredictable, fluctuating workloads, high vacancy and turnover rates, and high utilization of unplanned compensation time resulting from escalating stress and overload. It is estimated that six out of ten hospitals across the country are operating "at" or "over" capacity (JCAHO 2004).

2.3 Communication, Silos, and Queues

Current health care systems are under increasing stress loads. Current methods and practices of communication within these systems are less than ideal. Tools and strategies to manage flow and demand to capacity matching are not yet well understood. Since 1910, health care has evolved in silos. This resulted from extraordinary advances in health care technology which required each unit to develop unique areas of expertise which drove a sense of isolation or focus on monitoring and mitigating demand and capacity only within the "silo." Historically, communication across units has been challenging. It has been accomplished often by phone rather than face to face, allowing "silos" to continue to function in isolation. Problems arise from departments or services acting independently and considering their own demand and capacity issues without consideration for the upstream and downstream impact on other departments and services. Traditional allocation of resources has resulted in a capacity imbalance in which specific systems and units have over-capacity while others strive to deal with the stress of under-capacity, resulting in bottlenecks or under-utilization (Mango and Shapiro 2001).

From a patient's perspective, these "silos" are all part of the system that makes up their health care experience. Patient flow is defined as the observable process a patient experiences during their health care delivery process (Nacey 2004).

Attempts to solve the overcrowding problem with traditional change models has lead to increased episodes of ED diversion and waits and delays to access services. Significant analyses and process improvement initiatives targeted at EDs has forced health care administrators and consultants to identify overcrowding as actually being a hospital-wide system problem requiring a system-wide approach to managing waits and delays and resources (JCAHO 2004).

Waits and delays in the ED are caused by the inability to access needed resources such as inpatient beds for admissions and lab and x-ray services. Patients waiting in the ED for the next level of care, often an inpatient bed, reduce the functional capacity of the ED, limiting its ability to care for new arrivals. Patients waiting for treatment create a queue, which is often manifested by the use of waiting rooms and hallways for extended periods. Queues occur whenever the current demand for services exceeds the current capacity to provide those services. For the ED, the total service time for patients is the sum of time spent for all medical care and ancillary services provided (Siddharthan et al. 1996). Since patients consistently arrive at an uneven rate, multiple queues are a "constant state" for hospitals, particularly EDs. These queues identify the workload of the unit or service.

Queuing systems are defined by their input and arrival processes. Queuing defines the order in which patients entering a system are served. Use of queuing systems allows for the calculation of the average waiting time, the expected average number of patients waiting and the utilization of servers such as x-ray machines. Information drawn from studying queues supports long-term solutions including the addition of "servers" such as inpatient beds, radiology equipment, and technicians to allow for volume fluctuations to avoid the development of excessive queues. Planning for variabilities in access and matching demand to capacity in real-time are also operational imperatives (Jones et al. 2002). Much of this can be accomplished with some type of logistical communication across departments, which identifies the best use of valuable resources (Mango and Shapiro 2001). The key indicator of the hospital's ability to provide a bed is the admission cycle time. The admission cycle time measures the time from the decision by the physician to admit to the time the patient leaves the ED for an inpatient bed. Resolving extended admission cycle times is an issue that requires collaboration and workload redesign across the facility. It is an excellent opportunity to apply the principles of demand to capacity management.

2.4 Change, Crew Resource Management (CRM), and Microsystem Thinking

System-wide demand to capacity management requires the ability to implement large-scale change. Large-scale change is defined as any change that results in organizational processes and routines being fundamentally altered, facilitating philosophical change in practice. This kind of change is also referred to as discontinuous or transformational as it often causes the organization to deviate from previous approaches. It challenges the organization to reevaluate its vision, mission, identity, values and its strategic plan, at times precipitating a complete change in organizational direction (Narine and Persaud 2003). It should be noted that different cultural groups may require different approaches to communication and rewards to facilitate organizational change and job satisfaction. It is imperative that the large-scale change desired be congruent with the culture of the organization (Narine and Persaud 2003).

The role of leadership in creating a cultural change cannot be undervalued. Leaders must be visible; enforce the desired norms, vision, and values, and encourage others to do the same. They are responsible for assuring that the unit has the resources, skills, and training to achieve the organization's goals (Hawkins and Kratsch 2004).

Ultimately health care collaboration and teamwork are tied to the system's ability to work effectively and efficiently towards optimal patient flow and system outcomes across the health care experience. This is a result of a complex matrix of workload and resource management, staff empowerment, and effective communication targeted at matching demand to capacity in real time.

2.4.1 Crew Resource Management (CRM)

One successful approach for supporting system-wide change, particularly clinical change based on best practices, is Crew Resource Management (CRM). CRM is a communication methodology focusing on team-centered decision-making systems which was developed by the aviation industry in 1979 in response to a NASA workshop that examined the role of human error in air crashes. When CRM is applied to health care, the communication space of health care practitioners caring for critically ill patients can be viewed as resembling that of an aircrew engaged in complex flight operations. Use of team-centered decision-making systems enables teams to perform more efficiently.

Use of the CRM model does not presuppose that adequate communication is enough but instead supports a combination of communication, technology, and process change. CRM's primary building blocks include the use of backup systems: team communication and coordination, adequate briefings, availability and use of resources, leadership and adequate supervision, system knowledge, personal readiness, planning, correction of known problems, and issues and management support (Kosnik 2002).

Historically, effective medical practice depended on a small number of health care providers, which made communication and teamwork requirements simple. Today the health care system is composed of and dependent on many persons, each with unique knowledge and skill sets, which makes routine communication increasingly complex. With the emergence of patient safety, the importance of collaboration and a team approach to patient care has become paramount. Collaboration in providing patient care is more important than preserving an individual provider's professional boundaries or roles (IOM 2001). All members of the health care team must communicate effectively to coordinate care and meet the patient's needs. This expectation correlates with IOM's "New Rule" number ten, which states that, "Clinicians and institutions should actively collaborate and communicate to ensure an appropriate exchange of information and coordination of care (IOM 2001)".

CRM is all about shared knowledge and free flow of information. The model was specifically designed to promote team-based improvement initiatives and collaboration among clinicians for care that is safe and effective. This makes it an excellent tool to promote the behaviors necessary to create buy-in across multiple units and services towards matching demand to capacity.

2.4.2 Microsystem Thinking

Although CRM has been primarily used with teams and units, it has extraordinary potential to create synergy through communication and team building on the system level when coupled with the work of microsystem thinking. Microsystems are defined as the small, functional, frontline units that provide most health care to most people (Nelson et al. 2002). They are the place where patients and providers meet. Each microsystem has its own unique culture and customer population. From an operational perspective, it has clinical and business aims, policies and procedures, and shared information that produce the services and care measurable as performance outcomes. Macrosystems are the larger systems that are made up of microsystems. The emergency department and the radiology department are examples of microsystems. The average health care system is composed of a few basic parts: frontline clinical microsystems, overarching macrosystems, and patient subpopulations needing care (Godfrey et al. 2003). These systems evolve over time and are embedded in larger systems and organizations. As any living and adaptive system, the microsystem must (1) do the work, (2) meet staff needs, and (3) maintain itself as a clinical unit (Godfrey et al. 2003).

The microsystem framework provides practical steps to using microsystem thinking as strategic building blocks. Microsystem thinking makes several organizational assumptions. The first is that bigger systems (macrosystems) are made up of smaller systems (microsystems) which produce quality, safety, and cost outcomes at the front line of care (Nelson et al. 2002). Ultimately, the outcomes of the macrosystem can be no better than the outcomes of the microsystems of which it is composed. If strategically driven and the performance of each individual microsystem is optimized, a systematic transformation can be achieved to meet the organizational goals of the organization and the needs of the frontline care providers. The greater the linkage and collaboration between the different clinical and support microsystems, the more seamless, timely, efficient, safe and reliable will become the operations of the macrosystem (Kosnik and Espinosa 2003).

Microsystem thinking provides the structure and opportunity to drive strategic goals from the point where service is delivered to where the greatest value can be elicited. It is through the activation of the microsystems that there is the free-flow of information which drives the anticipation of needs supported by the collaboration of the providers. It is within the boundaries of the microsystem that the patient has the opportunity to meet with the providers and sculpt a common vision of the care desired and expected through shared decision making.

Integrating the concepts of CRM and Microsystems supports the creation of empowered teams who understand their position and relationship towards partnering with the organization (macrosystem) to achieving their mutual goals. For our purposes, this is the matching of resources and capacity to changing demand in a cost effective, efficient, patient oriented manner. The process of matching demand to capacity identifies the workload of a unit which drives patient flow.

3 Demand to Capacity Management System (DCMS)

Hospitals need to better understand the stress loads on their systems in order to identify, measure and mitigate stress loads. A system used to manage the relationship of demand on a system to the capacity of a system is called a Demand to Capacity Management System (DCMS). Mismatches between capacity and demand have significant consequences, including inpatient services meltdown, blocked patient flow, communication breakdowns, compromised patient safety, customer dissatisfaction, episodes of divert and bypass, and lost revenue.

A successful demand to capacity management system is a result of identifying and implementing interventions to critical stressors in the environment. The goals of a successful DCMS are to provide real-time communication, inter/intra-departmental and inter-disciplinary collaboration, staff empowerment, standardization of best practices, and institutional memory. The success of the work of matching demand to capacity is related to the success of six key factors:

1. Trust making
2. Staff empowerment
3. Collaboration
4. Common vocabulary
5. Mitigation of constraints and barriers
6. Reciprocity

A robust DCMS reduces incidents of overload, which is manifested by episodes of divert and bypass. These episodes are actually a result of inpatient services meltdown. A DCMS creates more reliable and stable systems with less variability because of the "smoothing" effect that can be achieved with monitoring, prevention and mitigation of stress loads and queues, returning control of the system (empowering) to the providers. It creates a synergy between the microsystems (units) and the macrosystems (hospital). Receptor sites become available, there are improved staffing rations and there is ready access to necessary supplies and processes that make the work of the microsystem efficient. The exciting by-products of demand–capacity matching include increased customer satisfaction from reductions in waits and delays, effective recruitment and retention from increased job satisfaction, and new avenues of collaboration and inter/ intradepartmental support resulting from collaboration. This collaboration is a result of the use CRM, the principles of which support the value, input, and empowerment of all the team members. For example, through these principles the importance and input of the environmental personnel to the admission process as they own responsibility for preparing the bed for occupancy are recognized. This recognition becomes a driver for nurturing better communication between environmental services and bed management services with the common goal of improving admission cycle time.

Now let's imagine a system that operationalizes matching demand to capacity (DCMS). This system must empower the providers to measure and mitigate stress loads and queues in real time. There are five basic concepts to developing such a system which we call, grids, categories, criteria, interventions and statuses.

3.1 Categories and Criteria

The categories for a unit or service organize the key criteria or stressors that are central to getting the "work" of the unit done. For this model, we use the categories of census, acuity, other and staffing (CAOS). Census is those criteria which describe what that unit or service "counts" to determine its workload. For example, an inpatient unit would count patients while a respiratory service might count treatments, Acuity criteria determine the level of stress associated with the population, procedure or specimens, which can be measured as workload and/or in time. This category often measures turnaround time. For example, an oncology inpatient unit may count the number of intravenous chemotherapy infusions scheduled while the laboratory may identify criteria around CBC turnaround. The oncology unit is using a number because from experience they know that each infusion takes a specific amount of time and requires a designated amount of resources. The laboratory, on the other hand, knows that CBC results can be routinely expected in a certain amount of time and are the most frequently ordered test. Therefore, for the laboratory, CBC is the test most likely to indicate the workload of the unit, in that an increase in turnaround time is most likely reflective of an increase in volume, a lack of resources or an access to information issue requiring mitigation.

"Other" criteria represent the factors that influence the productivity of the human services capacity such as availability of equipment, systems, and supplies. These criteria tend to center on access to information, particularly information systems downtime and on supplies, such as IV pumps. "Staff" criteria represent the capacity of a unit in terms of labor or human services. It is specific to the status and matching of staff to fluctuating demand and for the mitigation of staffing discrepancies.

All criteria must have valid values that indicate a "call to action." It is also important to limit the number of criteria to only those items that are reoccurring, specific to a significant stressor or representative of a performance improvement initiative that warrants tracking and higher visibility. If a criterion does not drive a response or action then it should not be used for a DCMS.

3.2 Interventions

Interventions to mitigate demand to capacity mismatches are divided into two concepts, my interventions and other interventions. My interventions are those actions that a designated unit can perform for themselves. These actions have been identified by that microsystem to mitigate and deescalate the stress caused by the corresponding criterion. Other interventions are those that require supportive actions or responses from other units, departments or services. These interventions have been identified prior to the event and negotiated collaborative. The microsystem must be able to depend on the identified response from the designated support unit under the specific conditions that triggered the intervention. In order to facilitate buy-in, reciprocal responses should have also been identified to create what can be metaphorically called a “Fair Trade Agreement.” The overarching goal of these interventions is to create a synergy or collaboration across all the microsystems within the macrosystem. By using the principles of CRM all individuals in the microsystem understand their value to the flow and “work” of the microsystem and the macrosystem. In this way, the strategic goals of the macrosystem can be realized and a culture of collaboration and trust can be established. Much of the trust building that occurs is simply a result of mutually sharing issues and working collaborative to identify solutions. It is very difficult to support a rationale to creating barriers or be obstructive towards individuals that you now know and have partnered with. Mutual respect for the workload of other “silos” drives positive relationships and the breakdown of silo thinking towards the use of CRM and microsystem thinking tools.

3.3 Status

The next step to developing a DCMS is to identify levels or statuses for each criterion, and to quantify the queues and the level of stress of the categories (CAOS). One of the most effective models is color coding, as promoted by the national security system—that of green, yellow, orange escalating to red with increased risk. This common vocabulary, derived mostly from everyday words, is the context of new or shared syntax. It encourages interactions and conversations about collaboratively seeking to solve problems and frames all activities as “shared.” The color coding and terminology of the DCMS can become part of a common vocabulary that galvanizes the microsystems to the macrosystems.

Green reflects an optimally functioning system, a state of equilibrium, homeostasis. Demand and capacity are matched. Staff describes it as “a good day.” *Yellow* reflects the status of early triggers, the first indications of demand without readily available capacity, and of developing queues. This is the most important status in that it allows for early intervention before the provider even recognizes escalating stress. This is the opportunity to mitigate—even eliminate—all the stressors that historically we ignore each day. If not mitigated, these stressors escalate in status and are compounded by each additional criterion triggered, ultimately creating work overload. It is usually not one event that generates a stress response in an individual, unit, service or facility, but more likely multiple events (French et al. 2000). Essentially, the goal is to act early before the system realizes it is under stress and while there is opportunity to match demand to capacity and maintain maximal system flow. *Orange* reflects escalating demand without readily available capacity. In this state, aggressive action is required to avoid system

Table 7.1 A generic DCMS grid

Unit name	Criteria	Green	Yellow	Orange	Red
Census		Values/ interventions	Values/ interventions	Values/ interventions	Values/ interventions
Acuity		Values/ interventions	Values/ interventions	Values/ interventions	Values/ interventions
Other		Values/ interventions	Values/ interventions	Values/ interventions	Values/ interventions
Staffing		Values/ interventions	Values/ interventions	Values/ interventions	Values/ interventions

Table 7.2 An example of escalating census criteria and number of patients held for inpatient beds in the ED

Emergency department	Criteria	Green	Yellow	Orange	Red
Census	Patients holding for admission (>1 h)	0–2	3–5	6–10	≥11

overload and ultimate gridlock. *Red* is a state of gridlock and system overload. This critical, high stress, status warrants the use of the organizations Disaster Plan, including such actions as canceling elective surgeries, admissions and procedures. This is the level manifested by significant waits and delays, admissions holding in the ED, customer dissatisfaction, high use of staff, unplanned benefit time, divert/ bypass and loss of revenue and market share.

Categories, criteria, interventions and statuses come together in a unit or service grid. Our current DCMS is made of 44 grids representing the microsystems that currently support each other in identifying and mitigating stressors in real time to avoid demand to capacity mismatches. Table 7.1, shows an example of a generic DCMS grid.

Each criterion is given values that demonstrate escalating stress corresponding to each color or status (Table 7.2). Developing the interventions for each criterion and status is the core to real-time matching of demand to capacity. These interventions create the institutional memory. Institutional memory is those interventions that consistently produce the desired results. They are identified by looking at the behaviors, actions and collaboration demonstrated by the institution's best supervisors, managers and charge staff when challenged by specific constraints and barrier to "getting the work done." By memorializing these "best practices" in DCMS unit grids, less experienced staff can be trained and empowered to utilize these practices successfully.

An example of escalating census criteria for the key criterion of the number of patients holding in the Emergency Department for an inpatient bed is displayed in Table 7.2.

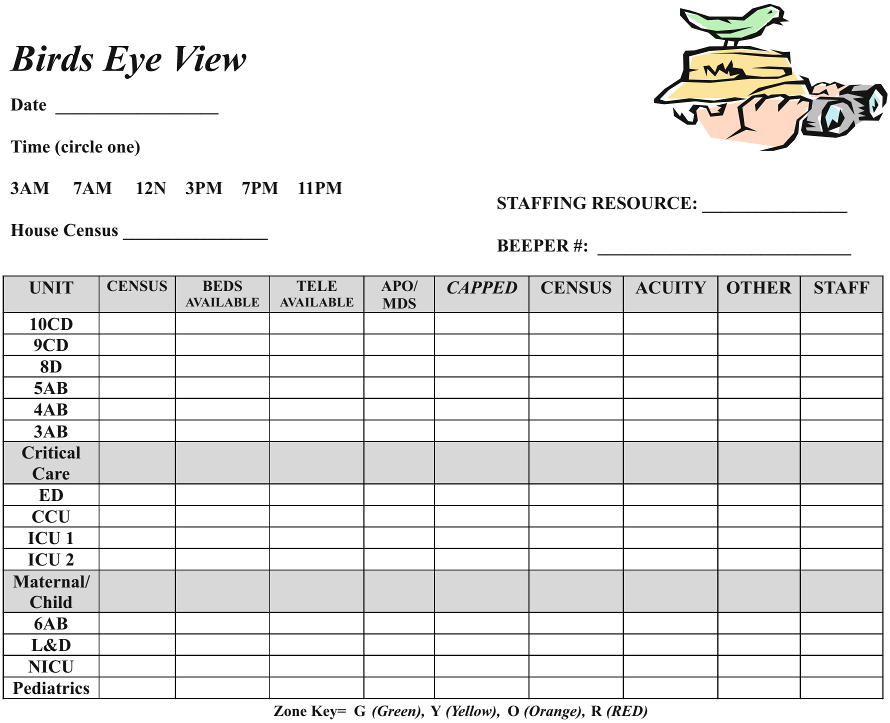
4 What You Don’t Measure Is Hard to Improve!

The use of real-time data is necessary to demand to capacity management because it is real-time feedback that eliminates the use of intuition and supports the allocation of resources to the right place, at the right time in the right amount. By identifying queues the work of mitigating bottlenecks before they impact patient flow can be achieved.

Virtual instrumentation has been applied to almost every industry including telecommunication, automotives, semiconductors, factory management and essentially any industrial operations program. Still relatively new in health care, programs have been developed and implemented in a wide range of research-based clinical applications and executive information tools particularly as related to financial management (Rosow et al. 2003). Unfortunately, health care has been slow to accept technological solutions and most institutions continue to manage complex processes such as patient flow using tradition methods such as paper, white boards and phone calls. Current “intuition” forecasting by managers is less than adequate, resulting in last minute adjustments to elective schedules, staffing and resources (Jones et al. 2002). These approaches lack the timely information necessary to match resources to changing patient needs. The result is the exacerbation of hospital-wide waits and delays, loss of admissions, decreases in revenue, provider job dissatisfaction, and an increase in all forms of health care resource wastes (Rosow et al. 2003). In addition, it creates real-time barriers to patient access and flow as hospitals struggle to deal with unprecedented increases in the demand for services. User-defined, customizable solutions facilitate decision making from the big-picture to transaction-level detail, while providing real-time knowledge, information, access and resources that can empower all levels of the organization (Rosow et al. 2003).

Harper (2002) demonstrated the effectiveness of an identified generic framework for modeling hospital resources. This framework was dynamic in that it used real-time data, patient flow and time-dependent demand profiles to support managerial decision making. The need for sophisticated, real-time tools is imperative to accurately reflect the complexity, uncertainty, variability and limited resources that drive the acute care facility’s ability to respond to fluctuating demand. The ability to access data regarding a system is the foundation to the developing of “best practices” for the matching of demand to capacity. Best practices result from the analysis of the data over time and the development of management behaviors that predict the practices and processes support optimal outcomes.

Much of the success of our DCMS resulted from a paper system initiated in 1997. This system used a report sheet called the “Bird’s Eye View” (Fig. 7.1) which was faxed to all participating units twice daily, with updates as needed. Since 2003 our system has been Web based and increasingly interactive to facilitate timely completion of interventions. This system called Acute Care Operations Management Systems (ACOMS) is a collaborative effort between Atlantic Health System and Vistaar Technologies, Inc. It is suggested that a system for driving and measuring the DCMS in real time be considered, particularly for improving operational efficiency.



Birds Eye View

Date ________________

Time (circle one)

3AM 7AM 12N 3PM 7PM 11PM

House Census _______________

STAFFING RESOURCE: ______________

BEEPER #: __________________________

UNIT	CENSUS	BEDS AVAILABLE	TELE AVAILABLE	APO/ MDS	*CAPPED*	CENSUS	ACUITY	OTHER	STAFF
10CD									
9CD									
8D									
5AB									
4AB									
3AB									
Critical Care									
ED									
CCU									
ICU 1									
ICU 2									
Maternal/ Child									
6AB									
L&D									
NICU									
Pediatrics									

Zone Key= G *(Green),* Y *(Yellow),* O *(Orange),* R *(RED)*

Fig. 7.1 Bird's eye view

4.1 *Our Results*

A robust DCMS, particularly one that is able to generate real-time and retrospective trended data, can provide the impetus necessary for buy-in, particularly from administration and physicians. We have been fortunate to have greater than 92 % retention, significantly decreased episodes of full divert (Fig. 7.2), and support for supplies and resources needed for matching demand to capacity across the macrosystem.

There are commonalities that drive collaboration and the DCMS across the continuum of care. Some units and departments are natural partners for each other, for example environmental services is the perfect partner for food services, volunteer services for transport. There also appears to be a hierarchy of impact with the first service to consistently become stressed being transport or the ability to deliver or return patients to and from services and procedures. It is also known that one service allowed to escalate in stress level and status will quickly cause the escalation of other services and units. For example, an increased turnaround to access laboratory results will delay diagnosis and treatment in the ED. The ability of services to provide support significantly decreases with the number of services in orange or red.

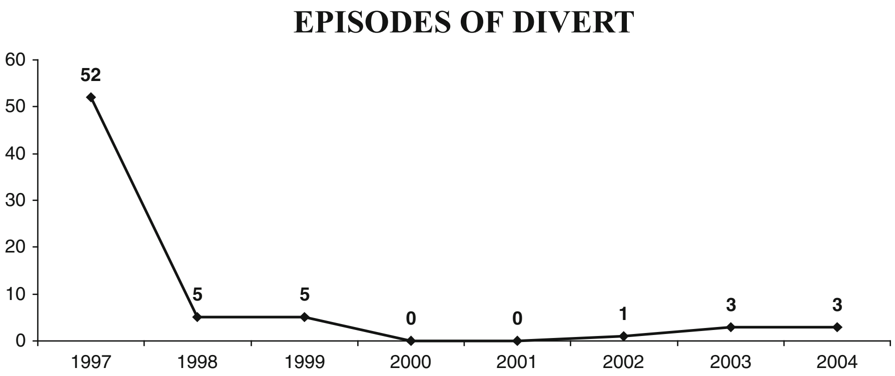


Fig. 7.2 Episodes of divert since the initiation of the DCMS in 1997

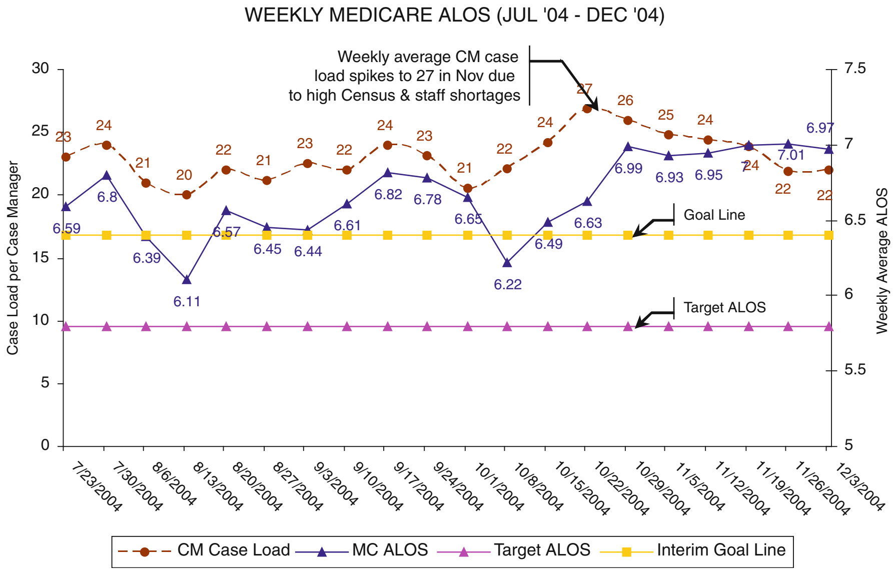


Fig. 7.3 Graphs and analysis provided by Vikas Phatak and Anik Roy. Base Demand & Capacity Data Captured by the ACOMS Software of Vistaar Technologies Inc. (WWW.Acoms.Vistaar.Corn)

Some of the new programs initiatives that we have realized from demand capacity management include; developing bridge orders for admission from the ED, opening of additional intermediate beds to decompress ICU, additional FTEs for Case Management (Fig. 7.3), additional transport staff, expedited ICU admissions, and additional ED staff for triage. ACOMS has also increased our ability to access capital dollars for purchasing equipment that had been found to negatively impact our ability to match demand to capacity including; IV pumps, PCA pumps and mini-infusers. Some of the most valuable results were those that improved communication across the system such as wireless phones and report sheets between the recovery rooms and the surgical units which were expanded to include components that improved antibiotic therapy (SIPS) compliance and pain management.

5 Discussion

Much of the research done on health care restructuring suggest that while workload may increase during restructuring there are other factors that may actually drive job dissatisfaction, such as diminished resources with increasing demand for services. Distress may, in fact, be the result of the amount of work expected or the way restructuring changes were implemented. Change initiatives that include placing staff nurses on multidisciplinary, multi-departmental task forces and committees and providing access to education and learning resources, offer growth and promotional opportunities that support empowerment and job satisfaction (Laschinger 2004). Better support, communication, resources and supervision during restructuring may be the key to increasing satisfaction with the change process (Burke 2003).

The work of Aiken (Aiken et al. 2002) demonstrated substantial difference in patient mortality and nursing job satisfaction and burnout related to patient-to nurse ratios. This work did not, however, identify the workload expectations of nursing pre and post mandated ratios. Many hospitals have responded to mandated patient-to-nurse ratios by eliminating supportive resources and staff and creating a more complex workload for nurses. If, instead, their work was redesigned to optimize resources and productivity, would there be a similar impact on job satisfaction?

Additional research should be done to quantify the impact of mental and physical workloads on job satisfaction, taking into consideration such factors as fluctuations in census and patient visits, staffing and nurse-patient ratios, ancillary support, particularly transport and secretarial support, patient acuity, and access to supplies and processes.

6 Conclusions

Efficiency demands on hospitals, driven by high costs and reimbursement and malpractice issues, have forced health care administrators to push operations to function at close to maximum capacity, resulting in a lack of resources, including staff, equipment and support services, particularly for accommodating unexpected surges in volume (Mango and Shapiro 2001). These organizational pressures coupled with increased stress, workload and empowerment have been the impetus for widespread job dissatisfaction created the most significant health care provider shortage in history.

Nursing staffing has been cited as being the primary driver of high quality hospital care and optimal patient outcomes. Staffing shortages have resulted in higher workloads with hospital nursing leading the country with a 40 % burnout rate (Aiken et al. 2002).

The perceived level of stress experienced by a health care provider also directly correlates with job satisfaction drives burnout. Core to health care provider job satisfaction are empowerment and organizational commitment. Workload has been identified as a major source of workplace stressors.

Key to operational performance and workload management is identification of potential bottlenecks and development of an action plan that allows for mitigation or smoothing of demand and capacity mismatches in real-time. This involves identifying the varying demand for services at any one time and matching it to the necessary resources efficiently and effectively. The random fluctuations of demand can be analyzed to identify real patterns that can be managed. Limits can be placed on services or practices to facilitate process improvement. The most successful strategies empower the frontline worker to handle matching demand to capacity in real time and thus prevent bottlenecks, queues and stress from occurring.

Information technology is just beginning to be accepted as a strategy for managing workload and matching demand to capacity across the continuum in real-time. Greater use of technology has the potential to support the efficient use of human resources, which are a valuable commodity. There is opportunity to use a real-time demand to capacity matching system (DCMS), to empower staff to manage their workload towards optimizing patient flow, satisfaction and outcomes and to increase staff job satisfaction. A successful DCMS is dependent six key principles, trust making, staff, empowerment, mitigation of constraints and barriers to flow between microsystems, development of a common vocabulary, collaboration between microsystems ideally based on the principles of CRM, understanding of the value of the individuals within the microsystem not only within the microsystem but the macrosystem and development of a sense and support of reciprocal or "fair trade."

Ultimately it is the elimination of waits and delays that will truly distinguish a hospital (Mango and Shapiro 2001). But it is the retention of our valuable human resources that will make it possible.

The right resources . . . in the right place . . . at the right time!

References

Aiken, L. H., Clarke, S. P., Sloane, D. M., Sochalski, J., & Silber, J. H. (2002). Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *JAMA, 288*(16), 1987–1993.

Bojtor, A. (2003). The importance of social and cultural factors to nursing status. *International Journal of Nursing Practice, 9*, 328–335.

Burke, R. J. (2003). Nursing staff attitudes following restructuring: the role of perceived organizational support, restructuring processes and stressor. *International Journal of Sociology and Social Policy, 23*(8/9), 129–157.

French, S. E., Lenton, R., Walters, V., & Eyles, J. (2000). An empirical evaluation of an expanded nursing stress scale. *Journal of Nursing Measurement, 8*(2), 161–178.

Godfrey, M. M., Nelson, E. C., Wasson, J. H., Mohr, J. J., & Batalden, P. B. (2003). Microsystems in healthcare: Part 3. Planning patient-centered services. *The Joint Commission Journal for Quality Improvement, 29*(4), 159–170.

Harper, P. R. (2002). A framework for operational modeling of hospital resources. *Healthcare Management Science, 5*, 165–173.

Hatcher, S., & Laschinger, H. K. S. (1996). Staff nurses' perceptions of job empowerment and level of burnout: A test of Kanter's theory of structural power in organizations. *CJONA, 9*(4), 74–94.

Hawkins, A. L., & Kratsch, L. S. (2004). Troubled units: Creating change. *AACN Clinical Issues, 15*(2), 215–221.

Holtom, B. C., & O'Neill, B. S. (2004). Job embeddedness: a theoretical foundation for developing a comprehensive nurse retention plan. *JONA, 34*(5), 216–227.

Institute of Medicine. (2001). *Crossing the quality chasm*. Washington, DC: National Academy Press.

Joint Commission. (2004). *Joint commission resources: managing patient flow*. Oakbrook Terrace, IL: Joint Commission on Accreditation of Healthcare Organizations.

Jones, S. A., Joy, M. P., & Pearson, J. (2002). Forecasting demand of emergency care. *Healthcare Management Science, 5*(4), 297–305.

Kosnik, L. K. (2002). The new paradigm of crew resource management: just what is needed to reengage the stalled collaborative movement? *Journal on Quality Improvement, 28*(5), 235–241.

Kosnik, L. K., & Espinosa, J. A. (2003). Microsystems in healthcare: Part 7. The microsystem as a platform for merging strategic planning and operations. *Journal on Quality Improvement, 29* (9), 452–459.

Kuokkanen, L., Leino-Kilpi, H., & Katajisto, J. (2003). Nurse empowerment, job-related satisfaction, and organizational commitment. *Journal of Nursing Care Quality, 18*(3), 184–192.

Laschinger, H. K. S (2004). Hospital nurses perceptions of respect and organizational justice. *JONA, 34*, 354–364.

Mango, P. D., & Shapiro, L. A. (2001). *The McKinsey quarterly: Number 2. Hospitals get serious about operations (pp. 74–85)*. Pittsburgh, PA: McKinsey & Company.

Nacey, G. E. (2004). *Maximizing hospital capacity*. Pittsburgh, PA: Tele-Tracking Technologies, Inc.

Narine, L., & Persaud, D. D. (2003). Gaining and maintaining commitment to large-scale change in healthcare organizations. *Health Services Management Research, 16*, 179–187.

Nelson, E. C., Batalden, P. B., Huber, T. P., Mohr, J. J., Godfrey, M. M., Headric, L. A., & Wasson, J. H. (2002). Microsystems in health care: Part 1. Learning from high-performing front-line clinical units. *The Joint Commission Journal for Quality Improvement, 28*(9), 472–493.

Rosow, E., Adam, J., Coulombe, K., Race, K., & Anderson, R. (2003). Virtual instrumentation and real-time executive dashboards. *Nursing Administration Quarterly, 27*(1), 58–76.

Siddharthan, K., Jones, W. J., & Johnson, J. A. (1996). A priority queuing model to reduce waiting times in emergency care. *International Journal of Healthcare Quality, 9*(5), 10–16.

Tonges, M. C., Rothstein, H., & Carter, H. K. (1998). Sources of satisfaction in hospital nursing practice. *JONA, 28*(5), 47–61.

Chapter 8
Managing Patient Appointments in Primary Care

Sergei Savin

Abstract In recent years, many US health care establishments have found themselves under increasing pressure to improve the cost-effectiveness of their operations in the face of tight competition, while maintaining high standards of care. Finding the best trade-off between these competing objectives is not an easy task, since the efforts to keep costs under control often result in overutilization of existing resources and, as a consequence, increased patient delays. The Institute of Medicine report *Crossing the Quality Chasm: A New Health System for the 21st Century* identifies "timeliness" as one of six goals that should drive the redesign of the health care delivery system in the coming years. For many patients, primary care is one of the most important settings for their contact with the health care system. In this setting, an appointment mechanism directly determines "timeliness" of received care. In this chapter we introduce a simple model that describes the evolution of appointment backlogs in a primary office setting and describe how expected value and the variance of the daily demand for appointments influences backlog buildup. We also discuss the popular practice of advanced access, a newly proposed approach for reducing and eliminating appointment delays. In particular, we develop a set of guidelines that any primary care office should use in determining the patient panel size to support the advanced access approach. Our guidelines are illustrated through a set of examples based on the demand and supply data taken from the surveys of the American Academy of Family Practice as well as 2002 National Ambulatory Medical Care Survey.

Keywords Access to care • Appointment scheduling • Advanced access • Optimal patient panel size

S. Savin (✉)
Graduate School of Business, Columbia University, 404 Uris Hall, New York, NY 10027, USA
e-mail: svs30@columbia.edu

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_8, © Springer Science+Business Media New York 2013

1 Introduction: Delays in Access to Primary Care

Primary care is the key part of any national health care system—a focal point of contact between an individual and health care professionals. The World Health Care Organization, in the Declaration of its International Conference on the Primary Health Care (1978), states that the primary care networks should ensure that "practical, scientifically sound and socially acceptable methods and technology" are "universally accessible to individuals and families. . .at a cost that the community and country can afford to maintain. . ." A body of evidence based on cross-country comparisons reported in Starfield (1991) suggests that strong primary care results in higher patient satisfaction scores, lower health care expenses, and fewer drug prescriptions. American Academy of Family Physicians (AAFP, www.aafp.org) defines primary care as "care provided by physicians specifically trained for and skilled in comprehensive first contact and continuing care for persons with any undiagnosed sign, symptom, or health concern (the "undifferentiated" patient) not limited by problem origin (biological, behavioral, or social), organ system, or diagnosis."

Timely access to care is, simultaneously, a key characteristic of an effective primary care system, and a well-documented problem area for the US primary care system. Difficulty in getting an appointment to see a physician in a timely manner is a widespread phenomenon. In its recent milestone report on the quality of health care, *Crossing the Quality Chasm: A New Health System for the 21st Century*, the Institute of Medicine (IOM 2001) identifies "the long waits for appointments which are common today" as a major factor increasing the probability of dangerous diagnosing delays and, consequently, "more advanced diagnosis." The same report defines "timeliness" as one of six characteristics that a working health care system should possess.

Timeliness is recognized as an important characteristic of service delivery in any service setting, including health care. However, it is hard to find any other business setting where the constraints on access to service are as persistent as they are in health care services. The IOM in *Crossing the Quality Chasm* characterizes prevalent practice as follows: "Lack of timeliness also signals a lack of attention to flow and a lack of respect for the patient that are not tolerated in consumer-centered systems in other service industries. It suggests that care has not been designed with the welfare of the patient at the center."

Patient delays in the primary care setting can be classified into two categories: *appointment* and *real-time*. Appointment delay can be defined as the number of days between the requested and scheduled appointment dates. Clearly, not every appointment scheduled for some future date necessarily indicates the lack of access to care: in some cases, patient may find it desirable (due to some previous commitments) not to see a physician "today" (Murray and Tantau 2000 estimate that the fraction of such patients in the total patient pool does not exceed 25 %). Such cases become a part of patient flow often characterized as a "good backlog" (another component of "good backlog" is formed by the appointments—such as follow-ups—which are prescheduled for a specific future date). On the other hand, when a patient is unable to obtain an appointment with her PCP on the day she selects, her access to service is clearly limited. The patient's service request has to be postponed and is forced to join the so-called "bad backlog." While the appointment delays can be as long as weeks or even months, the real-time delays that relate to the wait beyond the prespecified appointment time on the day of actual service, are measured in minutes and hours. In addition, real-time delays can affect both the patient and the health-care provider.

The reasons behind these two types of service delays can be quite different. Appointment delays may be indicative of an overall strategic mismanagement of the demand–supply balance in the primary care environment. Such mismanagement could stem from a primary practice's excessively large patient panel to ineffective approaches for allocating patient demand to appointment slots. In this respect, appointment delays can be virtually eliminated by ensuring that the overall demand for care meets adequate supply of capacity, and that both are actively managed to deal with unexpected short-term mismatches. The real-time delays, on the other hand, have a more tactical nature and are often a result of a complex combination of general service inefficiency, patient/provider lateness, potential mismatch between the average duration of an office consultation and the length of an appointment slot, and finally, due to sheer unpredictability of patient arrivals for their appointments and of actual consultation times. While there exist specific recipes aimed at minimizing the impact of most of these factors, the real-time delays cannot be entirely avoided due to the random, unpredictable nature of primary care service durations.

The focus of the majority of the operations research studies modeling delays in health care systems has been on the minimization of real-time delays (a comprehensive review of this literature stream is given in Cayirli and Veral 2003). At the same time, a more strategic (and, as one can argue, more practically important) task of appointment delay reduction has not received nearly as much attention. In this chapter we attempt to address this imbalance by focusing on the subject of appointment delays. In Sect. 2 we introduce a simple model that explains the role of patient demand characteristics in the creation and growth of appointment backlogs in primary care. Section 3 provides a detailed description of the advanced access, a recently developed approach to backlog elimination. In that section, our focus is on determining a patient panel size that is consistent with the practice of advanced access. We develop a set of practical guidelines that primary care offices can use to set their panel size targets and provide a number of examples based on the data reported in AAFP surveys and in the latest 2002 National Ambulatory Medical Care Survey.

2 Appointment Backlogs in Primary Care: How Are They Created

Long appointment backlogs in primary care are often a norm. In primary care facilities that employ traditional appointment systems, a patient with a “nonurgent” request may have to wait weeks and even months to be seen by a physician: 2004 National Healthcare Quality Report states that the fraction of patients who get timely appointments is only 43.8 % for routine care and 57.3 % for injury/illness related care (www.qualitytools.ahrq.gov/qualityreport/browse/browse.aspx?id=5080). Physicians’ schedules are almost always full, and the truly urgent cases are either diverted to urgent care centers or emergency rooms, or attended to by double-booking, delaying other patients and working overtime. Patients perceive the appointment capacity as being strictly rationed and employ various tactics trying to “game the system” and gain an early access to care. At the same time, as an appointment backlog grows so does the probability of cancellations for those appointments scheduled far in advance of actual service. The overall level of tension and frustration on the part of all participants in such service system can be quite high: providers work long hours, but the backlogs remain stable, and patients still cannot get care when they want to. All of these signs seem to indicate that the demand for care exceeds the supply of appointment capacity. In fact, physicians may use these dysfunctional dynamics to justify their discomfort about adopting any changes threatening the existing appointment system, which is viewed as a wall protecting physician time against an overwhelming flood of care requests (Gordon and Chin 2004).

As our example below will demonstrate, the “obvious” conclusion about the demand–capacity imbalance may be incorrect: if the appointment backlogs remain stable (albeit long) and the actual daily patient demand for appointments is uncertain, then the average daily demand for care is likely to be *lower* than the available daily appointment capacity. The uncertainty of the daily appointment demand is the key to understanding this seemingly paradoxical statement.

It is often convenient to describe uncertain demand for appointments in terms of its average daily value (sometimes also called expected or mean value) and the demand uncertainty as measured, for example, by the value of the standard deviation of the actual demand around its average value. Given that the daily demand for appointments is uncertain, two general rules apply to a typical primary care setting:

1. A backlog of unserved appointment demand can build up even if the average daily demand for appointments is *less* than the available appointment capacity.
2. The value of the appointment backlog grows not only with the average value of daily appointment demand, but also with the value of its standard deviation.

Below we present simple examples that illustrate these rules. Consider a newly opened primary care office with daily appointment capacity of $C = 20$ slots and the average daily demand for appointments equal to 19. We assume that such an office starts its operation on some “day 1” with an empty appointment book, i.e., that all appointment slots on day 1 and on any other day in the future are open. For simplicity we assume that all patients requesting appointments will accept same-day appointments if presented with such an option (our conclusions will also remain valid in the case when some patients insist on later appointments). Such an office has enough capacity to serve the average number of daily appointment requests, so if there were no variation in daily demand the office would have no patient backlog. However, consider a situation in which the actual demand for appointments on any given day is uncertain and takes on two possible values, 16 or 22, each with 50 % probability (we also assume that demands for appointments on different days are independent of each other). In this case, the average demand for appointments is 0.5 × 16 + 0.5 × 22 = 19, but on each particular day the demand is either four below the available capacity or two above it. Demand uncertainty can be characterized by a standard deviation, which is computed as follows for the example:

$$\sqrt{0.5 \times (22-19)^2 + 0.5 \times (16-19)^2} = 3.$$

Suppose that on days when the demand for appointments exceeds the appointment capacity the “overflow” (appointment requests in excess of capacity) is pushed to the next available day, adding to “bad backlog.” Figure 8.1 traces the possible values of patient backlog in such practice on days 1, 2, and 3 (numbers in the ovals show the actual appointment backlog values under each particular scenario): for example, at the end of day 1, the appointment backlog is either 2 (if demand for appointments on day 1 happens to be 22) or 0 (if demand for appointments on day 1 happens to be 16). Since each of these scenarios happens with 50 % probability, the average appointment backlog at the end of day 1 is 0.5 × 2 + 0.5 × 0 = 1. As we can see, when the office starts with “empty” appointment books, the average backlog in the system *grows* from one at the end of day 1–1.75 at the end of day 3, despite the fact that average demand is *lower* than appointment capacity.

Will this backlog continue to grow indefinitely? Certainly not! Imagine that at the beginning of day X, the backlog of appointments happens to be equal to 10. Figure 8.2 shows all possible changes to such backlog over the period of next 3 days: the average backlog drops to 9 at the end of day X, and then continues to drop to 7.25 at the end of day $X + 2$.

So far we have observed that the backlog grows when it is “too low” and drops when it is “too high.” Perhaps, then, there should be a “medium” backlog level that remains *stable* over long period of time. In Fig. 8.3 we extend the timeline depicted in Figs. 8.1 and 8.2 and use simulation to show what happens to the average appointment backlog over a period of 40 days when the office starts with no backlog or with a backlog of 10 appointments. In both cases, the average backlog converges to the same value—approximately 3.25 appointments—the value that remains stable over a long period of time (we will call this value *long-term backlog*). Intuitively, it is not hard to rationalize why such long-term backlog builds up in a

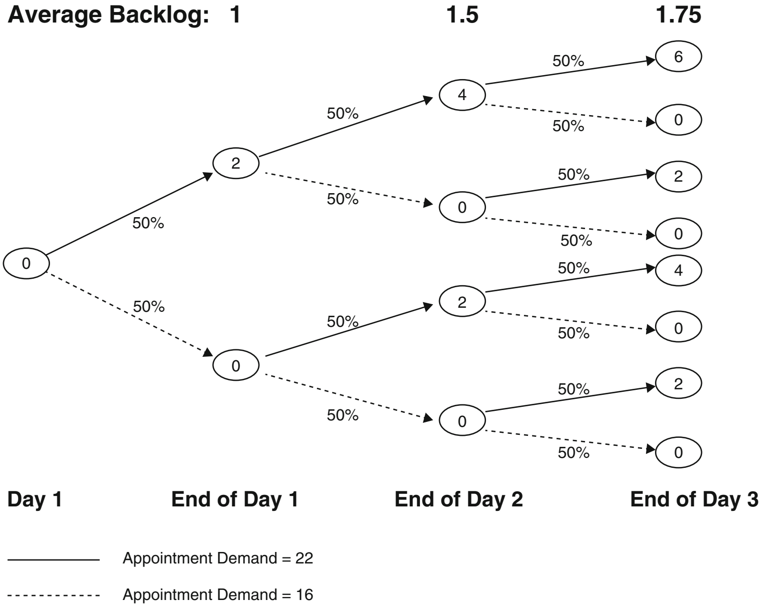


Fig. 8.1 The appointment backlog scenarios when office starts with backlog $= 0$

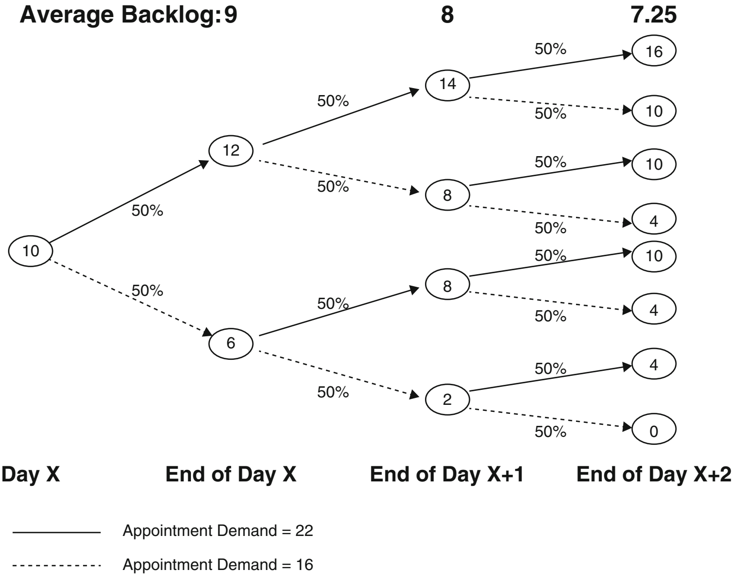


Fig. 8.2 The appointment backlog scenarios when office starts with backlog $= 0$

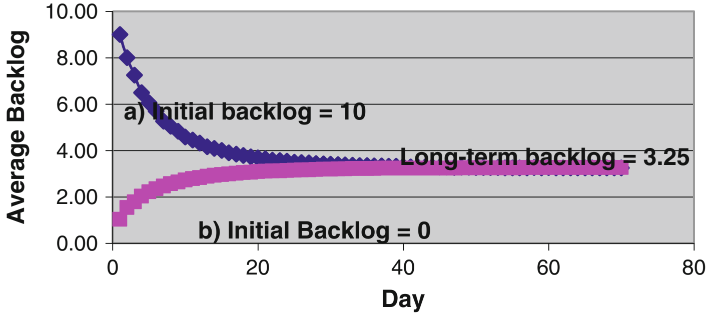


Fig. 8.3 Average backlog as a function of time: (*a*) initial backlog = 10, (*b*) initial backlog = 0. Average demand = 19, standard deviation = 3

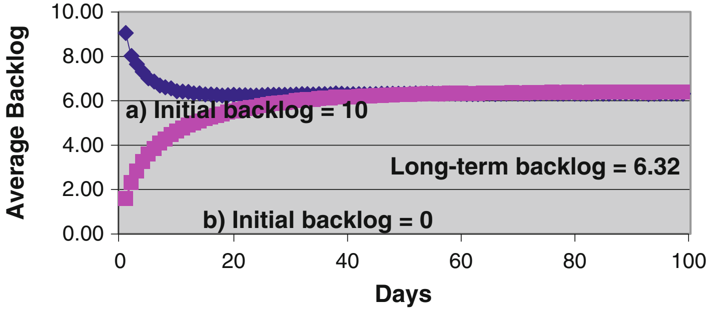


Fig. 8.4 Average backlog as a function of time: (*a*) initial backlog = 10, (*b*) initial backlog = 0. Average demand = 19, standard deviation = 4

system in which average daily demand is below the capacity, even if such system starts "empty." On the "good" days, when the patient demand is less than the appointment capacity, the extra service capacity cannot be "stored" and "transferred" to the next day to serve future patient demand, such extra capacity is simply lost, unless there is a backlog. On the other hand, on the "bad" days, when patient demand exceeds service capacity, the unserved demand is not lost and has to be satisfied in the future. So, if the system starts with no appointment backlog, the "good" days fail to clear the backlog created by the equal number of "bad" days.

The above example demonstrates the role of demand uncertainty in creating appointment backlogs. The greater is that uncertainty, the higher is the resulting long-term backlog. For contrast, Fig. 8.4 evaluates a case with larger variation in demand. In the example demand takes two values, 15 and 23, with equal probability. The average daily demand is still 19 (0.5 × 15 + 0.5 × 23), but the demand uncertainty, as measured by the standard deviation, is $\sqrt{0.5 \times (23-19)^2 + 0.5 \times (15-19)^2} = 4$, which is higher than in the previous example. Correspondingly, the long-term appointment backlog value grows to about 6.3 (as compared to 3.2 for the standard deviation of 3).

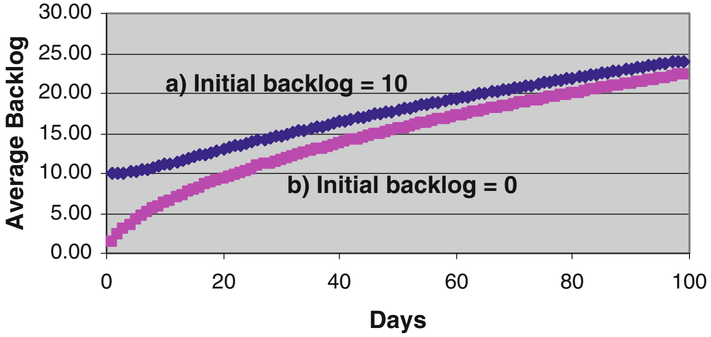


Fig. 8.5 Average backlog as a function of time: (*a*) initial backlog = 10, (*b*) initial backlog = 0. Average demand = 20, standard deviation = 3

As these examples indicate, the uncertainty in the daily demand for appointments (as measured, for example, by the standard deviation of daily demand around its average value) often creates a long-term appointment backlog, even in situations where the average daily demand is below the available appointment capacity. For a given value of the average demand, the larger the demand uncertainty, the longer is the resulting long-term backlog. An increase in the average daily appointment demand also leads to an increase in backlog. For instance, when the average value of uncertain demand is *equal to or larger than* the available appointment capacity, the appointment backlog continues to grow without limit. Figure 8.5 illustrates this fact by showing how the appointment backlog increases over the period of 100 days in cases when the initial backlog is equal to (a) 10 appointments, and (b) 0 appointments. Note the dramatic difference between this figure and Fig. 8.3 (where the average demand was strictly below the appointment capacity): now the backlog grows without bound, no matter how small the initial value, exhibiting no signs of "converging" to any limit.

The examples illustrate that appointment backlogs can occur in primary care practices even when appointment capacity is sufficient to serve average daily demand. However, appointment flexibility can eliminate this problem, as illustrated in the following section.

3 Eliminating Appointment Delays Through Advanced Access

In a seminal paper, Murray and Tantau (2000) describe the "advanced access," a new system for handling appointments in primary care. At the heart of the "advanced access" approach is the patient-centric goal of ensuring that each patient can be seen by his or her PCP on the day of patient's choice, even if this choice is "today." A transition from a traditional backlog-ridden system to the advanced access appointment handling requires a set of steps that adjust the way a primary care office matches patient demand with appointment capacity. While the details of the implementation of the advanced access may differ from practice to practice, the general transition plan includes the following steps: estimating daily appointment demand, adjusting patient panel size, instituting demand control techniques, working down the appointment backlog, and post-introduction management. Below, we provide a detailed review of each of these steps and focus on developing an analytical model that connects the choice of the panel size and the frequency of the overtime work required to sustain advanced access.

3.1 Estimating Demand for Appointments

Demand classification schemes play an important role in shaping up the structure of appointment systems. The traditional appointment scheduling approach is greatly influenced by a classification scheme where all patient care requests go through a triage system that sorts them according to the perceived urgency of the request. “Urgent” (or “same day”) requests are given priority and are often bumped-in on top of an already busy appointment schedule, while “nonurgent” requests are offered a future appointment slot. A commonly used alternative to this approach is the “carve out” appointment model, which explicitly reserves a fraction of daily appointments for urgent requests. A focus on urgency of care in these appointment systems is understandable. At the same time, a triage system is not perfect—and in many cases may disadvantage “stoic sick” (Murray and Berwick 2003) while yielding to the demands of “worried well.” As a result, some of the demand that can and should be dealt without much of a delay is pushed into the future, giving rise to long appointment backlogs.

The advanced access approach, on the other hand, reduces the role of triage by eliminating, for appointment purposes, the distinction between urgent and nonurgent cases. On any day, every patient requesting an appointment is offered a “same-day” option, irrespective of how urgent the demand is or when a patient prefers to be seen (on the same day or on some future day). Thus, under the advanced access approach, patients are encouraged to see their PCP as soon as possible, and the classification of daily demand for appointments is not based on the urgency of an appointment request, but rather on when a patient wants to be seen by a physician. In particular, daily demand for appointments is usually split into four components: “same-day,” “another-day,” “walk-in,” and “follow-up.”

The “same-day” component consists of patients who accept the offer of a same-day appointment. It is likely that as the “advanced access” mentality sets in, this group of patients will constitute a strong majority. On the other hand, for some patients it may be impossible to adjust their schedules and visit their PCP on the same day. While the number of patients declining the same day offer can be expected to diminish with time, it is likely that even in the long run there will always be a fraction of patients who would prefer to be seen some time in the future rather than on the same day (this fraction can be as high as 25 % according to Murray and Tantau 2000). These patients form an "another-day" component of the daily demand. "Walk-in" group consists of patients who "drop by" without contacting their primary care office in advance, while "follow-up" group is made up of patients who are seen on the *that day* and who require a follow-up appointment. Thus, daily demand for appointments is the total number of appointment requests *appearing* on a particular day, regardless of whether they are serviced on that same day or in the future. One important point to keep in mind is that the demand for appointments *cannot* be accurately estimated using past appointment data. Instead, a primary care office has to record all appointment requests as they happen for several weeks in order to establish repetitive daily demand patterns.

3.2 Finding the Right Panel Size

The process of establishing the right panel size for a particular primary care practice should proceed through the following five steps: (1) defining the current panel size, (2) estimating daily rate of appointment requests, (3) establishing the target number of daily appointment slots, (4) setting the target overflow frequency, and (5) computing the appropriate panel size based on the overflow frequency trade-offs. Below we provide detailed guidelines on implementing each of these steps.

3.2.1 Defining Current Panel Size

The estimation of patient panel size N in a managed care environment is easy: panel size is defined as a number of patients enrolled with a physician. On the other hand, in fee-for-service or mixed practices, the number of patients "on file" may be misleading since it is not uncommon to preserve files for patients who may no longer be using the practice's services. We suggest that in such an environment the panel size be estimated as a total number of distinct patients seen by a physician in the last 18 months (counting patients who visited practice over the last year may underestimate the effective panel size, while the 2-year count typically produces an overestimated value). For example, if the number of patients who visited a physician over the last 18 months is 2,500, we can use this number as an estimate of current panel size, N.

3.2.2 Estimating Daily Rate of Appointment Requests

Daily patient demand for care is based on the profile of the population served by the practice as well as the nature of the practice itself. To arrive at the most accurate assessment of total demand requires prospective measurement of the specific appointment dates that patients actually ask for including walk-ins (external demand) as well as the follow-up visit dates physicians request. This is accomplished by examining appointment logs for the last several months (it is best to look at the period of at least a year to capture all seasonal effects—we recommend 18 months, which is approximately equal to $7 = 315$ working days, assuming 210 Monday to Friday working days per calendar year) and count the total number of office visits over that period of time, A. Then, the average daily patient visit rate p is calculated as the ratio of the number of office visits A and the product of the number of patients on the current panel N and the number of days T in the period over which the appointments were counted (say, $T = 315$ for a period of 18 months): $p \sim (A/(N \times T))$. For example, consider a general/family practitioner with a current panel of $N = 2{,}500$ patients and suppose that the examination of her appointment log has established that the practice had $A = 6{,}500$ office visits over the last 18 months ($T = 315$ days). For this practice, we get $p = (A/(N \times T)) = (6{,}500/(2{,}500 \times 315)) = 0.008$ visits/day. Note that this value represents the average over a long period of time and is most appropriate for modeling long-term demand patterns. Over any short-term period, this value can underestimate or overestimate the actual demand rate. In Sect. 3.2.7 we will discuss the effect of predictable short-term variations in the demand rate on the recommended panel size values.

3.2.3 Establishing the Target Number of Daily Appointment Slots

To estimate the target supply of appointment slots C, the practice needs to decide upon the average length of an appointment slot and the average daily number of hours devoted to direct patient care. For example, if one assumes that a physician spends an average of 7 h per day in direct patient care and that appointments are scheduled at 20-min intervals, the target daily appointment capacity is $C = 7$ h x 3 appointments/h $= 21$ appointments. We use the term “target capacity” to reflect the subjective maximum length of the working day for a particular primary care physician. While under the advanced access approach daily fluctuations in patient demand may force a physician to work beyond this limit on any particular day, such extra work is considered undesirable.

3.2.4 Setting the Target overflow Frequency

In primary care, patient demand exhibits significant day-to-day variability. In part, this variability is predictable and can be attributed to changes in “calendar variables” (Batal et al. 2001): the expected number of daily patient requests to see a physician is often a *deterministic* function of the day of the week and time of the year. If such deterministic variability were the only source of day-to-day changes in patient demand, a primary care practice would potentially be able to provide an exact match between the patient demand and the supply of service capacity. However, a substantial part of the observed demand variability is *random*, i.e., it

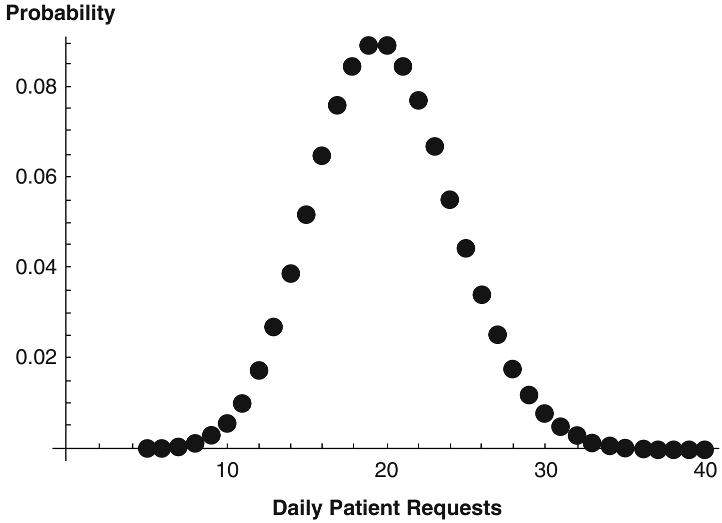


Fig. 8.6 Daily patient demand distribution under binomial model: panel size $N = 2{,}500$, demand rate $p = 0.008$

cannot be predicted in advance. In this section, we use a simple model of daily demand which accounts for this random variability component. Later, in Sect. 3.2.7, we extend our analysis to include the predictable weekly/seasonal demand variability.

In our demand model, we assume that for each patient a request for care is generated independently of any other patient's request, so that the total daily demand for primary care services can be modeled as a *binomial* random variable with the expectation equal to Np and the variance equal to $Np(1 - p)$. The patient demand rate p is considered to be constant and not subject to day-to-day variations—in other words, it represents the long-term average demand rate estimated in Sect. 3.2.2. Figure 8.6 illustrates this demand model by showing the distribution of binomial daily demand requests from a panel of $N = 2{,}500$ patients with $p = 0.008$. We observe that while the expected number of patient appointment requests on any day is equal to 20, the actual number of requests can very well be anywhere between 15 and 25.

Using our model and the target number of appointment slots available each day C, we can estimate the effect of panel size on the ability to offer same day appointments by calculating the probability that the demand for appointments exceeds the supply of appointment slots on any given day. We call this probability "overflow frequency." In particular, the overflow frequency for a primary practitioner who sets the target number of daily appointment slots to C and who serves a panel of size N with daily patient visit rate of p is equal to

$$f = 1 - \sum_{k=0}^{C} \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}. \tag{8.1}$$

The overflow frequency as expressed in (8.1) rapidly increases with the size of patient panel N and rapidly decreases with C. For $N = 2{,}500$, $p = 0.008$ and $C = 21$, the calculated value of the current overflow frequency is 41.1 %. This high value of overflow frequency indicates that with the given panel size a physician will not be able to "do today's work today" on a consistent basis without frequent overtime work. Clearly, the target level of overflow frequency should be much lower. In particular, a physician could consider setting it to 5 % (approximately once a month), 10 % (once in 2 weeks), or 20 % (once in a week). When defining the overflow frequency, it should be kept in mind that high overflows are equivalent to long overtime work: in our example ($N = 2{,}500$, $p = 0.008$ and $C = 21$, overtime frequency of 41.1 %) the average duration of overtime when it happens is more than an hour. Thus, limiting overflow frequency to 20 % may be advisable.

3.2.5 Computing the Appropriate Panel Size

The last step in calculating the appropriate panel size involves iterative adjustment of the trial panel size until the corresponding overflow frequency matches the target level. If the overflow frequency corresponding to the current panel size is higher than the target level, one should lower the panel size and repeat the overflow frequency calculation using (8.1). Similarly, if the computed overflow frequency turns out to be too low, the panel size should be increased. In our example, the computed initial value of the overflow frequency (41.1 %) is much higher than the target level of 20 %, so we need to diminish the panel size. Repeating the calculation of (8.1) for the trial panel size of $N = 2{,}000$, we obtain an overflow frequency of 15.9 %—which is lower than the level of overtime work a physician is willing to tolerate. Adjusting panel size upwards to $N = 2{,}250$ and recomputing the overflow frequency, we get the value of 32.9 %—above the target level. Going down to the panel of $N = 2{,}100$ patients, we get the overflow frequency of 21.8 %. Repeating these panel size iterations several more times, we finally achieve the overflow frequency of 20.04 % for the panel size of $N = 2{,}072$. This overflow frequency is pretty close to our target level of 20 % (due to discrete nature of the panel size it may not be possible to find the panel size that results in the overflow frequency of exactly 20 %). Thus, a panel size of 2,072 patients is recommended for the considered practice.

Table 8.1 Panel sizes (capacity utilizations) for different parameter values, primary care type: general and family practice and pediatrics

	General and family practice		Pediatrics	
Overflow frequency	Daily slots = 24	Daily slots = 20	Daily slots = 24	Daily slots = 20
5 %	2,294 (72 %)	1,852 (69 %)	1,475 (72 %)	1,191 (69 %)
10 %	2,468 (77 %)	2,006 (75 %)	1,586 (77 %)	1,290 (75 %)
20 %	2,697 (84 %)	2,211 (83 %)	1,734 (84 %)	1,421 (83 %)

3.2.6 Examples Based on NAMCS 2002 Data

While the NAMCS 2002 survey (www.cdc.gov/nchs/about/major/ahcd/ahcdl.htm) reports the total number of annual visits to general and family practitioners in the USA (215,466,000), the annual visit rate per patient is not easy to estimate since we could not find reliable statistics on the number of people who actually use (or even have) a primary care physician. The rate of 0.761 annual office visits per person, reported in NAMCS 2002 survey, was obtained by dividing the total number of visits to general and family practitioners by the entire US population (283,135,000), taken from 2000 Census data. Clearly, using this value would result in a gross underestimation of actual patient visit rates. The rate we use in our study (1.575 annual visits per patient) is calculated based on the assumption of 210 annual in-office days and on the assumption (used in Murray and Berwick 2003) that in an average patient panel not overly weighed with elderly and chronically ill patients, 0.7–0.8 % of patients will request visit on an average day. Table 8.1 shows the patient panel sizes (and attained capacity utilizations) for a "typical" general and family practitioner (on average, 1.575 annual visits per patient) and a "typical" pediatrician (on average, 1.98 annual visits per child according to 2002 NAMCS) which would result in an overflow frequency of 5 % (approximately, once a month), 10 % (twice a month), or 20 % (once a week).

2002 NAMCS reported that the average duration of the "face-to-face" part of the office visit is 16.1 min for general and family practice and for pediatrics, 18.1 min for OB/GYN practice and 20.0 min for internal medicine practice. In our calculations we considered appointment intervals of 20 min, which is likely to be a realistic estimate for the duration of a typical appointment in a primary care setting. An 8-h workday would produce 24 daily appointment slots (for a 5-day working week this number roughly corresponds to 40.2 h spent by a family physician on direct patient care or patient-related service during a complete week of practice, according to a recent AAFP survey, http://www.aafp.org/×769.xml). Since the actual daily appointment capacity is likely to be somewhat lower than this optimistic estimate, in our calculations we also consider an alternative daily capacity of 20 appointment slots. The calculations were performed using (8.1) under the assumption of 210 workdays per year. This value, in our estimate, is a good representation of the annual number of workdays for the majority of primary care practices.

3.2.7 Accounting for Weekly and Seasonal Variations in Patient Demand

A recipe for setting the size of patient panel described above is based on the assumption that patient demand rate is the same every day. In practice, however, the expected demand for primary care is subject to weekly as well as seasonal variation. Consider a general/family practitioner who has the target number of daily slots equal to 20 and who is serving a panel of size $N = 22/1$ for which the long-term average demand rate is $p = 0.075$, as defined in Sects. 3.2.1 and 3.2.2. This combination of parameters corresponds to one of the cells in Table 8.1 and results in an average overflow frequency of 20 %.

Now, suppose that the daily demand rates change from day to day during each week as well as from month to month. In particular, let p_{ij} be the demand rate from any patient on the panel on day i in month j: for example, $P_{\mathrm{TUE,APR}}$ stands for the demand rate on a Tuesday in April. Under these assumptions, patient demand on day i in month j is a binomial random variable with the mean Np_{ij} and the variance $Np_{ij}(1 - \mathrm{p}_{ij})$. Thus, demand rate dynamics is described by $5 \times 12 = 60$ demand rate parameters, such that their average is equal to p:

$$\frac{1}{60}\sum_{i=\mathrm{MON}}^{\mathrm{FRI}}\sum_{j=\mathrm{JAN}}^{\mathrm{DEC}} p_{ij} = p. \tag{8.2}$$

For a given value of p, it is convenient to describe this demand rate dynamics in terms of the calendar adjustment factors, which are defined as ratios of the average demand value on a particular day and the long-term average demand over all days. In our model, the calendar adjustment factors can be computed as $q_{ij} = Np_{ij}/(Np) = p_{ij}/p$, so that the daily expected demand is Npq_{ij} and its variance is $Npq_{ij}(1 - pq_{ij})$. Such adjustment factors reflect how much higher or lower the demand rate is on a particular day as compared to the long-term average demand rate. Figure 8.7 shows the calendar adjustment factors computed from the appointment data at Scott and White Killeen Clinic (TX) as reported in Forjuoh et al. (2001).

In this example, for most of the months, demand values are the highest on Mondays (average calendar adjustment factor is 1.234), sharply dropping on Tuesdays (0.998), before leveling off on Wednesdays (0.909), Thursday (0.921), and Fridays (0.938). Across months, daily demand rates follow mostly similar trajectories, reaching peak in colder months (average calendar adjustment factor from October through April is 1.088) and dropping in warmer months (average calendar adjustment factor in May through September is 0.887).

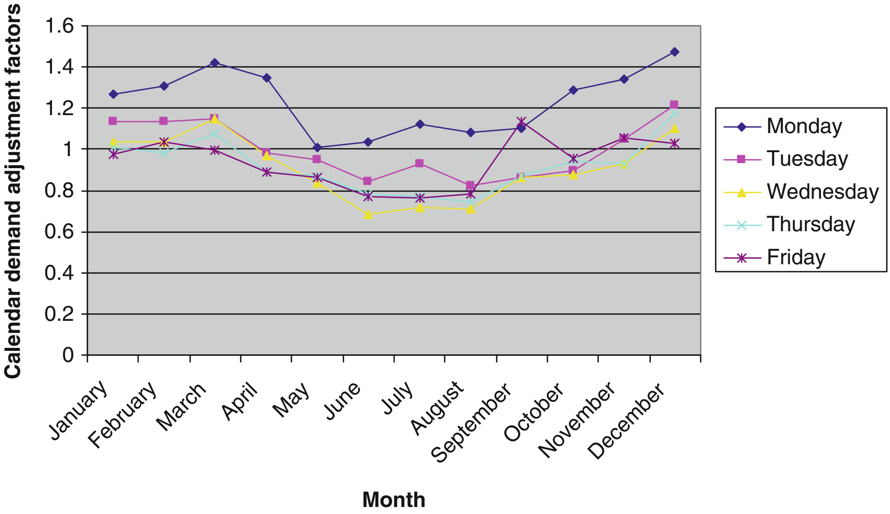


Fig. 8.7 Calendar demand adjustment factors based on the data from Scott and White Killeen Clinic (TX) as reported in Forjuoh et al. (2001)

Given these variations in demand rates, the resulting overflow frequencies differ from day to day. In particular, on day i in month j the overflow frequency can be computed using the generalization of (8.1):

$$f_{ij} = 1 - \sum_{k=0}^{C} \frac{N!}{k!(N-k)!} \left(pq_{ij}\right)^k \left(1 - pq_{ij}\right)^{N-k}. \tag{8.3}$$

Then, the average overflow frequency can be obtained as follows:

$$f^* = \frac{1}{60} \sum_{i=\text{MON}}^{\text{FRI}} \sum_{j=\text{JAN}}^{\text{DEC}} f_{ij}. \tag{8.4}$$

Would the resulting average overflow frequency f^* be far-away from the 20 % value reported in Sect. 3.2.6 for the case of time-independent demand rate? In the Table 8.2 below we show the overflow frequency values f_{ij} as well as the average f^*.

Two important observations can be made on the basis of the values in Table 8.2. First, uniform 20 % overflow frequency is replaced by a wide range of values: from 82 % on Mondays in December to 0 % on Wednesdays in June. Second, the average value of the overflow frequency rises to 24 %. The last observation indicates that in the presence of the demand rate variability advanced access approach cannot be sustained at the same level of overflow frequency. In this regard, primary care office will have to make certain adjustments—either in the size of patient panel it serves, or in the way the target appointment slot numbers are distributed across different days. Below we consider each of these possibilities in detail.

Table 8.2 Overflow frequency values for different days and months based on the data from Scott and White Killeen Clinic (TX) as reported in Forjuoh et al. (2001)

Overflow frequency	Monday	Tuesday	Wednesday	Thursday	Friday	Average
January	0.59	0.39	0.25	0.22	0.17	0.32
February	0.64	0.39	0.25	0.18	0.25	0.34
March	0.77	0.41	0.41	0.30	0.19	0.42
April	0.69	0.18	0.16	0.08	0.08	0.24
May	0.22	0.14	0.05	0.07	0.06	0.11
June	0.25	0.05	0.00	0.02	0.02	0.07
July	0.37	0.12	0.01	0.02	0.02	0.11
August	0.31	0.04	0.01	0.01	0.02	0.08
September	0.34	0.06	0.06	0.07	0.39	0.19
October	0.62	0.09	0.08	0.13	0.15	0.21
November	0.68	0.26	0.12	0.12	0.28	0.29
December	0.82	0.52	0.34	0.45	0.24	0.47
Average	0.52	0.22	0.15	0.14	0.16	0.24

The downward revision of the patient panel size may be necessary in practices where a physician is unwilling to compromise on the daily target amount of work. In the presence of daily/seasonal demand variability, a physician may select one of several criteria to limit the patient panel size.

Generalizing the approach of Sects. 3.2.5 and 3.2.6, one can use the average overflow frequency f^* as the appropriate measure of overtime effort. As Fig. 8.8 shows, a relatively small adjustment would be necessary to bring its value from the current 24 % to the 20 % level: panel size would have to be reduced from current 2,211 to 2,125 patients.

One possible downside of choosing the value of f^* as the guide for adjusting the panel size is that, as Table 8.2 shows, the daily overflow frequencies may be quite different from f^*: for example, when the panel size is set at 2,125 patients (and the average overflow frequency is 20 %), the maximum overflow frequency (namely, the one for Mondays in December) turns out to be 77 %, which roughly corresponds to working overtime three out of four Mondays in December. Thus, while for panel size of 2,125 the overtime work does not exceed the desired 20 % limit on average, there could be some short-term runs of nearly certain overtime.

To decrease the chance of such short-term overtime runs, a physician may want to focus on controlling a different measure of overtime work, for example, the fraction of days for which the chance of overtime exceeds the desired target of 20 %. An example of such measure is provided by Table 8.2, where out of 60 distinct day types we consider, the overflow frequency exceeds 20 % on 28 days (which is 46.6 % of 60 days). For the panel size of 2,125, such number turns out to be 21 day, or 35 % of all days. If such fraction is deemed too high, the panel size may have to be decreased further. Figure 8.9 shows how this overtime measure, the fraction of days for which the overflow frequency exceeds the target level of 20 %, changes with panel size. We observe that if a physician would like to

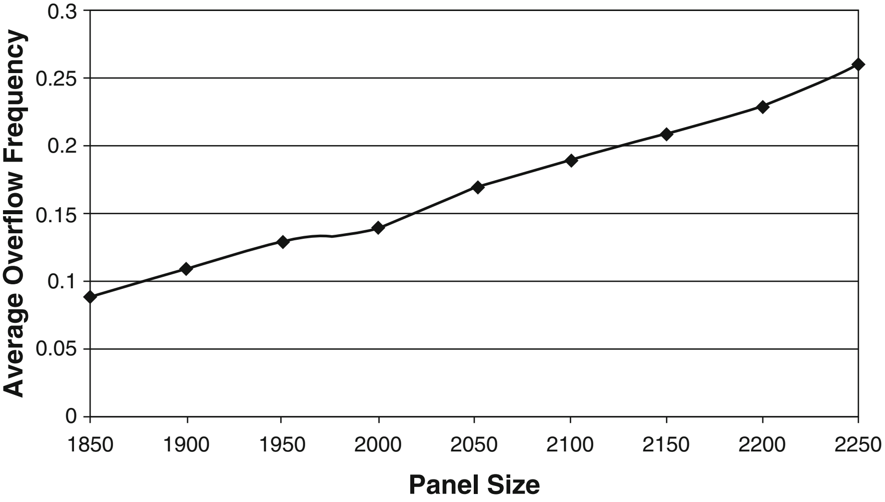


Fig. 8.8 Average overflow frequency *f** as a function of patient panel size

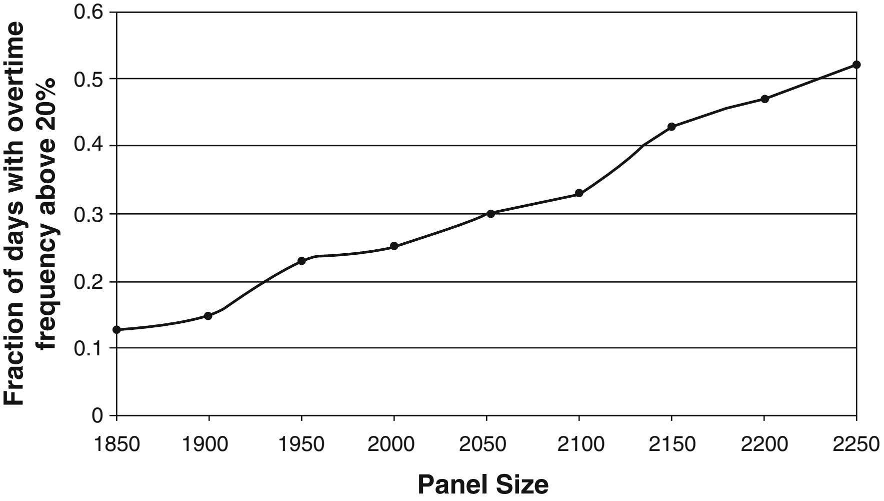


Fig. 8.9 The fraction of days for which the overflow frequency is above the target level of 20 %, as a function of patient panel size

limit the fraction of days for which the overflow chance exceeds the target to, for example, 25 %, the panel size should not exceed 2,000 patients.

The analysis conducted above relies on the assumption that the primary care provider is willing to reduce the size of patient panel, but remains rather inflexible with respect to day-to-day changes in the target duration of daily work. An alternative assumption would describe an environment where provider would like

Table 8.3 Target values for appointment capacity for different days and months based on the data from Scott and White Killeen Clinic (TX) as reported in Forjuoh et al. (2001)

C_{ij}	Monday	Tuesday	Wednesday	Thursday	Friday	Average
January	25	23	21	20	20	21.8
February	26	23	21	20	21	22.2
March	28	23	23	21	20	23.0
April	26	20	19	18	18	20.2
May	20	19	17	18	17	18.2
June	21	17	14	16	16	16.8
July	22	19	15	16	16	17.6
August	22	17	15	15	16	17.0
September	22	17	17	18	23	19.4
October	25	18	18	19	19	19.8
November	26	21	19	19	21	21.2
December	29	24	22	23	21	23.8
Average	24.3	20.1	18.4	18.6	19.0	20.1

to retain given panel by adjusting her day-to-day target work duration. In such environment, day-to-day demand rate variability is matched by the corresponding variability in the target appointment capacity. In particular, on day i in month j, this target C$_{ij}$ has to be selected so that the overflow frequency

$$\bar{f}_{ij} = 1 - \sum_{k=0}^{C_{ij}} \frac{N!}{k!(N-k)!} \left(pq_{ij}\right)^k \left(1 - pq_{ij}\right)^{N-k}$$

is as close as possible to 20 % (due to the discrete nature of C_{ij} it may not be possible to exactly match this value). Table 8.3 shows the resulting values of the target appointment capacity for the panel size $N = 2{,}211$ and the demand rates taken from Forjuoh et al. (2001).

Appointment capacity values from Table 8.3 show how the target primary care capacity should be adjusted in order to match patient demand from a panel of 2,211 patients for any day of the year. In particular, while the primary care provider can expect short office days on Wednesdays in June (14 appointments, or about 4 h and 40 min), but very long days on Mondays in December (29 appointments, or 9 h and 40 min). Note that, on average, the target length of the workday (20.1 appointments) is virtually the same as in the case of stationary demand considered in Sect. 3.2.6.

In summary, significant day-to-day variability in the patient demand rates may require adjustments in either the panel size (as indicated in Figs. 8.8 and 8.9) or in the distribution of the target appointment capacity (as shown in Table 8.3).

3.3 Demand Control Techniques

In the primary care setting, physician time is usually the most constrained resource. Effectiveness of any appointment scheduling approach depends on how this valuable resource is managed. Murray and Tantau (2000), Murray and Berwick (2003), and Oldham (2001) outline several approaches to improving the match between demand and supply in primary care settings: (1) enforcing the continuity of patient care, (2) increasing the effectiveness of each appointment, and (3) reducing the demand for face-to-face patient–physician interactions.

Continuity of care plays the major role in reducing the unnecessary demand for future care in advanced access settings. When a patient is attended by a physician who is not her PCP, the probability of an extra follow-up visit (for which patient requests to be seen by the PCP) increases, creating an avoidable future demand: Houck (2004) refers to data from Kaiser Permanente which indicate that as many as 48 % of patients who were seen by a physician other than their PCP, return within 2 weeks to see “their” physician. In addition, continuity of care can increase profitability of primary care operations: O’Hare and Corlett (2004) report that the relative value unit (RVU) per patient visit was up to 17 % higher for visits where patients were treated by their PCPs, with an average increase in physician’s compensation of about $4.50 per visit.

Maximizing the value of each appointment (“max-packing”) is another way of reducing the need for future appointments. Gordon and Chin (2004) describe a “combing” technique that could be used to facilitate max-packing of appointments. “Combing” is used every time a patient requests an appointment: the schedule is checked for any appointments and/or some anticipated needs (annual checkup, flu shots, etc.) for the same patient in the near future. This way, a single appointment can be used to attend to multiple patient needs. “Max-packing” is clearly appropriate in a managed-care environment but, as Murray and Tantau (2000) argue, it can also be useful in fee-for-service settings, since a more service-intensive appointment would correspond to a higher CPT code.

Effective demand reduction techniques may include broad use of phone and e-mail to substitute for various components of face-to-face interaction between patients and primary care office. It can be argued that advanced access reduces patients’ “anxiety” about getting an appointment and, in a paradoxical way, reduces the need to book face-to-face appointments, opening the way to handling a larger fraction of demand through e-mail or phone interactions. E-mail can be used for repeat prescriptions, checking the test results, appointment reminders. In this regard, Oldham (2001) argues for the use of separate e-mail addresses for receptionists, nurses and physicians: some advanced access primary care offices report that patients use e-mail to query receptionists nearly as often as physicians (for example, see Patient Online system at Dartmouth-Hitchcock Medical Center in New Hampshire, www.dhmc.org). Phone consultations (with a nurse or a physician) may be used for managing same-day demand, follow-up appointments, and other queries. Oldham (2002) reports as much as a 30–50 % reduction in face-to-face consultations as a result of phone management of same day demand, and a 15–20 % reduction in follow-up consultations.

Demand can also be reduced via increasing the length between patient visits as long as this does not contradict the requirements of medical necessity. Gordon and Chin (2004) describe an implementation of advanced access system under which some of the patients with chronic conditions are seen every 3 or 4 months instead of "standard" 2 months, in cases when a physician felt that a patient can manage her condition on her own through medication and monitoring. It could also be argued that patients do not necessarily associate an increase in inter-appointment intervals with lower service quality: a recent study by Wick and Koller (2005) indicates that patients prefer longer (by 6 days, on average) intervals for return visits than their physicians.

Finally, group patient consultations, or cooperative health care clinics (CHCC), can be used to combine visits for patients with similar chronic conditions. Houck et al. (2003) provides detailed instructions on how such group sessions have to be organized and run. Selecting the "right" type of patients for group visits is a key factor determining the effectiveness of this approach for overall reduction of the demand for appointments: group visits could work best for patients with chronic conditions characterized by potentially high rate of office visits (e.g., hypertension, asthma, diabetes, depression) and/or geriatric patients with multiple comorbidities. CHCC can also be useful for patients without established chronic conditions who nevertheless generate high number of annual office visits. Existing empirical evidence points out that the use of group visits combines improved demand management with greater patient satisfaction, lower medical costs, and better medical outcomes: Beck et al. (1997) report the patients participating in group sessions had fewer emergency room visits, visits to sub-specialists, and repeat hospital admissions (on a per patient basis), and a higher rate of flu and pneumonia vaccinations. An important consequence of the use of group sessions is a decreased use of physician's time for services that can be delivered by other personnel: group participants made more visits and calls to nurses and fewer calls to physicians, while exhibiting higher overall satisfaction with care and reduced cost of care. Masley et al. (2000) report that the introduction of group sessions for patients with poorly controlled type 2 diabetes lead, after a year, to a 32 % average reduction in total cholesterol/HDL ratios, a 30 % average reduction in $\mathrm{HbA_{ic}}$ levels, and a 7 % average reduction in medical expenses.

3.4 Working Down the Appointment Backlog

Before the advanced access can be effectively implemented, it is important that the "bad" appointment backlog accumulated in the system is eliminated. This one-time backlog clearing requires a temporary increase in service capacity that is needed to absorb new appointment demand, while working down a backlog. A "Backlog Reduction Worksheet" developed by Batalden, Godfrey, and Nelson in 2003 and publicized by the Institute for Healthcare Improvement (www.ihi.org/IHI/Topics/OfficePractices/Access/Tools/, login required) outlines practical steps involved in successful implementation of this process. Typically, the extra capacity is gained by extending regular working hours; other strategies include the use of weekend appointment sessions, employing locum tenens, even (temporarily) rejecting new prescheduled (not same-day) appointments and off-loading other duties, such as teaching.

The length of this transition period varies from practice to practice and depends on how big the initial backlog is and how much extra appointment time can a practice afford. For one of the very first implementations of advanced access, Murray and Tantau (2000) report that it took 6 weeks to work down a 2-month appointment backlog in a six-physician clinic. On the other hand, Grandinetti (2000) states that Community Pediatric-and Adolescent Medicine section of Mayo Clinic in Rochester, MN had to use between 3 and 6 months to eliminate its appointment backlog; and in small practices, the transition period can be even longer.

3.5 Post-introduction Management

Demand for primary care is inherently random and appointment capacity of a primary care office can change at a short notice. After the introduction of the advanced access approach, a set of robust contingency plans should be developed for dealing with both the expected as well as the unexpected temporary mismatches between supply and demand: staff sickness, vacations, demand surges (e.g., due to epidemics), etc. The UK National Primary Care Development Team (NPDT) advocates delegating the coordination and real-time management of such plans to a designated "contingency" person whose responsibilities include constant monitoring of the state of the appointment system, and activating a contingency plan when the situation warrants (www.npdt.org/Pre-Bookable.pdf). Contingency planning could include personnel cross-training (Nolan et al. (1996)) as well as the use of demand smoothing techniques before the predictable surge in appointment requests, such as "staggering" of demand for physicals near the beginning of a school year (Murray and Tantau 2000).

Monitoring the degree of mismatch between supply and demand is an important component of the overall management of advanced access. Oldham (2001) provides a detailed discussion on the practical use of two measures of patient access to primary care: time to third next available appointment (TTNAA) and percentage of patients receiving an appointment on the day of their choice, which we call "access fraction."

TTNAA is an access measure directly related to the length of the appointment backlog. When appointment cancellations are frequent, the position of the first and even second available appointments may not be indicative of the typical backlog. Third available appointment, on the other hand, is a much more stable measure of access, not easily affected by cancellations. According to the Institute for Healthcare Improvement, the goal of advanced access is to reduce the wait for the third next available appointment to 24 h for general and family practices and to 2 days for specialty practices. An extensive set of reports on the achieved values of TTNAA for a number of different practices (primary as well as specialty) is available on the site of the Wisconsin Collaborative for Healthcare Quality (www.wiqualitycollaborative.org).

The fraction of patients obtaining an appointment on the day of their choice (whether it is the "same day" or some other day in the future) could serve as a proxy for patient satisfaction with the timeliness of the received care. While the exact relationship between this access fraction and TTNAA depends on the details of patients' preferences for the timing of their appointments, it is likely that high values of the access fraction correspond to the low values of TTNA. In this regard, Oldham (2001) suggests that a 90 % value of access fraction would be roughly equivalent to TTNAA of 1 day.

4 Conclusions

The lack of timely access to primary care is a well recognized problem of the US health care system. In the UK, the National Primary Care Collaborative, the first program of the state-sponsored National Primary Care Development Team, has reached over 5,000 practices and more than 32 million patients in the effort to reduce patient waiting and to improve patient service. In the USA, the growing number of primary care offices is adopting the advanced access practice developed by Murray and Tantau in order to reduce or even eliminate appointment backlogs. One of the important enablers of the advance access approach is the overall match between the demand for primary care services and the supply of the appointment capacity. In our analysis, we explicitly model the connection between the patient panel size and the daily demand for appointments and develop analytical expressions for the frequency of overtime work which is required to sustain advanced access for a given level of appointment capacity. Using our model, we design a set of guidelines that can be used by primary care offices to determine the patient panel size to match the preset target level of overtime work.

The advanced access is quickly transforming itself from a new concept to a day-to-day routine. In our opinion, many primary care practices that contemplate adopting advanced access would greatly benefit from a decision support system that codifies the basic rules of advanced access and helps with its implementation and maintenance. Such a system could successfully complement the functionality of existing office management software (appointment recording, patient databases, billing, etc.) by adding new patient flow management capabilities.

References

Batal, H., Tench, J., McMillan, S., Adams, J., & Mehler, P. S. (2001). Predicting patient visits to an urgent care clinic using calendar variables. *Academic Emergency Medicine, 8*(1), 48–53.

Beck, J., Scott, J., Williams, P., Robertson, B., Jackson, P., Gade, G., et al. (1997). A randomized trial of group outpatient visits for chronically ill older HMO members: The cooperative health care clinic. *Journal of American Geriatric Society, 45*(5), 543–549.

Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: A review of literature. *Production and Operations Management, 12*(4), 519–549.

Declaration of the International Conference on Primary Health Care. (1978). Retrieved from http://www.euro.who.int/AboutWHO/Policv/20010827-1.

Forjuoh, S. N., Averitt, W. M., Cauthen, D. B., Symm, B., & Mitchell, M. (2001). Open-access appointment scheduling in family practice: Comparison of a demand prediction grid with actual appointments. *Journal of the American Board of Family Practice, 14*(4), 259–265.

Gordon, P., & Chin, M. (2004). Achieving a new standard in primary care for low-income populations: Case study 2: Advanced access learning. *The Commonwealth Fund Report*, New York, NY.

Grandinetti, D. (2000). You mean I can see a doctor today? *Medical Economics*, 77(6):102-4, 109, 113-4.

Houck, S. (2004). *What works: effective tools and case studies to improve clinical office practice* (p. 90). Boulder, CO: HealthPress Publishing.

Houck, S., Kilo, C., & Scott, J. C. (2003). Improving patient care. Group visits 101. *Family Practice Management, 10*(5), 66–68.

Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century* (p. 51). Washington, DC: National Academy Press.

Masley, S., Sokoloff, J., & Hawes, C. (2000). Planning group visits for high-risk patients. *Family Practice Management, 7*(6), 33–37.

Murray, M., & Berwick, D. M. (2003). Advance access: reducing waiting and delays in primary care. *JAMA, 289*(8), 1035–1039.

Murray, M., & Tantau, C. (2000). Same-day appointments: Exploding the access paradigm. *Family Practice Management, 7–8*, 45–50.

Nolan, T. W., Schall, M. W., Berwick, D. M., & Roessner, J. (1996). *Reducing delays and waiting times throughout the healthcare system*. Cambridge, MA: Institute for Healthcare Improvement.

O'Hare, C. D., & Corlett, J. (2004). The outcomes of open-access scheduling. *Family Practice Management, 11*(2), 35–38.

Oldham, J. (2001). *Advanced access in primary care*. National Primary Care Development Team. Retrieved from http://www.npdt.org/AccesstoPrimaryCare/advancedaccess.pdf.

Oldham, J. (2002). Telephone use in primary care. *British Medical Journal, 325*(7363), 547.

Starfield, B. (1991). Primary care and health. A cross-national comparison. *JAMA, 266*, 2268–2271.

Wick, A., & Koller, M. T. (2005). Views of patients and physicians on follow-up visits. *Swiss Medical Weekly, 135*, 139–144.

Chapter 9
Waiting Lists for Surgery

Emilio Cerdá, Laura de Pablos, and Maria V. Rodriguez

Abstract Health waiting lists in general and surgical waiting list in particular are a problem for the majority of the European countries with a National Health System. In this chapter, the problem of the waiting lists for surgery from a general perspective in the scope of the health management in the European Union (EU) is analyzed. Also, applying mathematical programming techniques, we intend to design the real performance of surgical services at a local general hospital offering the decision maker a suitable methodology that allows us to analyze whether or not it is possible to improve the running of the services, taking into account all the real constraints, e.g., space, staff availability, waiting time upper limit, or financial support.

Keywords Waiting lists • National Health Systems • Hospital management • Mathematical programming

1 Introduction

The health systems in the European Union are aligned in two groups:

- The system inspired by the Beveridge Report of 1942, which formalized the health organization adopted by Sweden in the year 1930, establishing a National Health System. The UK, Denmark, Italy, Portugal, Finland, Sweden, Norway, Ireland, Greece, and Spain can be considered in this group.

We thank B. Readman for the linguistic revision of the text.

E. Cerdá (✉) • L. de Pablos
Departamento de Fundamentos del Análisis Económico I, University Complutense of Madrid, Campus de Somosaguas, 28223 Madrid, Spain
e-mail: ecerdate@ccee.ucm.es

M.V. Rodriguez
University of Oviedo, Calle Doctor Fernando Bongera, s/n, 33006 Oviedo, Spain

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_9, © Springer Science+Business Media New York 2013

- The Bismarck System, inspired in the social legislation of Germany in 1883, and which is traditionally known as Social Security. Austria, Belgium, France, Germany, and the Netherlands can be considered in this group.

However, it is necessary to notice that neither of the models is strictly applied. Although each has its preponderant characteristics, each takes something from the other.

The existence of waiting lists is a usual fact in the countries with a National Health System. In some ways, lists are a method for managing the health services. In Spain the surgical waiting lists are a priority from both the political and social point of view.

This chapter is organized as follows: In Sect. 2, we review the dimension of waiting lists in some European countries, the main features of the Health Systems of the countries with and without waiting lists and the factors that could explain waiting time differences. In Sect. 3 we describe the major policies used to reduce waiting times. The issue of Sect. 4 is the choice between public and private sector in relation to waiting lists. The introduction of different standardized prioritization rules for waiting lists is discussed in Sect. 5. Section 6 proposes a specific model to manage the surgical waiting lists. This model has been applied to an important public hospital in Madrid, and is capable of adaptation to any other hospital. It is a mathematical programming program that has elements of linear programming, integer programming, dynamic optimization and multi-objective programming. Finally, Sect. 7 presents conclusions and future directions.

2 National Health Systems and Surgery Waiting Lists in European Countries

2.1 Waiting Lists and Health Systems in Some European Countries

An important feature of EU countries is that while some countries report significant waiting times for non-emergency surgery, others do not. Waiting times are a serious health problem in Denmark, Finland, Ireland, Italy, Sweden, the UK, Spain, Portugal, Norway, and Greece. However, waiting times are not reported administratively because they are low and insignificant in a second group of countries, such as Germany, Belgium, Luxembourg, and France.

Tables 9.1 and 9.2 show average and median waiting times in some European countries. The data were collected through a questionnaire submitted to 12 countries involved in the OECD Waiting Time Project (Hurst and Siciliani 2003). The first question resolved in that project was the identification of a common definition of waiting time. This allows comparisons of waiting times across countries. In this sense, inpatient waiting time for patients admitted for treatment is defined by Hurst

Table 9.1 Mean inpatient waiting times of patients admitted for surgical procedure (days). Year 2000. (*Source*: Siciliani and Hurst 2003)

	Denmark	Finland	Norway	The Netherlands	Spain	Sweden	UK
Hip replacement	112	206	133	96	123		244
Knee replacement	112	274	160	85	148		281
Cataract	71	233	63	111	104	199	206
Varicose veins	99	280	142	107			227
Hysterectomy		100	64	61	102		159
Prostatectomy		81	75	60	62		52
Cholecystectomy	75	159	103	71	107		156
Inquinal femoral hernia	73	125	109	75	102		150

Table 9.2 Median inpatient waiting times of patients admitted for surgical procedure (days). Year 2000. (*Source*: Siciliani and Hurst 2003)

	Denmark	Finland	Norway	UK
Hip replacement	87	148	99	211
Knee replacement	90	202	132	262
Cataract	36	189	28	182
Varicose veins	69	155	110	178
Hysterectomy		70	37	110
Prostatectomy		39	47	37
Cholecystectomy	57	90	63	97
Inquinal and femoral hernia	46	74	74	95

and Siciliani (2003) as "the time between specialists' assessment and the surgical treatment". A more comprehensive measure of waiting time for surgery would cover the whole period from the time a general practice refers the patients to a specialist, to the date they are admitted for the surgical procedure. This latter measure also will include any delay between a general practice referral and the specialist's treatment (outpatient waiting time). The data reported by the different countries are waiting time of patients admitted for surgical treatment.

Measures of waiting times are often aggregated through the utilization of statistics. The most commonly utilized are the mean and the median. If the waiting time is distributed according to a normal (Gaussian) distribution, the two measures coincide. However, as is often observed, the waiting-time distribution tends to be positively skewed. This implies that there is a low proportion of patients with consistently long waits. In this case, using the median is usually recommended because the mean tends to be heavily influenced by the few patients with long waits.

The countries with highest waiting times are England and Finland, followed by Denmark, Norway, and Spain. It is interesting to note how the waiting times for less urgent procedures are systematically higher than the waiting times for more urgent procedures.

The main characteristics of The Health System of the countries with waiting lists are:

- They have universal health coverage and are financed mainly through general taxation. Health care is provided by a National Health System. However: (1) In Ireland the health system is a mix of public and private institutions and funders. (2) In Norway, prior to 2002, the Health System was organized into three political and administrative levels: national government (legislation and regulation), counties (secondary care) and municipalities (primary care). The counties were responsible for financing and planning. Since January 2002, the responsibility for hospital care was removed from the counties to the state. Five regional enterprises have been established and the objectives and the basis for management of these enterprises are determined by central government. (3) In Spain the responsibility for provision of health services was, until 2002, partly centralized under INSALUD (the National Institute of Health), and partly decentralized to some Autonomous regions (Catalonia, Andalusia, Land of Valencia and Balearic Islands). From January, 2002, it has been wholly decentralized to all 17 Autonomous Regions. (4) In Sweden the health system is organized into three political and administrative levels. National government is responsible for legislation and control; county councils for primary and secondary care; and municipalities, for care of the elderly and nursing homes.
- Generally, in countries with waiting lists, the patients cannot access elective surgery without general practice referral.
- The majority of the hospitals are owned and financed by the public sector. The private sector plays a marginal role as a supplier of health care. Sometimes budgets for hospitals have been based on historical funding (Ireland, Denmark, and Spain fix budget bases on past expenditure and the case mix of the hospital). In England public hospitals are remunerated according to contracts or arrangements that specify the services that must be provided. From 2002 Norway has been implementing the activity-based system for financing hospital care (this system is also applied in Sweden).
- Hospital doctors are salaried in the public sector. Often the specialists working in publicly funded hospitals are not allowed to treat patients within the same hospital as private doctors and have serious restrictions on working in the private sector. Only "part-time" specialists in publicly funded hospitals are allowed to work in privately funded hospitals (Sweden, Spain, Italy, and England).
- Usually there are no co-payments for receiving publicly funded surgery. However: (1) In Finland, an outpatient visit costs 20 euros and each day in the ward costs 25 euros. (2) In Ireland, there are two categories of patients. The first category, public inpatients, receives services free of charge. The second one is subject to a daily overnight charge of 40 euros (in 2003), subject to an overall annual limit of 400 euros. (3) In Norway, patients pay 114 NOK for a general practice visit and 200 NOK for a specialist visit. The overall annual limit is 1,350 NOK. (4) In Sweden, the patient usually pays a token daily fee for each day spent in the hospital. For a general practice, he/she pays 15 euros and for a specialist visit, 20 euros. The overall annual limit for the last two treatments is 90 euros.

- The countries that do not report waiting time administratively, because it is low and insignificant, are Germany, Belgium, Luxembourg, and France. The health system of these countries is based on health insurance, generally composed of a basic compulsory public insurance and a supplementary insurance provided by private insurers. There is no universal coverage and the health system is also characterized by an important freedom for people to choose and use public and private health care services without a referral system. It is the patients who can access elective surgery without general practice referral. Public hospitals are important, but they are not the only ones: about 70 % of beds are in public hospitals in France and around 55 % in Germany, and they are funded through global budgets that are usually set annually during negotiations between the sickness funds. In Germany the budget specifies targets in terms of activity as well as per diems to be reached by the end of the financial year. The hospital specialist can be salaried or paid by fee for service. This latter option is applied generally in private hospitals and in treatment. The countries without waiting times usually apply low co-payments. Some examples of co-payments in 1999 were:

France	11 euros per day
Germany	7,16 euros per day
Belgium	250 FB per day
Luxembourg	5,43 euros per day

- The Netherlands represents an interesting case for analyzing waiting times because, despite the funding being based on a mix of public and private health insurance like France or Germany, waiting times for surgery are a significant problem. The reason could be the strong central control over the last two decades on total health expenditure fees.

2.2 The Factors That May Explain Waiting Time Differences Between European Countries

The National Health System seems to have more problems with surgery waiting lists. We want to know which factors influence this question. Waiting time may be determined by demand factors, which affect the inflow to the waiting list, and by supply factors, which affect the outflow.

2.2.1 The Demand for Elective Surgery Depends on

Health status of population. One of the main factors that could influence the health status is the share of the population that is elderly. In Europe, however, the population older than 65 is more or less the same in all countries. The age structure seems to be very similar in countries with and without waiting lists.

The proportion of population with private insurance, the price of private health insurance and the price of private surgery also influence public demand. In any case, there are feedback effects from prices to quantities demanded and supplied in private markets, so there are likely to be feedback effects from waiting times to quantities demanded and supplied in the public provision of elective surgery. For example: higher waiting time may encourage demand for private health (Besley et al. 1998). But also higher waiting times may discourage public demand for reducing referrals and deterring surgeons from adding patients to lists. At the same time, higher waiting times may raise supply by encouraging public authorities to allocate more money to public hospitals with longer queues (Gravelle et al. 2003).

Co-payments. Some financial measures, such as the extent of cost sharing, for example: co-payment can reduce demand. The countries with waiting times apply usually low co-payments but also countries without waiting times also have very low co-payments or no co-payments at all.

Doctors. It is also important to give a key role to the doctors in managing demand, the thresholds for referrals and addition to the list. General practitioners often act as gatekeepers in the countries with waiting times. The term gatekeeper suggests that the general practitioners should control the demand for access to specialists, avoiding unnecessary referrals. However, it may be that where there is a clear division of labor between generalists and specialists, general practitioners can consider that the best treatment is to pass on to surgeons any patients creating an upward pressure on demand. By contrast, where specialists can be approached directly by patients, they may become skilled at handling excess demand.

Most of the policies consider that waiting times can be reduced through supply-side policies because the volume of surgery is considered inadequate.

2.2.2 The Supply of Elective Surgery Depends on

Public resources, mainly beds and surgeons (see Table 9.24 of the Appendix). In this sense, some evidence on the impact of waiting on public capacity is provided by Lindsay and Feigenbaum (1984), Martin and Smith (1999) and Siciliani and Hurst (2003), who concluded that the waiting list is associated negatively with the number of beds and surgeons. A low endowment of acute beds may constitute a binding constraint for countries with waiting lists, limiting, in the short term, the opportunity to increase output. The number of doctors can also be essential on waiting times. However, authors like Siciliani and Hurst (2003) think that a larger staff is usually associated with lower waiting times if combined with other inputs. Also higher expenditure per capita is associated with a higher rate of surgery and with lower waiting time. Finland and England have low expenditure and report high waiting times. Countries such as Germany and the USA have high expenditures and do not report waiting times (see Tables 9.22 and 9.23 of the Appendix). However, Norway is a high-expenditure country and reports high waiting times. The middle expenditure countries like Denmark, the Netherlands, Austria, Belgium, or France are also equivocal, while the first two report waiting times, the others do not. Consequently health expenditure may be important on waiting times but it is not the only factor.

Other differences. Another possibility is that other differences may play a role. For example, incentives and remuneration systems may encourage higher productivity in countries without waiting times, but induce a high propensity to add patients to the list. Productivity depends, among other things, on the way in which surgeons and hospitals are paid. There are some studies that have investigated the relationship between methods of paying physicians and productivity. The results suggest that fee-for-service payment shows better rates than the salaried staff model. Incentives to hospitals seem to be effective. Activity-based funding is likely to encourage higher productivity compared to funding based on fixed budgets (Clemmesen and Hansen 2003). Some experiences support this idea (Spain, Norway, Denmark, etc). Productivity also depends on the percentage of patients treated by day-surgery. Martin and Smith (1999) show that the waiting time is negatively associated with the percentage of day-surgery cases and the elasticity is −0,252.

Consequently, waiting times may be explained by these variables: health expenditure per-capita, the number of practicing specialists and physicians, the number of acute care beds, and remuneration systems of hospitals and physicians.

3 The Major Policies Used to Reduce Waiting Time

The main ways used to reduce waiting times are provided below.

Maximum waiting time guarantee. One of the most common policies introduced to reduce waiting time is the maximum waiting-time guarantee. These guarantees are intended to regulate waiting times so that patients should never wait beyond a certain time limit. Almost every country with waiting lists has regulated these guarantees, but the formulation of the guarantee differs substantially across countries. An alternative is an unconditional guarantee that is provided to patients (England and Sweden between 1992 and 1995). Another alternative is a conditional guarantee that is given only to a number of patients, for example “all the patients with most need should be treated within ‘x’ months” (Norway 1990–2000, New Zealand and Sweden between 1992 and 1996). Another possibility of conditional guarantee is to regulate that a fixed percentage of patients should be treated within “x” months (The Netherlands, Italy, and Denmark).

A general point of criticism of a maximum waiting time guarantee is that it may be obtained at the expense of increasing the outpatient waiting time. In this sense some countries have also set maximum “outpatient waiting time guarantees”, such as Norway (between 1997 and 2000), the Netherlands, Sweden, and England. Also, as Siciliani and Hurst (2003) comment, the introduction of guarantees may produce conflicts between policy maker and surgical specialist, especially if they are not accompanied by extra resources. This and other problems have led to some countries, like Norway, to replace the maximum waiting time guarantee with a "right to necessary health care". In this case, the patient has the right to receive the treatment in an appropriate time and the waiting time for a specialist visit is 30 working days after the referral from the general practitioner.

Increasing productivity policies. There are a large number of policies whose main objective is to increase productivity. In many countries, public hospitals have been funded according to fixed budgets. In the Netherlands fixed budgets for specialists in replacement for service arrangements were piloted with bad results. Admission declined and waiting times increased. Several governments have tried to tackle high waiting times with extra funds. This fund has been tied to the achievement of waiting time objectives in several ways: (1) To raising the productivity of the hospitals in terms of number of treatments per surgeon or bed. (2) To fund hospitals which perform extra activity such as Sweden and the Netherlands.

Activity-based remuneration system. In Spain the concept of target population is important. It was introduced in 1996. The idea is to know what share of the target population could be treated in normal hours. In this way it is possible to calculate supplementary funding for the residual patient, who has not been treated in normal hours. The funds have been allocated to hospitals on the basis of the achievement of the different targets either in terms of the activity performance and achievement of maximum waiting time and mean time (monthly targets).

To encourage specialists to reduce waiting time to achieve maximum use of waiting time, Spain is using bonuses for specialists who have achieved waiting time reductions. In England from 2001, the "Performance Fund" has included rewards for staff (new equipment, improved facilities and cash incentives), for individuals and teams.

Increasing resources. An alternative to increasing capacity in the public sector is to use the existing capacity in the private sector. Usually in these cases this can take the form of a purchaser of health services contracting out to privately owned providers some volume of activity. This scheme presents some advantages: it may be the quickest way to increase capacity compared to other options. Second, contracting with private providers may introduce an element of competition with public providers.

Some countries, such as Norway, Denmark, Ireland, England, and the Netherlands have increased elective health surgery services by purchasing extra activity abroad. Usually in these countries the private sector may be fairly small and already working at maximum capacity.

Reducing waiting times by improving management of the waiting list. Australia has introduced an important system: pre-admission services, optimization of patient's health status prior to admission; education of the patient and family about hospital procedures; reducing cancellations and the number of unused sessions, and facilitation of day-surgery admissions.

Several governments (such as England) have taken steps to encourage day surgery. For example, England plans to introduce "Diagnostic and Treatment

Centres" to increase the number of elective operations that can be treated in a single day. These centers will focus on routine hospital surgery and not on hospital emergency work so they can concentrate on reducing waiting times.

Reducing average hospitalization time. As can be seen in the data that appear in Table 9.25, in most of the countries of the European Union the average hospitalization time has declined progressively in the last 30 years. For example, in the UK the average hospitalization time was 25.7 days in 1970, 15.6 days in 1990, 10.2 days in 1998, and 8.1 in 2002. In Denmark the evolution of the average hospitalization time went from 18.1 days in 1970 to 8.2 in 1990, 6.7 in 1998, and 5.7 in 2002. See Appendix, Table 9.25.

Increase patient choice. England, Denmark, Norway, and Sweden have recently introduced more choice for patients often in conjunction with activity-based payment. It is possible that this kind of measure may encourage hospitals to compete for patients and revenues. A prerequisite of this type of policy is the dissemination of information on waiting times. In Denmark patients have had free choice of treatment in any publicly funded hospital. However, it has been estimated that only 5 % of the patients exercised this right.

Reduce demand by subsidizing voluntary "private health insurance". The main idea is that by lowering the price for private insurance, many citizens will be induced to purchase private health. However, this simple reasoning should be qualified. First, the substitution effect is likely to be strong when some dimension of the quality of the public provision is low (for example long waits) and it is the feature that induces a shift of patients from public to private; reducing waiting times may decrease the incentive for the population to buy voluntary private health insurance (feedback effect). Second, waiting times are very important in this decision, but, in fact, there is some evidence that suggests that other factors like age, income, and political affiliation are important. Third, if private hospitals have no ability to expand in the short or medium term to respond to increases in demand, due to markets access regulations or to shortages of capacity or medical workforce, the expected reductions in waiting times may be delayed.

4 Waiting Time and the Public and Private Sector

The concept of a private sector has several dimensions. One dimension is related to the ownership of the means of production. Another dimension is whether services are paid by the client or covered by the public sector. Generally it is considered as private provision if it is private in both aspects.

There are several possibilities as to how the existence of a queue might reduce the flow of the demand for health treatment, one of them is that the longer the waiting time, the more people choose the private alternative. In this sense, it could say the waiting time is an equilibrating mechanism making the demand for public treatment equal the supply. Theoretical models like the one developed by Besley and Coate (1991), suggest that the determinants of the demand for private health insurance must be investigated as a function of the quality of public sector provision and the individual characteristics, especially income. This theory has been contrasted by Besley et al. (1999) with an empirical specification in two stages. The first model that they consider does not differentiate between sources of health insurance. This was legitimate provided that the workers face the full cost of purchasing insurance or when the employers purchase insurance on their behalf. The general results of the model revealed that health insurance demand rises with age, tailing off for those older than age 65, it also depends strongly on income. Larger households are less likely to buy insurance, probably reflecting the equivalent income effect. Finally, and one of the most important factors is the long-term waiting list. The results suggest that if the long-term waiting list were to rise by one person per thousand, then there would be a 2 % increase in the probability that an individual with the same characteristics would buy private insurance. The length of waiting lists is used as an indicator of health system quality and appears as one of the main factors demanding private health. This assumes that individuals know the length of the waiting list. This information is used as a barometer for the performance of the NHS.

It is supposed that the choice between public and private health services is an aspect with many consequences that have been studied by several authors. One consequence is that the public health system should improve because it would have fewer patients. However, a long-term system could go in the opposite direction, if there are feedback effects from private insurance demand to waiting lists through the political process. If the lobbying pressure to keep waiting times short declines in areas where there is a large privately insured segment, then this could lead to a positive correlation between private insurance and waiting lists. Iversen (1997) concluded that the effects on waiting time from the private causes are rather indeterminate. He developed a model with a long-term perspective: first the queuing model solution was applied; second, an elastic supply of health personnel was assumed.

The main conclusions can be summarized as follows: If the waiting times are not rationed, the effect of the private sector on a public hospital's waiting time is in general indeterminate. If the demand for public treatment is more elastic with respect to waiting time, then it is more likely that the private sector causes a longer waiting time for public patients. The reason is that a more elastic demand for public treatment makes possible a large reduction in public expenditure by increasing the waiting time (Iversen 1997). When the admission to waiting time list is rationed, the waiting time will increase if the public sector consultants are permitted to work in the private sector (Iversen 1997). Otherwise, waiting time will not change. Aaron and Schwartz (1984) also say the same. The private option motivates the consultants to reduce their work effort in the public sector and some patients probably choose the private sector. Other empirical studies came to the same conclusion. For example, in England Besley et al. (1998) have investigated the extent to which areas with high health insurance coverage had low waiting lists. Results were unexpected in the sense that areas with high private coverage had higher waiting lists. The authors suggest the government may under-fund public services in areas with private insurance coverage.

The choice of public or private sector also could have distributional implications. Individuals who opt out of public sector treatment free up resources for those who continue to rely exclusively on the National Health System. Assuming a fixed budget, the provision of public health should improve. However, as Besley et al. (1999) say, a significant fraction of the gain in any increased resources devoted to the National Health System could be taken by high income individuals who will choose to remove their private insurance coverage. Also, even high income individuals who are privately insured will continue to use the public service.

Hoel and Saether (2003) in the same line as Besley and Coate (1991) use a model describing the choice between treatment in the public and private sector. They use the model within a framework of standard welfare theory and the most important result is that if distributional objectives (equity) are sufficiently strong, it may be optimal to have waiting time for public treatment. There is a self-selection mechanism that gives the desired results, because the high-income persons choose to buy health in the private sector.

Another important issue in a system with predominantly public health care is how the government should treat the alternative private treatment. It is sometimes argued that the private alternative may undermine the public system. So the government ought to discourage any private alternative. There are different ways for discouraging the private alternative, such as regulation or taxes. Against this position one could think that those who choose the private alternative should be subsidized by the public sector. Cullis and Jones (1995) say that the argument above for subsidizing private health was based on fairness. This alternative is interesting from the point of view of the costs, because they are lower. The cost saved could be used to expand the treatment capacity of the public system. Australia has been the most active country in subsidizing voluntary private health insurance. For example, several policies have been included in the "1997 and 1998 private Health Insurers incentive schemes" and in 2000 "the lifetime health cover", which introduced tax rebates. As a result, the percentage of population covered by private health increased sharply from 30.5 % in 1999 to 44.1 % in 2002. The effects of these incentives still are unknown.

5 Setting Priorities for Waiting Lists

Several countries—Denmark, Finland, the Netherlands, Norway, Portugal, Spain, Sweden, New Zealand, Canada, and the USA—have introduced a profound debate in order to establish standardized prioritization rules for patients' admission to health waiting lists and in this sense, several actions have been carried out and many works have been developed in this field because it is a sensitive one (see Rodríguez Sendín 2000; Ortún-Rubio et al. 2001; Noseworthy et al. 2003).

Traditionally, the usual prioritization rule for nonurgent patients was "first in, first out" but now new factors, such as doctors' opinions, are taken into account for the management of waiting lists in most of the developed countries. In this new context the necessity for the introduction of different standardized prioritization rules has been pointed out in order to preserve equity.

In principle, patients with more urgent conditions should receive services ahead of those with less urgent conditions, and patients with approximately the same degree of urgency should wait about the same length of time. Nevertheless, as has been stated by Hadorn et al. in 2000, standardized measures to assess patients' relative priority are needed.

Hadorn (2000) proposes several key concepts underlying the development of criteria for assessing patients' relative priority on waiting lists. These concepts are:

- *Severity*: the degree or extent of suffering, limits to activities or risk of death. The more a patient is suffering, the more severe is his/her condition, other factors being equal. But how can severity be measured? How can we measure and compare pain and suffering? These last questions show how severity cannot be the main prioritization criteria in health services.
- *Urgency*: extent to which immediate clinical action is required. Usually, in elective surgery severe cases are considered as urgencies but these two criteria can diverge in other situations, such as in the setting of many terminal conditions when there is no pain and no intervention is available to forestall death, in the presence of patients with a low or middle level of severity but which if not treated might become more severe. Therefore, urgency may be, in Hadorn's opinion, defined as severity in addition to considerations of the expected benefit and the natural history of the condition.
- *Need*: urgency.
- *Expected benefit*: extent to which desired outcomes are likely to exceed undesired outcomes.

Other key concepts appearing in the literature referring to the development of prioritization rules are the effectiveness and cost-efficiency of the treatment and some patients' social characteristics (employment status, for instance, is taken into account in some health systems).

The Council of Europe (1998) has published several recommendations related to the criteria for the management of waiting lists and waiting times in health care. Among them, there are several criteria for admitting patients to waiting lists which mainly coincide with the ones proposed by Hadorn in 2000. But although priority is recommended to be given to patients with the greatest need for the services, "(...) waiting times should not be so long that the patients' health is at risk of deterioration". In this sense, acceptable waiting times have to be determined transparently trying if it is possible, to respect patients' preferences and principally patient necessities; but this is a key point because both concepts are difficult to define.

What seems to be clear in most of the public health systems is that the need and urgency for treatment should not be established on the basis of race, sex or religion of the patient. But unlike what is happening in some health systems (see Kee et al. 1998), the Council of Europe does not recommend prioritizing patients on the basis of their socioeconomic status or in general based on their age, although it could be taken into account as "an aspect of a patients' general medical condition and as a risk factor for particular treatments". The European Systems use a two-level to four-level classification system. In Spain high priority and low priority; in Sweden very urgent, urgent, and nonurgent; and in Italy admission within 30 days, within 90 days and 12 months.

Other kinds of prioritization have been considered in non-European countries. New Zealand recognized that the public resources are limited and it has been decided that patients on the waiting list should be prioritized according to need, and the public treatment is only provided for patients with the greatest need. One prerequisite is necessary to implement this policy, the introduction of guidelines to prioritize patients. These guides may also serve to pursue an efficiency goal.

6 A Model of Optimal Management

6.1 Introduction

In this section we will study the initial challenge that this research team faced some time ago, which would be the starting point for our line of work. The long stay of patients on waiting lists for surgery in public hospitals is a problem that worries the health authorities, public opinion and the professionals of health care in Spain. The health authorities of the different public administrations in our country have been taking steps and establishing requirement levels increasing over time. Specifically, in the year 1998, the maximum limit of stay of a patient on a waiting list for an operation changed from 9 to 6 months, implemented as follows: (1) For surgeries before the first of July of 1998, the maximum limit of stay was 9 months. (2) Patients who entered a waiting list before the first of July of 1998, had to be off the list before the first of January of 1999; (3) Patients who entered a waiting list after the first of July of 1998 had 6 months as maximum limit of stay on the list.

By the end of 1997 a specific public hospital of Madrid had long waiting lists for the following surgical processes: Cataracts, Hallux Valgus, Knee Operations and Osteoarthritis. The first one depends on the hospital service of Ophthalmology and the rest on the service of Orthopedic Surgery and Traumatology. The authorities of the hospital, knowing the situation of the waiting lists for these four surgical processes and in the light of the new and larger requirements about the new limits of stay to be reached, were very worried. The four surgical processes had long waiting lists not only in the hospital we are considering but in all the public hospitals of Madrid. In fact, they were first place in the accumulation of patients, using aggregated data in Madrid.

When the Operations Research team went to work in 1998, the hospital had previously established its agreements for the year both at an internal level, with the different services, and at an external level with the Spanish National Health Service (SNHS) (INSALUD, in Spanish). In our Hospital, according to the decision maker, the bottleneck was neither in human resources nor in the number of beds available (at least initially), but in the operating rooms. The number of operating rooms in the hospital was appropriate from the point of view of the relationship with the rest of the facilities and teams.

The challenge for all hospitals involves getting waiting times down while maintaining costs within certain limits (Bitran and Valor-Sabatier 1987; Chae et al. 1985). To attain the previously mentioned objectives, hospitals are allowed to use several methods of operating scheduling: (1) Within regular-operating hours, (2) Overtime, and (3) Private hospital contracts. Specifically, in our Hospital, in accordance with previously established agreements:

- Cataract surgeries could be in regular operating hours or in overtime, but private hospital contracts were not allowed (Methods 1 and 2, but not 3).
- Hallux valgus: Methods 1 and 3 were possible, but not 2.
- Knee operations: Methods 1 and 3 were possible, but not 2.
- Osteoarthritis: Method 1 was possible, but not 2 and 3.

The problem consisted of deciding how many operations can be performed in every month of the year 1998 for each of the four types, in regular time, overtime and private hospitals under contract, in such a way that all the constraints (to be introduced) are satisfied and the objective function (to be introduced) is optimized. Therefore, we have a problem of annual planning, to be solved with mathematical optimization. Before introducing the mathematical program we will point out several general features of the hospital in which the study was done.

6.2 General Features of the Hospital

The data we consider correspond to the first of January of 1998, when the problem was solved. The general characteristics of the hospital are provided below, and in Tables 9.3 and 9.4.

- The Hospital is located in Madrid.
- Population dependent on the Hospital: 305,000 people.
- Number of beds available: 407.
- Number of operating rooms for planned operations: 8.
- Number of consulting rooms: 70.

6.3 The Mathematical Problem to be Solved

In this section a mathematical programming problem is formulated in order to plan the surgical activity of the hospital for the four processes considered. First the decision variables will be defined, then the initial relevant data used in our work will be presented, the constraints and the objective functions will be defined and, finally, from the previous subsections, the mathematical problem to be solved will be defined.

Table 9.3 Structure of the Hospital in 1997

	Hospital	Global TSNHS
Beds/1,000 inhabitants	1.34	2.67
Human resources/bed	3.62	2.96
Doctors/bed	0.62	0.48
Operating rooms/100,000 inh.	3.5	5.9
Surgeons/operating rooms	8.5	7.7
Anesthetists/operating rooms	1.4	1.4

Table 9.4 Activity in 1997

Activity	Number
Admissions	14,518
Stays	108,109
First examinations	132,821
Successive examinations	237,426
Total examinations	370,247
Urgencies	98,539
Planned surgical operations	3,111
Urgent surgical operations	2,061
Surgical operation without stay for the night	4,935
Total surgical operations	10,107

We need to introduce considerable notation. Let

CL_1: Number of patients on the waiting list for cataracts the Jan, 1, 1998.
HL_1: The same for hallux valgus.
KL_1: The same for knee operations.
OL_1: The same for osteoarthritis.

These quantities are known for January 1, 1998.

6.4 Variables

Let us define the following state variables:

CL_k: Number of patients on the waiting list for cataracts the first day of month k.
HL_k: The same for hallux valgus.
KL_k: The same for knee operations.
OL_k: The same for osteoarthritis,

for $k = 2, 3, \ldots, 12, 13$, where $k = 2$ corresponds to February, 1998, $k = 3$ to March, 1998, …, $k = 12$ corresponds to December, 1998, and $k = 13$ to January, 1999.

Let us define the following control variables:

CR_i: Number of cataract operations to be carried out in month i in regular-operation hours.
HR_i: The same for hallux valgus.
KR_i: The same for knee operations.
OR_i: The same for osteoarthritis.
CO_i: Number of cataract operations to be carried out in month i in overtime.
HP_i: Number of hallux valgus operations to be carried out in month i through private hospital contracts.
KP_i: The same for knee operations.

for $i = 1, 2, \ldots, 12$, where $i = 1$ corresponds to January of 1998, $i = 2$ to February of 1998, …, $i = 12$ to December of 1998.

Any month $i \in \{1, 2, \ldots, 12\}$ starts with a number of patients on the waiting list for each one of the four surgical processes CL_i, HL_i, KL_i, OL_i (using the terminology of Dynamic Optimization, these are the state variables, including then $i = 13$, but excluding $i = 1$ which are given). In month i the following operations will be carried out: CR_i, CO_i, HR_i, HP_i, KR_i, KP_i, and OR_i (the control variables). Moreover, during the month i new patients will come on the waiting lists and some patients will leave the waiting lists without an operation for some reason (the forecast is among the data in the next subsection). The month i will finish (or equivalently the month $i + 1$ will begin) with the numbers of the waiting lists given by CL_{i+1}, HL_{i+1}, KL_{i+1}, and OL_{i+1}. Therefore, we have a problem with the following number of variables:

- State variables: $4 \times 12 = 48$.
- Control variables: $7 \times 12 = 84$.
- Total number of variables: 132.

6.4.1 Data

In this subsection the relevant data for the problem are incorporated. We present the data in the way they were given by the Hospital. Table 9.5 contains, for each pathology, the number of patients on the waiting list by December, 31, 1997. Table 9.6 shows the month in which patients enter the waiting lists.

Table 9.7 contains the entries on the waiting lists estimated by the hospital for each month of 1998.

Table 9.8 contains the exclusions in the waiting lists (without an operation) estimated by the hospital for each month of 1998.

Table 9.9 contains the distribution of surgical sessions for the year 1998 (within regular-operating hours).

Table 9.10 contains the operating room time necessary for each operation. Is the time that elapses from the moment when the patient enters the operating room until the moment when he or she leaves.

Table 9.11 contains the number of operations and the total time of operating room used in the hospital in 1997, for each process under study.

Table 9.5 Patients on waiting list by 31 December 1997

Cataracts	480
Hallux valgus	199
Knee operations	132
Osteoarthritis	128
Rest of processes of traumatology	511
Rest of processes of ophthalmology	97

Table 9.6 Month of entrance on the waiting lists

	Month of 1997									
Process	April	May	June	July	August	September	October	November	December	Total
Catar.	11	15	24	66	37	71	85	89	82	480
H.V.	9	15	38	30	11	27	23	12	34	199
K.O.	4	19	12	14	10	23	18	19	13	132
Ost.	3	14	17	4	4	19	33	13	21	128
Rest T.	31	56	61	53	26	62	69	67	86	511
Rest O.	2	10	12	8	0	16	21	16	12	97

Table 9.7 Estimation of entries on the waiting lists

	Month of 1998												
Pro.	1	2	3	4	5	6	7	8	9	10	11	12	Total
C.	84	85	82	94	78	104	125	42	78	98	94	86	1,050
H.V.	28	28	22	22	34	45	31	12	20	24	12	33	311
K.O.	21	22	18	15	30	18	15	12	24	18	21	13	227
O.	10	22	15	14	30	24	5	5	17	34	14	21	211
R.T.	130	145	120	122	159	169	116	65	151	162	122	57	1,618
R.O.	111	113	100	112	107	139	144	50	102	137	119	104	1,338

Table 9.8 Estimation of exclusions

	Month of 1998												
Pro.	1	2	3	4	5	6	7	8	9	10	11	12	Total
C.	8	13	16	16	20	37	53	12	19	20	17	7	238
H.V.	4	8	13	6	10	31	22	3	5	12	2	19	135
K.O.	3	5	4	3	10	14	4	0	7	9	1	5	65
O.	5	2	9	7	9	7	7	5	7	13	2	5	78
R.T.	26	37	56	58	42	72	105	17	61	75	26	61	636
R.O.	15	17	20	19	31	43	58	13	27	31	24	10	308

Table 9.9 Distribution of surgical sessions

	Month											
Service	1	2	3	4	5	6	7	8	9	10	11	12
Traum.	26	27	28	24	27	28	21	19	18	31	27	24
Ophtha.	19	20	20	17	20	20	9	11	9	23	20	16

Table 9.10 Operating room time

Process	Time (min)
Cataracts	60
Hallux valgus	65
Knee operations	100
Osteoarthritis	140

Table 9.11 Number of operations and total time of operating room

Process	Number of operations	Total time (min)
Cataracts	450	26,557
Hallux valgus	27	1,795
Knee operations	68	6,868
Osteoarthritis	105	15,173
Total traumatology	1,018	102,506
Total ophthalmology	904	54,749

Table 9.12 Maximum number of possible operations

	Month											
	1	2	3	4	5	6	7	8	9	10	11	12
C.o	0	0	68	40	64	72	0	0	44	52	48	24
H.p	0	20	25	35	35	35	35	35	35	35	35	35
K.p	0	0	8	21	21	20	20	20	20	10	10	0

C.o: Cataracts overtime. H.p: Hallux valgus in private hospital contracts. K.p: Knee operations in private hospital contracts

We have to point out that some of these operations were not on any surgical waiting list, being urgent operations but carried out in programmed operation rooms. Table 9.12 contains the maximum number of possible operations in each month, for each one of the possibilities out of the regular-operating hours.

These maximum limits correspond to previously established agreements (with financing guaranteed) between the management of the hospital and the service of ophthalmology of the hospital (for cataracts in overtime) and the regional health direction of the SNHS (for hallux valgus and knee operations in private hospital contracts).

Table 9.13 contains the costs for this hospital, for each process, both in regular time and in overtime. The hospitalization costs are included.

Table 9.13 Costs for the hospital

Process	Cost in regular time	Cost in overtime
Cataracts	110,852 pesetas	123,733 pesetas
Hallux valgus	125,899 pesetas	138,781 pesetas
Knee Operations	287,338 pesetas	313,273 pesetas
Osteoarthritis	853,338 pesetas	887,071 pesetas

Table 9.14 Rates of the SNHS for 1997

Process	Rates for the SNHS for 1997
Cataracts	146,971 pesetas
Hallux valgus	106,605 pesetas
Knee operations	141,120 pesetas
Osteoarthritis	925,000 pesetas

The costs of the processes to be carried out in private hospitals with contracts are considered maximum rates for 1997 (rates of the SNHS for 1997), in accordance with the order of the government published the day May, 8, 1997. These are presented in Table 9.14.

6.4.2 Constraints

State Equations

For each one of the surgical processes under study, the number of patients on the waiting list on the first day of month $i + 1$ is equal to the number of patients that were on the list on the first day of month i plus those who came on the list during the month i minus those excluded from the list without operation during the month i, minus those operated (in regular time, overtime or in a private hospital under contract), during the month i. That is, for $i = 1, 2, \ldots, 12$,

$$\begin{aligned}
\mathrm{CL}_{i+1} &= \mathrm{CL}_i + \mathrm{CA}_i - \mathrm{CE}_i - \mathrm{CR}_i - \mathrm{CO}_i,\\
\mathrm{HL}_{i+1} &= \mathrm{HE}_i + \mathrm{HA}_i - \mathrm{HE}_i - \mathrm{HR}_i - \mathrm{HP}_i,\\
\mathrm{KL}_{i+1} &= \mathrm{KL}_i + \mathrm{KA}_i - \mathrm{KE}_i - \mathrm{KR}_i - \mathrm{KP}_i,\\
\mathrm{OL}_{i+1} &= \mathrm{OL}_i + \mathrm{OA}_i - \mathrm{OE}_i - \mathrm{CR}_i,
\end{aligned}$$

subject to the following initial conditions:

$$\mathrm{CL}_1 = 480,\ \mathrm{HL}_1 = 199,\ \mathrm{KL}_1 = 132,\ \mathrm{OL}_1 = 128.$$

Operating Rooms Allocated to Each Service

Ophthalmology

$$80\,\mathrm{CR}_i \leq \mathrm{OCQ}_i \text{ for } i = 1, 2, \ldots, 12.$$

The previous inequality shows that for each month the number of minutes of operating room needed to operate cataracts in regular time has to be smaller or equal to the number of minutes of operating room available to operate cataracts. It is

Table 9.15 Parameter values

Month	OCQ_i	TEQ_i
$i = 1$	5,520	3,255
$i = 2$	5,840	3,392
$i = 3$	5,920	3,486
$i = 4$	4,880	2,982
$i = 5$	5,840	3,392
$i = 6$	5,840	3,486
$i = 7$	2,560	2,572
$i = 8$	3,040	2,384
$i = 9$	2,400	2,247
$i = 10$	6,720	3,892
$i = 11$	5,840	3,434
$i = 12$	4,560	3,024

assumed that each operation requires 80 min (60 min of operation plus 20 min to clean the operating room). The amount of time available in each month for operations of cataracts in regular time (OCQ_i) is collected in Table 9.15.

Traumatology

$$85\ \text{HR}_i + 120\ \text{KR}_i + 160\ \text{OR}_i \leq \text{TEQ}_i \text{ for } i = 1, 2, \ldots, 12.$$

The previous inequality shows for each month the operating room time (in minutes) needed to operate on hallux valgus, knee operations and osteoarthritis in regular time (where to the required time for each operation 20 min have been added for the cleaning of the operating room) has to be smaller or equal to the operating room time that the traumatology service of the hospital has to carry out these three types of surgical processes (TEQ_i). The values of OCQ_i and TEQ_i appear in Table 9.15.

In the Table 9.15, time (the numbers that appear in columns 2 and 3) is expressed in minutes. It is very important to explain how the values that appear in Table 9.15 have been obtained.

Ophthalmology: Let us see how the values OCQ_i have been obtained. Every surgical session is assumed to (theoretically from 8 to 15 h) last 6 h and a half, which is more realistic than the seven theoretical hours. In this way, as Ophthalmology has for example 19 sessions in January, it has initially 19 × 390 = 7,410 min of operating room in January. In the same way the initial available minutes of operating room are worked out for each month of the year. The time contained in some of these sessions coincides with time reserved for scientific sessions and during this time operations are not carried out. Therefore, from the previously obtained minutes, it is necessary to subtract the time devoted to scientific sessions. We have been checking day-by-day for each month, ensuring that we have to subtract 480 min in January, February, April, May, June, August, September, October, November, and December, 360 min in March and 240 min in July. From the remaining time we have deduced that 80 % of time is devoted to cataract operations, thus obtaining the values of OCQ_i.

Traumatology: Let us see how the values TEQ_i have been obtained. In the same way as for the other service, we start multiplying for each month the number of sessions for Traumatology by 390 min (which correspond to 6 h and 30 min) and subtracting the minutes devoted to scientific sessions, which correspond to surgical sessions of Traumatology. Specifically, we have to subtract 840 min in January, February, April, May, and July; 960 in March, June and October; 600 in August and September; and 720 in November and December. From this time we have to keep the 70 % that the service devotes to programmed operations that are on the waiting list (the remaining 30 % is devoted to "delayed urgencies"). From the remaining time we have deduced that it is necessary to give 50 % to these three traumatology operations, thus obtaining the values of TEQ_i.

If it were possible to pool the operating room times devoted to Cataracts and the set of hallux valgus plus knee operations plus osteoarthritis, the following constraint would substitute the constraints in Sect. 6.4.2.2:

$80\ CR_i + 85\ HR_i + 120\ KR_i + 160\ OR_i \leq OCQ_i + TEQ_i$ for $i = 1, 2, \ldots, 12$.

Limits to the Number of Processes in Private and Overtime Scheduling

The following constraints have to be satisfied:

$$\begin{aligned} CO_i &\leq l_i, \\ HP_i &\leq m_i, \\ KP_i &\leq n_i, \end{aligned}$$

for $i = 1, 2, \ldots, 12$, where the values for l_i, m_i, and n_i, appear in Table 9.16.

Waiting List Time Upper Limit: No More Than 9 Months

With the following constraints we reflect that throughout the year the maximum time for patients to be on the waiting list should be 9 months.

$$\begin{aligned} &(CR_1 + CO_1) + (CR_2 + CO_2) + \ldots + (CR_k + CO_k) \geq a_k, \\ &(HR_1 + HP_1) + (HR_2 + HP_2) + \ldots + (HR_k + HP_k) \geq b_k, \\ &(KR_1 + KP_1) + (KR_2 + KP_2) + \ldots + (KR_k + KP_k) \geq c_k, \\ &OR_1 + OR_2 + \ldots + OR_k \geq d_k, \end{aligned}$$

for $k = 1, 2, \ldots, 12$, where the values for a_k, b_k, c_k and d_k appear in Table 9.17.

The values contained in Table 9.17 are constructed from Table 9.6. The meaning of these values is the following: in January, at least the 11 patients that entered the waiting list for cataracts in April on 1997 (and are on the waiting list on January 1) have to be operated on. In the same way, the nine patients that entered the waiting list for hallux valgus in April, 1997, have to be operated on January. The same for the four patients that entered the list of those needing knee operations and the three patients that entered the list for osteoarthritis in April, 1997. In February, at least the

Table 9.16 Limits to the number of processes

	1	2	3	4	5	6	7	8	9	10	11	12	Total
l_i	0	0	68	40	64	72	0	0	44	52	48	24	412
m_i	0	20	25	35	35	35	35	35	35	35	35	35	360
n_i	0	0	8	21	21	20	20	20	20	10	10	0	150

Table 9.17 Waiting time upper limit: no more than 9 months

	1	2	3	4	5	6	7	8	9	10	11	12
a_k	11	26	50	116	153	224	309	398	480	556	628	694
b_k	9	24	62	92	103	130	153	165	199	223	243	252
c_k	4	23	35	49	59	82	100	119	132	150	167	181
d_k	3	17	34	38	42	61	94	107	128	133	153	159

15 patients that entered the waiting list for cataracts in May, 1997, and have not been operated on in January. That is the reason why the number of cataract operations in January and February has to be greater than or equal to the number of patients who entered the waiting list for cataracts in April and May, which is equal to 26. Reasoning in this way, the values that appear in Table 9.17 are obtained.

No More Than 6 Months Waiting at the End of 1998

The following constraints have to be satisfied:

$$\begin{aligned} \mathrm{CL}_{13} &\leq 395, \\ \mathrm{HL}_{13} &\leq 69, \\ \mathrm{KL}_{13} &\leq 77, \\ \mathrm{OL}_{13} &\leq 57. \end{aligned}$$

These values are obtained from the addition of the estimated entries minus the estimated exits without operations corresponding to the last 6 months of 1998, Tables 9.7 and 9.8.

All the Variables Have to be Non-negative Integers

6.4.3 Objective Functions

The problem has two objective functions.

First objective. Minimize the waiting list pending by the end of 1998 (measured in operating room time).

$$\mathrm{Min}\, f_1 = 80\ \mathrm{CL}_{13} + 85\ \mathrm{HL}_{13} + 120\ \mathrm{KL}_{13} + 160\ \mathrm{OL}_{13}.$$

It is clear that a fundamental objective of the planning to be done is to leave the waiting list by the end of the year as small as possible. Specifically, the objective function to minimize is the operating room time (in minutes) that remains on the waiting list by the end of the year we are planning.

Second objective. Minimize costs.

Min f_2 = 110,852 (CR_1 + . . . + CR_{12}) + 125,899 (HR_1 + . . . + HR_{12}) + + 287,973 (KR_1 + . . . + KR_{12}) + 853,338 (OR_1 + . . . + OR_{12}) + + 123,733 (CO_1 + . . . + CO_{12}) + 106,605 (HP_1 + . . . + HP_{12}) + + 141,120 (KP_1 + . . . + KP_{12}) + 90,584 CL_{13} + 58,035 HL_{13} + + 148,157 KL_{13} + 537,603 OL_{13}.

Let us explain how this second objective function was obtained. We have to add the costs of the operations carried out in the hospital in regular time (the unit values of which appear in Table 9.13), the costs of the cataract operations carried out in the hospital in overtime (the unit values also appear in Table 9.13), the costs of the operations carried out in private hospitals with contracts (for which it has been assumed that the unit values are those given in Table 9.14) and an assessment, in costs terms, of the operations pending for the next year, where for each process an expected approximated unit cost has been introduced, bearing in mind that some entries on the list will leave the list without an operation and that there are several possibilities for operation (regular time, overtime and private contracts).

6.4.4 The Mathematical Program

In previous sections the elements of a mathematical program have been introduced. It is a program with two objectives and several constraints, where the decision variables are integer. Among the different possibilities to broach the bi-objective program, one of the more immediate is to ask the decision maker (in this case, the manager of the hospital) if it was possible to give weighting to the objectives in accordance to their importance from the hospital's perspective. The reply of the decision maker was emphatic: give a weighting of 0.8 to the first objective (to minimize the waiting list by the end of 1998) and a weighting of 0.2 to the second one (to minimize costs).

After the introduction of the usual technical adjustments in multiobjective programming, specifically

$$0.8\left(f_1/\left|f_1^* - f_{*1}\right|\right) + 0.2\left(f_2/\left|f_2^* - f_{*2}\right|\right),$$

where f_1^* and f_2^* are the ideal of the first and second objective and f_{*1} and f_{*2} are the anti-ideal of the first and second objective, respectively,

$$\begin{aligned} f_1^* &= 34,379, & f_{*1} &= 55825, \\ f_2^* &= 431,561,300, & f_{*2} &= 462,946,208. \end{aligned}$$

We have the following objective function:

Min $0.0706\ (CR_1 + \ldots + CR_{12}) + 0.0802\ (HR_1 + \ldots + HR_{12}) + + 0.1835\ (KR_1 + \ldots + KR_{12}) + 0.05437\ (OR_1 + \ldots + OR_{12}) + + 0.0788\ (CO_3 + \ldots + CO_6 + CO_9 + \ldots + CO_{12}) + 0.0679\ (HP_2 + \ldots + HP_{12}) + + 0.0899\ (KP_3 + \ldots + KP_{12}) + 0.3561\ CL_{13} + 0.3539\ HL_{13} + + 0.5420\ KL_{13} + 0.9393\ OL_{13}.$

The problem is subject to the constraints given in Sect. 6.4.2. It is a linear program with integer variables, 132 variables and 160 constraints (apart from the non-negativity of the variables).

6.4.5 Results

The program HIPERLINDO has been used to solve the problem. We introduced the data of the mathematical programming problem in HIPERLINDO, and found that no feasible solution exists. That is, it is not possible to satisfy all the requirements of maximum limit of permanence on the waiting list with the resources of the hospital, with the established agreements and with the hospital's usual way of working.

After the analysis of the problem and its solution we found that the problem has an optimal solution if the following constraint is removed,

$$OL_{13} \leq 57.$$

Therefore, the hardest constraint is the maximum limit of 6 months of permanence on the waiting list for osteoarthritis, which occupies the most operating room time and is the most expensive; which cannot be done in overtime and which cannot be sent to a privately contracted hospital (as had been previously decided in the agreements of the hospital). We have studied the problem without that constraint, obtaining the minimum limit for OL_{13} in order to assure feasibility. All requirements can then be satisfied if it is possible to renegotiate, and if it is possible to send 30 processes of osteoarthritis to privately contracted hospitals.

A second possibility consists of transferring some planned operating room sessions from ophthalmology to traumatology. In that case, an optimal solution is obtained, satisfying all the constraints, and the hospital could satisfy all the requirements without the necessity of renegotiating with the Spanish National Health Service to send 30 osteoarthritis processes to privately contracted hospitals. Specifically, 13 operating room sessions initially allocated to ophthalmology should be allocated to traumatology in the following way: 1 in January, 1 in February, 1 in March, 2 in May, 1 in June, 1 in July, 1 in August, 2 in September, 2 in October, and 1 in November.

The results obtained with the first option, to have the Spanish National Health Service finance 30 operations of osteoarthritis in private hospitals, appear in Table 9.18.

In this situation the evolution of the waiting list is as recorded in Table 9.19, where for each process there appears the number of patients on the waiting list for

Table 9.18 Results obtained with the first option

	Month											
Process	1	2	3	4	5	6	7	8	9	10	11	12
Cat RT	69	73	74	61	73	73	32	38	30	84	73	57
CatOT	0	0	68	40	64	72	0	0	44	52	48	24
HV RT	10	1	6	0	0	0	0	0	0	0	0	0
HV PC	0	20	25	35	35	35	35	35	35	35	35	35
KO RT	15	10	2	2	0	1	8	1	0	23	0	5
KO PC	0	0	8	21	21	20	20	20	20	10	10	0
ORT	4	13	17	17	21	21	10	14	14	7	21	15

Table 9.19 Evolution of the waiting list with the first option

	Month											
Process	2	3	4	5	6	7	8	9	10	11	12	13
Cataracts	487	486	410	387	308	230	270	262	247	189	145	143
H.V.	213	212	190	171	160	139	113	87	69	46	21	0
K.O.	135	142	146	135	134	117	100	91	88	64	74	77
Osteoar.	129	136	125	115	115	111	99	85	81	95	86	87

Table 9.20 Results obtained with the second option

	Month											
Process	1	2	3	4	5	6	7	8	9	10	11	12
Cat RT	65	69	70	61	65	69	28	34	22	76	69	57
CatOT	0	0	68	40	64	72	0	0	44	52	48	24
HV RT	10	0	7	0	0	0	0	0	0	0	0	0
HV PC	0	20	25	35	33	35	35	35	35	35	35	35
KORT	5	18	14	4	0	4	2	19	1	0	1	1
KO PC	0	0	8	21	21	20	20	20	20	10	10	0
ORT	14	10	10	15	26	21	17	3	18	29	23	18

the first day of the corresponding month. The optimal value of the objective function is equal to 404.9.

With the second option (rearranging the allocation of operating rooms to the different services of the hospital), the results are recorded in Table 9.20.

In this situation the evolution of the waiting list is as recorded in Table 9.21, where for each process there appears the number of patients on the waiting list for the first day of the corresponding month. The optimal value of the objective function is equal to 407.2.

Table 9.21 Evolution of the waiting list with the second option

	Month											
Process	2	3	4	5	6	7	8	9	10	11	12	13
Cataracts	491	494	422	399	328	254	298	294	287	23	197	195
H.V.	213	213	190	171	160	141	115	89	69	46	21	0
K.O.	145	144	136	123	122	102	91	64	60	59	68	75
Osteoar.	119	129	125	117	112	108	89	86	78	70	59	57

7 Conclusions and Future Directions

The revision of the health systems in operation in Europe permits us to affirm that the problem of waiting lists is usual in countries that have a National Health System. However, these systems also have important advantages. For example, they permit good control of costs; they are less expensive than the systems based on health insurance; and they permit high levels of coverage in all benefits. In the European Union the countries that apply this kind of health system, mainly characterized by supplying universal coverage and obtaining financing via taxes, are increasingly numerous. To England and the Northern Countries have been added little by little some countries from Southern Europe, such as Italy, Portugal, Greece, and Spain. In these systems waiting lists work as management instruments for health resources. They have always been the object of special attention, particularly if certain waiting times are exceeded or some kind of collapse is produced. It is for that reason that waiting lists are always on the agenda of health reforms. Some measures have been used to try to shorten the waiting time and the number of patients in queues. For example: maximum waiting time guarantee, trying to diminish the length of stay of hospitalized patients, increasing productivity policies, increasing resources, improving the management of the waiting lists, increasing patient choice or reducing public demand by subsidizing voluntary private health insurance.

Continuous and appropriate management of the waiting lists is essential. The operations research techniques used to control the waiting lists can be enumerated: queuing theory (Worthington 1987), simulation (H&SSSG 1994; Wisniewski 1997), data envelopment analysis (DEA) (O'Neill and Dexter 2004), mathematical programming (Cooper 1981), and multicriteria decision making (Arenas et al. 2002).

One important factor in demand for private health insurance seems to be the long-term waiting lists. At the same time, the choice between public and private health services should be an aspect with many consequences for waiting lists. One consequence of high private coverage could be that the public health system should improve because it would have fewer patients. However, a long-term system could go in the opposite direction, if there are feedback effects from private insurance demand to waiting lists. In this empirical sense it has been observed that some areas with high private coverage had higher waiting lists because the government may under-fund public services in areas with private insurance coverage.

The choice of public or private sector could also have distributional implications. Individuals who opt out of public sector treatment free up resources for those who continue to rely exclusively on the National Health System.

Another important scheme is to establish priorities for managing waiting lists. Traditionally, the usual prioritization rule for nonurgent patients was “first in, first out”, but now new factors are taken into account for the management of waiting lists in most of the developed countries. In this new context the necessity of the introduction of different standardized prioritization rules has been pointed out in order to preserve equity. Some of these rules are based on severity, urgency, need, expected benefit, etc. Another kind of prioritization has been considered in some non-European country. For example, New Zealand recognizes that the public resources are limited and the public treatment is only provided for patients with the greatest need.

Section 5 contains the application of a mathematical problem of dynamic optimization to the management of the surgery waiting lists in a public hospital of Madrid. This model is adapted to the particular circumstances of that hospital, being capable of adaptation to any other hospital. The mathematical program has elements of linear programming, integer programming, dynamic optimization in discrete time and multi-objective programming. The computer program HIPERLINDO has been used to solve the problem. An important property in our model is that it can easily be made adaptive, in the sense that in every month in the year, where we have new information about current waiting lists or updated forecasting for admission/exit of patients, it is possible to adapt the model in such a way that it incorporates the new information in substitution of the old and we can obtain updated values after optimization, from that month to the end of the year. The results obtained confirm our belief that this kind of mathematical technique is very useful in the management of surgery waiting lists.

In the work presented in Sect. 5, the agreements of the hospital, both with external authorities and between the different services of the hospital, were given. In future research it would be interesting to study the problem in two stages: first, a model as an aid to the decision maker of the hospital in the negotiations both internal and external; second, take the results of the negotiations as given (as is the case in the model presented here). In other situations it will be necessary to include the number of available beds as an additional constraint. Also it would be interesting to improve the forecasting of demand, entries of new patients to the waiting lists and patients that leave the waiting lists without an operation. It would also be interesting to introduce random elements in the model, especially when the operating room time of some of the surgical processes to be considered has high variance. Finally, it would be interesting to introduce in the model different criteria about priorities for waiting lists.

Appendix: Additional Data

Table 9.22 Total and public health expenditure (*Source*: O.C.D.E. 2005: Health data)

	Total expenditure on health per capita US$ PPP			Public expenditure on health per capita US $ PPP		
	1998	2000	2002	1998	2000	2002
No waiting times						
Austria	1,953	2,147	2,220	1,362	1,495	1,551
Belgium	2,041	2,288	2,515	1,433	1,613	1,790
France	2,231	2,416	2,736	1,696	1,832	2,080
Germany	2,470	2,640	2,817	1,942	2,080	2,212
Luxembourg	2,291	2,682	3,065	2,117	2,406	2,618
With waiting times						
Portugal	1,290	1,493	1,646	866	1,091	1,201
Denmark	2,141	2,351	2,580	1,755	1,940	2,142
Finland	1,607	1,698	1,943	1,225	1,276	1,470
Ireland	1,487	1,774	2,367	1,138	1,300	1,779
Italy	1,880	2,001	2,166	1,293	1,474	1,639
Norway						
Spain	1,371	1,493	1,646	990	1,056	1,176
Sweden	1,961	2,243	2,517	1,682	1,904	2,148
UK	1,607	1,839	2,160	1,292	1,392	1,801
Greece	1,517	1,617	1,814	743	810	980
The Netherlands	2,016	2,196	2,843			

Table 9.23 Total public health expenditure, % GDP; and public expenditure on in-patient care, % GDP (*Source*: O.C.D.E. 2005: Health data)

	Total expenditure on health % GDP			Public expenditure on in-patient health % GDP		
	1998	2000	2002	1998	2000	2002
No waiting times						
Austria	5.4	5.4	5.4	2.9	2.9	
Belgium	6	6.2	6.5	2.2		
France	7.1	7.1	7.4	3.8	3.6	3.7
Germany	8.3	8.3	8.6	3.3	3.2	3.3
Luxembourg	5.4		5	1.7	2.1	2.1
With waiting times						
Portugal	5.6	6.4	6.5			
Denmark	6.9	6.9	7.3	4.3	4.2	4.2
Finland	5.3	5	5.5	2.6	2.4	2.7
Ireland	4.7	4.7	5.5		3.3	
Italy	5.6	6	6.4	3.9	3.2	3.3
Spain	5.4	5.3	5.4	1.9	1.8	1.8
Sweden	7.2	7.2	7.9	3.5	3.9	2.8
UK	5.5	5.9	6.4			
Greek	4.9	5.2	5			
The Netherlands						

Table 9.24 Resources of the health systems (*Source*: O.C.D.E. 2005: Health data)

	Acure care beds/1,000 pop.		Hospital physicians		Total hospital employment	
	1998	2002	1998	2002	1998	2002
No waiting times						
Austria	6.40	6.1	3.3	3.3		
Belgium			3.7	3.9		
France	4.3		3.3	3.3	18.7	
Germany	9.2		3.2	3.3	15	15
Luxemburg	6	5.8	2.4	2.6		
With waiting times						
Portugal	3.2		3.1		10.1	
Denmark	3.6		3.1	3.3		18.1
Finland	2.6	2.3	3	3.1		
Ireland	3.1	3	2.2	2.4	14	16.7
Italy	5		4.1	4.4		
Norway						
Spain	2.9		2.8	2.9	10	
Sweden	2.6					
UK	4.1	3.9	1.9	2.1	22.4	23
Greece	4		4.3		9.5	
The Netherlands	3.7		2.9	3.1	16.6	16.6

Table 9.25 Total surgical cases/1000p and average length of stay (*Source*: O.C.D.E. 2005: Health data)

	1998	2002	1970	1990	1998	2002
No waiting times						
Austria			2.2	13	10.9	8.1
Belgium				13.8	11.4	
France			18.3	13.3	14.1	
Germany			23.7	17.2		
Luxembourg	216.6	214.1	27	17.6	15.3	
With waiting times						
Portugal	48	58	23.8	10.8	9.8	
Denmark	170	207.6	18.1	8.2	6.7	5.7
Finland	89	91.2	24.4	18.2	11.8	
Ireland	136.7	197.8	13.3	7.9	7.8	7.6
Italy	64.4	73.3	19.1	11.7	10.1	
Norway					10	
Spain	63.6			12.2		
Sweden			27.2	18	6.7	6.2
UK	124.9	125.3	25.7	15.6	10.2	8.1
Greece			15	9.9	8.2	
The Netherlands	71.8	75.1	38.2	34.1	32.8	

References

Aaron, H. J., & Schwartz, B. (1984). *The painful prescription: Rationing hospital care*. Washington, DC: Brookings Institution.

Arenas, M., Bilbao, A., Caballero, R., Gómez, T., Rodríguez, M. V., & Ruiz, F. (2002). Analysis via goal programming of the minimum achievable stay in surgical waiting lists. *Journal of the Operational Research Society, 53*(4), 387–396.

Besley, T., & Coate, S. (1991). Public provision of private goods and the redistribution of income. *American Economic Review, 81*(4), 979–984.

Besley, T., Hall, J., & Preston, I. (1998). Private and public health insurance in the UK. *European Economic Review, 32*(3–5), 491–497.

Besley, T., Hall, J., & Preston, I. (1999). The demand for private health insurance: Do waiting lists matter? *Journal of Public Economics, 72*, 155–181.

Bitran, G. R., & Valor-Sabatier, J. (1987). Some mathematical programming based measures of efficiency in health care institutions. *Advances in Mathematical Programming and Financial Planning, 1*, 61–84.

Chae, Y., Suver, J., & Chou, D. (1985). Goal programming as a capital investment tool for teaching hospitals. *HCM Review, Winter*, 27–35.

Cooper, R. C. (1981). A linear programming model for determining efficient combinations of 8, 10 and 12-hour shifts. *Respiratory Care, 26*, 1105–1108.

Clemmesen, F., & Hansen, M. (2003). Erfaringerne med meraktivitetsfinansiering af sygehuse. *Samfundsøkonomen, 3*, 11–16.

Council of Europe. (1998). *Recommendation No R(99) 21 on the criteria for the management of waiting lists and waiting times in health care*. Retrieved from http://www.coe.int.

Cullis, J. G., Jones, P. R. (2000). Waiting lists and medical treatment: Analysis and policies. In *Handbook of health economics* (Vol. 1, Part B, pp. 1201–1249).

Gravelle, H., Smith, P. C., & Xavier, A. (2003). Waiting lists and waiting times: A model of the market for elective surgery. *Oxford Economic Papers, 55*(1), 81–103.

H&SSSG. (1994). *Simulation in health care management*. Operational Research Society.

Hadorn, D. C., & The Steering Committee of the Western Canada Waiting List Project. (2000). Setting priorities for waiting lists: defining our terms. *Canadian Medical Association Journal, 163*(7), 857–860.

Hoel, M., & Saether, E. M. (2003). Public health care with waiting time: The role of supplementary private health care. *Journal of Health Economics, 22*(4), 599–616.

Hurst, J., & Siciliani, L. (2003). *Tackling excessive waiting times for elective surgery: A comparison of policies in twelve OECD countries*. OECD Health Working Papers n. 6.

Iversen, T. (1997). The effect of a private sector on the waiting time in a National Health Service. *Journal of Health Economics, 16*, 381–396.

Kee, F., Mcdonald, P., Kirwarn, J. R., Patterson, C., & Love, A. H. G. (1998). Urgency and priority for cardiac surgery: A clinical judgment analysis. *BMJ, 316*, 925–929.

Lindsay, C. M., & Feigenbaum, B. (1984). Rationing by waiting list. *American Economic Review, 74*(3), 404–417.

Martin, S., & Smith, P. (1999). Rationing by waiting lists: An empirical investigation. *Journal of Public Economics, 71*, 141–164.

Noseworthy, T. W., McGurran, J. J., & Hadorn, D. C. (2003). Waiting for scheduled services in Canada: Development of priority-setting scoring systems. *Journal of Evaluation in General Practice, 9*(1), 23–31.

O.C.D.E. (2005). *Health data*.

O'Neill, L., & Dexter, F. (2004). Evaluating the efficiency of hospitals perioperative services using DEA. In M. L. Brandeau, F. Sainfort, & W. P. Pierskalla (Eds.), *Operations research and health care. A handbook of methods and applications*. Boston, MA: Kluwer Academic Publishers.

Ortún-Rubio, V., Pinto-Prades, J. L., & Puig-Junoy, J. (2001). *El establecimiento de prioridades en al cirugía electiva*. Madrid: Ed. Ministerio de Sanidad y Consumo.

Rodríguez Sendín, J. J. (2000). Responsabilidad y listas de espera. El problema de las listas de espera en el SNS. Discordancias, ética y equidad. Actas del VII Congreso de Derecho Sanitario, Madrid.

Siciliani, L., & Hurst, J. (2003). *Explaining waiting times variations for elective surgery across OECD countries*. OECD Health Working Papers n. 7.

Wisniewski, M. (1997). *Quantitative methods for decision makers*. London: Pitman Publishing.

Worthington, D. J. (1987). Hospital waiting list management models. *Journal of Operational Research Society, 42*(10), 833–843.

Chapter 10
Triage in Nonemergency Services

Katherine Harding and Nicholas Taylor

Abstract Triage systems are traditionally associated with emergency services, but are also commonly used in a much broader range of health care settings. This chapter explores some of the arguments for introducing triage systems, as well as some of the pitfalls associated with their use. Lessons from patient flow literature suggest that there may be better ways to make decisions about patient priority and to maintain throughput, without defaulting to long waiting lists and associated triage systems. These principles are demonstrated using a case study of an alternative model of triage that successfully reduced waiting time in a community rehabilitation program.

Keywords Triage • Prioritization

1 Introduction

Triage and prioritization systems have become widely used as methods to assist with the management of waiting lists, and allocation of services to patients by treating clinicians. Triage is traditionally associated with emergency medicine, but in recent years these systems have become common across a broad spectrum of health services (Harding et al. 2011). Despite the extent of their use, there is limited evidence for the effectiveness of triage, and issues surrounding its use have been identified in the nonemergency setting. Many triage systems have been shown to lack reliability, and also have other potential pitfalls. Furthermore, there is a growing body of literature suggesting that there may be alternative ways of prioritizing patients that improve patient flow without adversely affecting patient

K. Harding (✉) • N. Taylor
Allied Health Clinical Research Office, Eastern Health, Melbourne, VIC, Australia

La Trobe University, Melbourne, VIC, Australia
e-mail: katherine.harding@easternhealth.org.au

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_10, © Springer Science+Business Media New York 2013

care. This chapter aims to present evidence regarding the use of triage systems in nonemergency services, explore issues surrounding the use of triage systems for these services, and presents a case study of one alternative to the traditional wait-list and triage model that was successfully applied to a community rehabilitation program.

Waiting lists are often considered to be an inevitable part of modern health care systems. Lengthy waits for health care are common in publicly funded services, and affect a broad spectrum of services across the health sector. The physical and psychological consequences of waiting for care have been described in a range of patient groups, including patients seeking elective surgery (Gimbel and Dardzhikova 2011; Hodge et al. 2007; Oudhoff et al. 2007), emergency departments (Molyneux et al. 2006), pain management services (Lynch et al. 2008), child development and rehabilitation services (Feldman et al. 2008; Miller et al. 2008; Russell et al. 2011), and veterans' health care programs (Pizer and Prentice 2011).

When demand exceeds supply, there is a need to make decisions about what types of services will be provided to which patients, and in what time frame. These decisions may take the form of prioritizing certain treatments or patient groups, or rationing services so that some services become unavailable to some types of patients. Both rationing and prioritizing of services take place at all levels of health delivery: politicians and policy makers set directions for health care funding (the macro level); managers of health services and authorities make specific decisions about where dollars will be spent (the meso level); and frontline clinicians make decisions about which patients are to be seen and in what order (the micro level). Prioritizing one patient or group of patients over another inevitably creates winners and losers, with ethical, political and economic consequences (Williams et al. 2012). There is therefore a need to ensure that decision making about the provision of services is transparent, and to ensure that limited services reach those who need them most in a timely manner.

1.1 What is Triage?

Triage was originally developed in the field of emergency medicine as a systematic method of assigning priority for medical treatment (Iserson and Moskop 2007). Triage systems are still commonly used in emergency departments, but have also become widely used by a variety of other settings to categorize patients in terms of urgency and/or the type of treatment required. The word triage comes from the French verb "trier" meaning "to sort," and is defined by the Oxford Dictionary as "the process of determining the most important people or things from amongst a large number that require attention." "Prioritization" (as a derivative of "prioritize") is defined in a more general sense: "to determine the order for dealing with a series of items or tasks according to their relative importance".

However, as the term triage has become more widely applied to the allocation of services beyond emergency care, "prioritization" and "triage" are often used interchangeably in the context of allocating services to patients at the point of service delivery. Irion (1997) attempts to make the distinction in relation to physical therapy services by describing prioritization as a process to rank patients in terms of need for services, and triage as the process for reaching a decision about the type of services provided (for example, the need for a physical therapist versus a therapy assistant) or the need for services at all (Irion 1997). Parkin, Frake and Davison (2003) describe triage in relation to mental health services as a process of "determining clinical need, the likely response to intervention and the degree of urgency required in providing that intervention." This process of categorizing patients on the basis of need or urgency can vary from an intuitive, ad hoc process conducted by clinicians as part of management of a caseload, to a formal process based on predetermined protocols or assessment tools completed by a designated triage provider (protocol-driven triage). The definition of triage used by Iserson and Moskop (2007) takes this into account, suggesting the use of the term "triage" should be specific to situations in which there is (1) at least a modest scarcity of health care resources, (2) a specific person who assesses each patient's needs based on a brief evaluation, and (3) an established system or criteria to distinguish treatment priority for each patient.

Triage, then, is a defined process designed to sort patients into groups based on variable criteria including urgency, type of treatment or suitability for a service; prioritization is limited to ranking patients due to receive a specified service in order of priority.

1.2 Triage in Nonemergency Services

The concept of triage began as a way of sorting wounded soldiers on the battle field (Iserson and Moskop 2007), but over time its application has become far more widespread to the point where triage is now found in many places that are far removed from its origins. Many people still associate triage systems with emergency departments, and there is no doubt that triage systems continue to be widely used in these settings. However, protocol-driven triage systems have also become popular in many services that are not dealing with patients at imminent risk of death or serious decline. These nonemergency services include outpatient clinics (Christie et al. 1997; Rastall and Fashanu 2001), community health services (Brown and Pirotta 2011), rehabilitation programs (Passalent et al. 2010), and mental health services (Inglis and Baggaley 2005; Jones et al. 2000) to name a few.

The use of triage systems is widely reported in services provided by allied health professionals, including those from single disciplines such as physical therapists, occupational therapists, or psychologists as well as multidisciplinary allied health services (Gauthier et al. 2006; Harding et al. 2010a, b; Hardy et al. 2011; Rastall and Fashanu 2001). Allied health services may sometimes be provided as a single session incorporating an assessment, advice and treatment, but are commonly provided over a series of consultations or therapy sessions. In this sense they are different from other services typically associated with triage systems such as emergency departments and elective surgery services where services are provided as a single event. Nevertheless, waiting lists managed by protocol driven triage systems still appear to be widely used in allied health services, both in ambulatory services (Brown and Pirotta 2011; Wright and Ritson 2001) and inpatient settings (Gauthier et al. 2006; Lowe and Barber 2005; Porter and Jamieson 2012). An example of one such triage system, designed for use by an acute hospital occupational therapy service is shown in Box 10.1.

Despite the wide use of triage systems in nonemergency services, their benefits to patients and health care providers have not been well established. In emergency settings, triage systems have one obvious aim of ensuring that very urgent patients (such as those with life threatening conditions) receive rapid care. However, nonemergency services such as those provided by many allied health professionals often have a stronger focus on restoration of function, and are less likely to be dealing with critical or life threatening situations. When clinicians are not making decisions about patients at imminent risk of death, triage systems are usually set up to help clinicians to make decisions about which patients should be seen first when demand is great and resources are limited. Priority decisions in these settings may need to take into account a range of other factors such as pain, loss of function, risk of deterioration, impact of a disability on work or social roles, dependence on others, or economic impacts. Many of these factors may be difficult to evaluate, and subject to varying opinions about their relative importance. Designing triage criteria under these circumstances becomes a complex task.

Box 10.1: The Ottawa hospital occupational therapy prioritization guidelines for physical medicine inpatients: An example of a triage system for a nonemergency service

1. *Acute priority* (Service within 24 h): Immediate action is required to prevent deterioration or exacerbation of a medical condition.
2. *High priority* (service within 48 h): Medically stable. Anticipating discharge home alone or home with limited support.
3. *Moderate priority* (service within 72 h): Anticipating discharge home with caregiver or supportive environment (e.g., rehabilitation program, convalescence home), or anticipating changes in discharge destination.
4. *Low priority*: Medically unable to participate in occupational therapy; needs can be met in community; or conditions are longstanding and will not change in the acute care setting.

Gauthier et al. (2006)

2 Benefits and Pitfalls of Triage in Nonemergency Services

2.1 Rationale for the Use of Triage Systems in Nonemergency Services

Triage systems may be introduced for a variety of purposes, with the primary aim being to benefit the patient, the clinician or the health service. Some of the more common reasons for introducing triage systems include:

2.1.1 Ensuring that the Patients with Greatest Need Receive Rapid Service

Services that do not deal routinely with emergency or high acuity patients still have patients with varying degrees of need, limited resources and high demand. Triage systems may therefore be introduced in an attempt to ensure that the patients with the most urgent needs are sorted from those presenting with more routine problems and receive rapid access to the service. In this sense triage systems for nonemergency are not dissimilar to those for emergency services, except that they may operate over very different timeframes and with different types of patients. Triage systems for emergency services may seek to ensure urgent patients are seen in minutes rather than hours, where as some nonemergency services may attempt to ensure access within weeks rather than months for urgent cases (Jones et al. 2000; Woodhouse 2006). Under either scenario, however, triage systems are implemented with the expectation that meeting these timeframes will improve patient outcomes.

2.1.2 Transparency in Decision Making

Triage systems are used to increase transparency in the allocation of resources, by providing a systematic method for prioritizing one patient over another according to objective rules or guidelines. A clear policy that dictates who will receive priority can support clinicians to make difficult decisions that need to be explained to patients, families, colleagues, managers and possibly the legal system. While setting up a triage system for this purpose may seem to be a worthwhile activity, it can only be considered successful if the triage system has demonstrated reliability and validity. If providers disagree on triage ratings or do not apply the criteria in the same way every time, the system cannot be considered reliable. Similarly, if the triage system does not identify patients who are at the most risk or have the most urgent needs, it will add little value to patient care (Harding et al. 2009). A system that lacks reliability and validity therefore fails to provide the transparency for which it was designed.

2.1.3 Translation of Organizational Priorities to Frontline Workers

Triage systems are one way of translating priorities set at the “meso level” (mangers and policy makers within health services) to the “micro level” where services are delivered. Priority categories in nonemergency services are not always based solely on the needs of individual patients, but may also be influenced by funding arrangements, consumer pressures, special interest groups or pressure on specific aspects of the health service. For example, organizational priorities may include a high throughput for orthopedic elective surgery to keep operating rooms at full capacity and minimize surgical waiting lists. Physical therapy services to these patients may be given higher priority over others with more urgent need for therapy in order to make sure that these patients are discharged as quickly as possible to maintain patient flow. A protocol based triage system can help to communicate these priorities to clinicians making day to day decisions about how to allocate their services.

2.1.4 Describing Patient Populations for Allocation of Clinical Resources

Data collected from protocol driven triage systems is sometimes used as a method of describing the acuity of patient populations, or dividing patients into groups for measuring performance indicators. For example, studies undertaken in emergency departments frequently describe patients and analyze data according to triage categories (Tanabe et al. 2004; Sethuraman et al. 2011). Similarly, such data is sometimes used in nonemergency services for organizational decision making, such as the basis for the allocation of resources (Porter and Jamieson 2012). A dietetics service in an acute hospital, for example, may choose to provide a higher level of resources to Ward A that has a high number of patients deemed to be priority category 1 (recommended for assessment with 24 h) compared with Ward B with more category 2 patients (with a target assessment time of 3 days).

However, reliability and validity issues described above in relation to transparency of decision making can also be significant problem in the use of triage data as the basis for operational decision making. If the apparent difference in caseload between ward A and B in this example is actually explained by differences in the way the dietitians on those two wards interpret the triage protocols rather than actual differences in the patient population, using this data as the basis for allocation of resources is fundamentally flawed.

2.2 Potential Pitfalls in the Use of Triage Systems

Although triage systems can come in many varieties, one of the more traditional and widely used approaches in nonemergency services is to assign a triage category to each incoming patient according to a set of protocols or triage criteria, and then place that patient on a waiting list to be contacted when the service becomes available. We will call this the “triaged waiting list” approach, and describe here some of the potential pitfalls in the use of this model.

2.2.1 Diversion of Resources from Frontline Care

Reducing waiting times has been recognized as a priority across a range of health settings. Waiting for care not only has direct consequences for patients; it also creates inefficiencies in the delivery of services. Once a waiting list exists, a new layer of activity is needed to monitor and prioritize the waiting patients, directing resources away from direct patient care (Kreindler 2008). The formation and implementation of triage systems is a typical of this type of activity, which includes the initial creation of criteria and protocols (in itself a time-consuming exercise), assessing new referrals and assigning triage categories, and monitoring the changing needs of people on the list. These processes may be conducted by someone within the team with major responsibility for this task (such as a team leader or administrative assistant), by treating clinicians on an ad hoc basis, or within a whole separate structure within the health service, such as a centralized referral office or access unit.

The dedicating of resources to the management of waiting has been described not only across the spectrum of health services, but is also recognized in other industries. Operations Management literature describes strategies such as reducing “works in progress,” and the “just in time” approach (Vissers and Beech 2005). Both are about reducing the amount of work in the system at any time, and doing the work when it needs to be done. These ideas acknowledge that there are costs involved with every piece of work that is in the system; components need to be stored, creating a need for additional space and double handling in and out of storage spaces, resources held in parts awaiting assembly reduce cash flow, administration systems need to track large numbers of items, orders have the potential to be lost and so on. Similarly, the “lean thinking” approach has also developed from experiences in the manufacturing industry, and focuses on removing processes that do not add value to the final product (Bowen and Youngdahl 1998).

Within health systems, people on waiting lists could, in a sense, be considered to be the health care equivalent of “works in progress” with the systems associated with managing waiting not adding value to the final service the patient is waiting to receive. Children waiting many months for speech therapy services, for example, receive little value from the work of administrative and clinical staff who are developing prioritization criteria, organizing and monitoring the waiting list and spending time contacting clients to check that services are still required. These resources are diverted from “value adding” activities such as direct clinical care or other activities that directly improve the quality of service delivery.

2.2.2 The Creation of an Expectation of Waiting

Systems that are built around the management of waiting lists inevitably focus attention on working around (and therefore accommodating) long waiting times, rather than the bigger picture of why patients are waiting in the first place.

There may be times when there is no alternative to asking some patients to wait, and when a triage system may be of benefit in assisting service providers to ensure that the patients in greatest need get the fastest service. However, it is possible that triage and other “waiting” systems have become embedded in some services that could operate without them, leading to excessive complexity in booking processes that inadvertently contribute to longer waiting times (Kreindler 2008).

Many health providers will be familiar with services that have always had a waiting list, but the size of the waiting list or typical time spent waiting varies relatively little outside of monthly or seasonal fluctuations. Some of these services may have a 6 month wait, others a 3-week wait, but whatever the typical waiting period the variation over time is minimal. A service that has always had a 6-week wait clearly has a balance between supply and demand. If more patients were arriving than could be seen, the waiting list would continually grow longer. Why then, do patients in this hypothetical, but not unusual, service have to wait 6 weeks for an appointment?

The answer to this question is not altogether clear, but may be partly due to habit, expectations and embedded behaviors that become entrenched in health systems and are resistant to change. Once there is an expectation that patients will wait, attention may become constantly focused on methods to manage waiting (such as the development of triage categories) rather than questioning the need for waiting in the first place.

2.2.3 Triage Systems Frequently Lack Reliability

Triage systems may provide some comfort to the clinicians that patients are not being ignored, and some reassurance that access systems are fair, objective and transparent. Unfortunately, there is considerable evidence that any such reassurance may be misplaced.

Inter-rater reliability has proven to be a challenge in triage systems, with frequent findings of low to moderate levels of agreement. For example, a study examining the inter-rater agreement of prioritization decisions in a community rehabilitation service resulted in a weighted kappa of 0.6 and showed that raters agreed on approximately 70 % of referrals (Harding et al. 2010a). Although this is better than chance alone, it still means that three of every ten referrals will receive different priority rating depending on who picks up the referral. It could be argued that lack of agreement could simply be due to a failure of the raters to properly apply the criteria, but a follow-up study in the same service found no improvement following a program of rater training (Harding et al. 2010b). Similar issues with lack of reliability in triage systems have also been identified in a range of other health service settings (Creaton et al. 2008; Cunningham et al. 2000; Dennett and Parry 1998; Gravel et al. 2007; Leonard 1993; O'Cathain et al. 2003; Wright and Ritson 2001).

The difficulty in attaining high levels of reliability in triage systems is not surprising when one considers the complexity of triage decisions. These decisions are essentially about who will receive priority over others, and often involves the weighing up of many different factors. With the possible exception of conditions that can be triaged according to very specific, quantifiable criteria, even the most well thought out triage protocols are open to a degree of interpretation. Those assigning triage categories do so against a background of their own experience and values, and in an environment that presents external factors (such as vocal relatives threatening formal complaints, or highly respected medical specialists advocating for their patients) that may influence triage decisions.

The difficulties in establishing high levels of reliability in triage systems not only threaten to undermine the fairness and transparency they are designed to uphold, but also have implications for the use of triage data for operational decision making. Using triage scores that lack reliability as an indicator of caseload complexity, for example, is a flawed basis on which to allocate clinical resources. Furthermore, reliability is an essential prerequisite before validity can be established (Streiner and Norman 2003), so lack of reliability has important implications for the value of triage systems.

2.2.4 Difficulties with Establishing Validity

The question of validity in assessment tools is essentially about whether or not a tool measures what it intends to measure. There are many types of validity, a full discussion of which is beyond the scope of this chapter. However, if the aim of a triage system is to identify patients with the most urgent needs, to establish whether the system works it is necessarily to know who the patients with the most urgent needs are. As previously discussed, priority decisions (particularly in nonemergency services) are often subjective and based on values rather than measurable indicators, making it very difficult to come up with a gold standard of urgency against which to compare triage decisions. Evaluations of triage systems are therefore often based on indirect measures of performance such as agreement between clinicians or with an expert panel, service outcomes like the percentages of people seen within target timeframes, or secondary outcomes that may be considered to be markers of successful decision making (such as adverse events due to inappropriate service times).

Another difficulty in establishing the validity of triage systems is a tendency for the users to avoid using the categories at the extremes of the range, and to allocate the vast majority of cases to the middle categories (Harding et al. 2012). There is some evidence to suggest that triage systems have value in sorting the patients with very urgent needs from the rest, but the value of dividing into additional categories is more questionable. For example, a triage system in a community rehabilitation service showed that waiting times were significantly less (mean 4.8 days) for patients allocated to the most urgent category, but the choice of triage categories (2, 3 or 4) for the vast majority of patients who were referred to the less urgent categories made little difference to waiting time (19.6 days, 26.4 days and 19.4 days respectively) (Harding et al. 2012). Simplifications of triage have therefore been advocated, avoiding complex systems with multiple categories and instead allocating patients into two groups. For example; urgent versus routine cases (Kreindler 2008), or separating patients who are likely to require hospital admission from those who are not (King et al. 2006).

2.2.5 Traditional Triage Systems Can Limit the Scope of Decision Making

The triaged waiting list approach usually involves assigning triage categories to incoming referrals, thereby making a judgment about the urgency in comparison to other patients who are arriving at the same time or already on the waiting list. However, this type of triage system often does not compare the needs of new patients to those who are already receiving a service. This issue becomes an important consideration in nonemergency services in which patients may be receiving care over an extended period. For example, a triage process for an allied health professional working in the field of childhood disability may describe in detail the relative priority of incoming referrals, but may not consider the question of whether a new referral for Child A is a higher priority than continuation of weekly therapy sessions for Child B who is already receiving treatment.

3 The Effect of Triage on Patient Flow

Health service managers are keenly interested in identifying factors that improve or impede the movement of patients through health systems (Walters and Dawson 2009) and therefore have an interest in the effect of triage systems on patient flow. Triage systems are frequently used to assist clinicians to make decisions about allocation of resources, and efficient allocation of resources has the potential to improve patient flow by reducing waiting time and length of stay for patients accessing health services. However, at the same time triage systems add additional processes in the access to care that may have a contrary effect.

3.1 How Does Triage Affect Patient Flow?

A systematic review by Harding et al. (2011) considered the question of whether triage systems across a broad range of health services improve patient flow. The review included 25 studies that reported comparative outcome data on triage systems with outcome measures related to patient flow. Studies were not limited by health setting, and the included papers were from emergency departments, mental health services, dental surgery clinics, outpatient sexual health clinics and an obstetrics unit.

The findings of the review were inconclusive in regard to the impact of patient flow on simple triage systems that allocated patients to triage categories compared with no formal triage system. Some studies reported improvements in patient flow with a triage system in place, whereas others found that triage added to overall waiting time. Several studies reported improvements to patient flow with adjustments to triage criteria, suggesting that there are multiple factors that may influence how a triage system affects patient flow.

This review also reported on a group of studies that compared a traditional triage system that only allocated patients to categories, to an enhanced system in which management options were available at the point of triage. These options may have included the ability to provide initial advice, commence treatment or discharge simple cases. Such systems have been tested most extensively in hospital emergency departments, by placing a doctor at the triage desk in conjunction with the traditional nurse triage role. However, similar concepts have also been reported in community mental health services (Lynch and Hedderman 2006) and outpatient clinics (Tideman et al. 2003). Although many of the studies used observational designs and were considered to be of low to moderate methodological quality, they consistently reported improvements in patient flow when compared with traditional approaches to triage. For example, Lynch and Hedderman (2006) reported a fall in average waiting time from 122 days to 38 days following the introduction of face to face triage assessments that were designed to "make the first encounter with service beneficial in its own right." In this regard, the findings of this review concurred with previous literature reviews of emergency department flow (Bond et al. 2006; Cooke et al. 2004) but also suggested that this principle is also likely to be applicable to nonemergency services.

Treating simple cases or redirecting those who do not require services at the point of triage addresses both the aims of treating some patients in a shorter time frame, and rapidly removing some patients with relative minor needs from the list. Triage providers must collect sufficient information about a patient's needs in order to make a decision about priority. Sometimes a simple and rapid intervention is identified that would meet these needs and the triage provider has the skills to implement it. Providing this service immediately rather than duplicating the process later is likely to benefit both service efficiency and patient satisfaction. Combining triage and initial management generally requires triage to be brought to the point of service delivery, rather than conducted as a separate process isolated from frontline clinicians.

3.2 Triage in the Context of Supply and Demand

Given that waiting lists are often assumed to be the result of an imbalance between supply and demand, most interventions to reduce waiting times address one of these two factors (Rotstein and Alter 2006). Interventions to increase supply can come in various forms. One obvious supply side intervention is an injection of additional resources to reduce the backlog, but this approach has often been found to be ineffective in achieving long term reductions in waiting time (Kenis 2006). Other interventions that act on supply aim to encourage an ongoing increase in activity (for example, through fee for service models of funding), looking for additional capacity (for example, supplementing with services from the private sector) or strategies to increase the efficiency of use of existing resources (Kreindler 2010).

Triage or prioritization processes can be considered to be demand side strategies, as they moderate demand by influencing who is entering the service and the priority that will be given to each new arrival. Demand side interventions can also involve rationing of services, such as tightening eligibility criteria, or reducing the service to each patient by decreasing the duration or frequency of treatment (Williams et al. 2012). Other ways of reducing demand include interventions that aim to reduce the use of unnecessary or ineffective activity, such as reducing inappropriate referrals or procedures (Hobbs et al. 2011; Isouard 1999) or limiting the use of specialist services to those who really need them (Maddison et al. 2004).

Many approaches to waiting list interventions therefore tend to treat supply and demand separately, with interventions aimed at addressing one or the other. However, supply and demand also interact with each other, a factor that can be exploited to maximize efficiency of patient flow. The principle can be illustrated with a simple analogy from the retail industry. If four customers arrive at a store simultaneously, the shopkeeper will make a different choice about how to serve them than would be the case had they arrived one by one over half an hour. One or two can perhaps be pointed to the relevant part of the store and given a few minutes to browse, another with a question about the price of goods may be answered in a moment and on her way, while the fourth is provided with more comprehensive service. By the time this customer is attended to, the first two may be better placed to know what they need, making more efficient use of the shopkeeper's time. It would make little sense, on the other hand, to keep customers waiting in a queue out of sight of the shopkeeper, and have an independent gate keeper allow them through one at a time as each selected a product and completed a transaction. Priority decisions (influencing demand) about who to serve next are therefore made with some knowledge of the current level of available supply (that is, the needs of existing customers), and decisions about the supply of services to each individual customer also take into account the level of demand at that particular time.

Triage systems that combine triage with initial treatment can also take advantage of this principle, particularly if the provider making the triage decision also has some influence over the supply of the service. The following section illustrates this concept through the description of a triage intervention that was successful in reducing waiting time in a community rehabilitation service.

4 An Alternative Approach to Triage in Nonemergency Services: A Case Study from Community Rehabilitation

The contribution of alternative approaches from health operations management literature, such as the lean thinking approach (Bowen and Youngdahl 1998; King et al. 2006), have led some health services to question traditional models of triage and try more innovative approaches to prioritize patients. These include the various enhancements to traditional triage models that have been tried in emergency departments in recent years, including commencing investigations or treatments at the point of triage or introducing fast track systems to manage simple cases quickly (Oredsson et al. 2011). Other systems have moved away from separate triage systems, to a first come, first served, model aiming to see all patients within a short time. These include the Advanced Access approach designed for general practice clinics (Murray and Berwick 2003) or more comprehensive "up front" assessment and triage clinics to enable early decision making and care planning (Parkin et al. 2003). The following case example illustrates another model aimed at improving the triage process while also enhancing patient flow that was successfully applied to a community rehabilitation program (CRP). Some of the key features of this approach together with other evidence-based models of access and triage that have been shown to improve patient flow across a variety of health care settings are summarized in Table 10.1.

4.1 Study Setting

This study took place in a publicly funded adult musculoskeletal CRP operating across two sites in a large metropolitan health service, offering multidisciplinary outpatient rehabilitation to patients following elective joint replacements, fractures or soft tissue injuries, as well as less specific conditions such as debility or deconditioning. Each of the two CRP teams included physical therapists, an occupational therapist, a social worker, dietitian, and allied health assistant with patients seen by any number of disciplines according to need. The service was using a traditional "triaged wait list" in which patients were allocated to one of four categories according to urgency from 1 (highest) to 4 (lowest) at the point of referral. The patient was placed on a waiting list, and therapists accepted new patients as they discharged others off their caseload. Previous studies in this setting suggested the existing waitlist and triage system lacked reliability (Harding et al. 2010a, b) and made minimal difference to waiting time (Harding

Table 10.1 Examples of models of access and triage that have been demonstrated to improve patient flow. All focus on prompt face-to-face assessment/triage with immediate initiation of treatment

Strategy	Setting of design/ evaluation	Demonstrated outcome	Key references
STAT (*Specific and Timely Appointments for Triage*) aims to provide a rapid assessment appointment for all patients. Clinicians are given autonomy to make triage decisions within their caseload	Community Rehabilitation	Reduced time from referral to first appointment	Harding et al. (2013a, b)
Advanced access involves reducing prebooked appointments, instead opening the schedule for same day appointments	General practice surgeries	Reduction in time to first available appointment	Murray and Berwick (2003)
Triage clinics, in which patients have an initial, brief assessment with a multidisciplinary team with subsequent triage to further assessment and treatment, specific treatment streams, or brief intervention and discharge	Community mental health services	Reduction in time from referral to first appointment	Jones et al. (2000), Parkin et al. (2003)
Combining the triage role with initial assessment and management by placing medical staff at triage	Hospital emergency departments	Reduced waiting time and reduction in the number of patients who leave without being seen	Oredsson et al. (2011), Hodge et al. (2007)

et al. 2012). The mean time from referral to first appointment was consistently around 3 weeks over the previous 2 years, suggesting supply and demand were not out of balance, but the delay had become an accepted feature of the service.

4.2 Specific and Timely Appointments for Triage

A new model, referred to as Specific Timely Appointments for Triage (STAT) was developed based on evidence of successful features of triage systems (Harding et al. 2011). The key feature of STAT was that all clinicians created a specified number of assessment times in their weekly schedule with the aim of allocating an appointment immediately on referral. The number of appointment slots required was calculated by dividing the average number of referrals received per week per discipline by the number of equivalent full time clinicians of that discipline, and adjusting to account for anticipated loss of assessment slots due vacation time and unplanned leave. On receipt of a referral the team leader immediately allocated the patient an initial appointment with at least one member of the team.

At the first visit, the clinician was given the autonomy to make a decision about the patient's priority and ongoing needs within the context of their caseload. For example, they could begin treatment immediately, or could provide a home program and review in several weeks. Following a period of baseline data collection, STAT was introduced at one of two of the CRP sites utilizing well established principles of change management (Fernandez and Rainey 2006). The impact of the model on waiting time, as well as secondary outcomes including length of stay in the service, quality of life scores at discharge and adverse events was tested in a prospective controlled before and after study. The comparison group was a control site in the same health service that continued to use the traditional triaged wait list approach. The triaged waitlist and STAT models are summarized in Fig. 10.1.

4.3 Results of the Trial

The STAT model was tested in a controlled before and after trial involving 971 patients, in which baseline data was collected from two CRP sites using a traditional "wait list and triage" model, and the intervention was introduced at one site maintaining the other as a control (Harding et al. 2013b). Using specific and timely appointments rather than a waiting list for managing referrals for community rehabilitation resulted in a 43 % decrease in overall time to first appointment (17.5–10 days) with no change at the control site. Patients ready to begin rehabilitation received their first appointment in a mean of 7.7 days, and were 4.5 times more likely to receive an appointment within 7 days under this system compared with the control site. The STAT model had no impact on total length of stay in the program, adverse events or quality of life scores at discharge. Semi-structured interviews with 32 staff and patients suggested that it was well received by both groups, with the majority of staff stating that they preferred STAT to the previous model of care (Harding et al. 2013a). Some expressed that they had initial doubts, but any loss of autonomy related to having patients automatically allocated to their care were outweighed by the benefits of not having to organize appointments and having more structure to their week. The intervention was implemented with considerable attention to good change management principles, and it was apparent from the qualitative data that this was also an important factor in the successful implementation of the change.

In this study the community rehabilitation clinicians were given prompt and direct knowledge of all patients being referred, as well as the autonomy to make their own decisions about the management of their caseload. The STAT approach to triage encourages staff to adjust their practices depending on the number and type

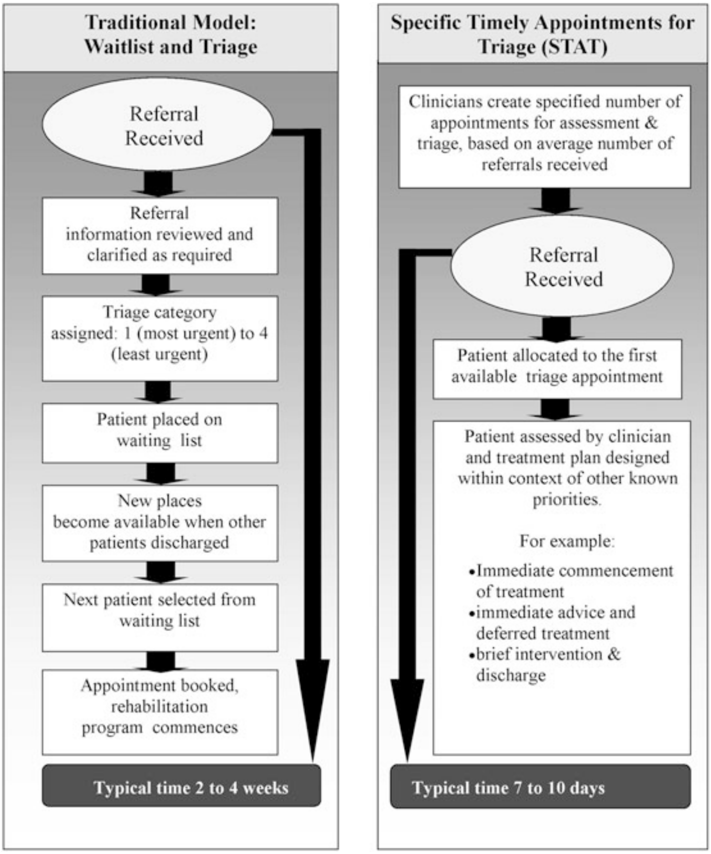

Traditional Model: Waitlist and Triage
Referral Received
Referral information reviewed and clarified as required
Triage category assigned: 1 (most urgent) to 4 (least urgent)
Patient placed on waiting list
New places become available when other patients discharged
Next patient selected from waiting list
Appointment booked, rehabilitation program commences
Typical time 2 to 4 weeks
Specific Timely Appointments for Triage (STAT)
Clinicians create specified number of appointments for assessment & triage, based on average number of referrals received
Referral Received
Patient allocated to the first available triage appointment
Patient assessed by clinician and treatment plan designed within context of other known priorities.
For example:
•Immediate commencement of treatment
•immediate advice and deferred treatment
•brief intervention & discharge
Typical time 7 to 10 days

Fig. 10.1 Comparison of traditional "waitlist and triage" and STAT models, with sample time frames observed in a community rehabilitation service

of new patients being referred into the service. Outcome data indicated that there was some increase in the use of group versus individual therapy, suggesting that this was one strategy used to increase flow during busy periods. Other strategies used included programs for patients to work on at home for a week or two until regular therapy sessions could commence or a reduction in frequency of sessions. During a lull in referrals, however, therapists could offer an extra session in areas of need.

4.4 Potential Reasons for Success

Unlike traditional waitlist and triage models which separate the management of supply and demand, such that those supplying the service have minimal knowledge of those who are waiting (Fig. 10.2a), STAT recognizes that supply and demand are linked and that each can constantly be adjusted in response to the other (Fig. 10.2b).

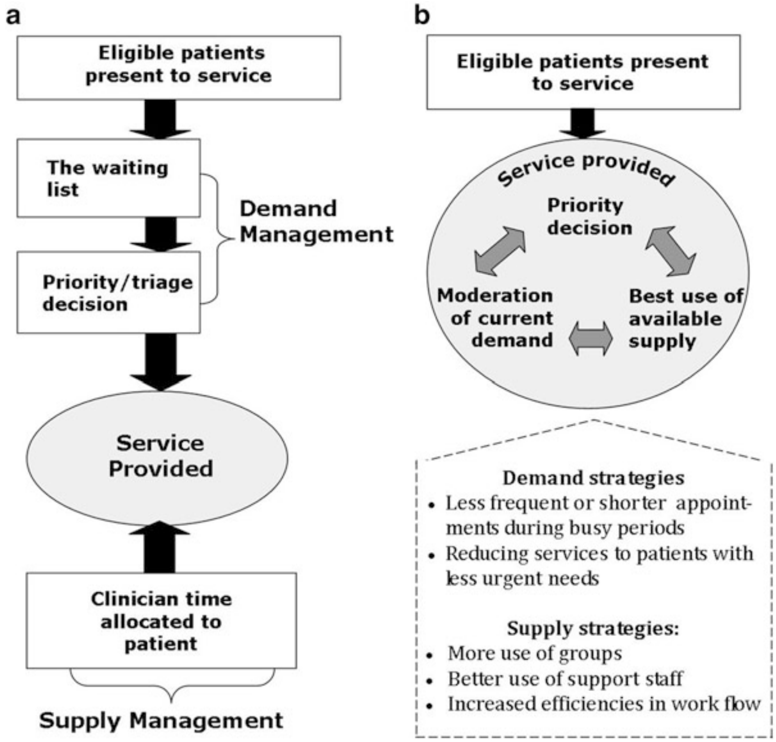

a
Eligible patients present to service
The waiting list
Priority/triage decision
Demand Management
Service Provided
Clinician time allocated to patient
Supply Management
b
Eligible patients present to service
Service provided
Priority decision
Moderation of current demand
Best use of available supply
Demand strategies
• Less frequent or shorter appointments during busy periods
• Reducing services to patients with less urgent needs
Supply strategies:
• More use of groups
• Better use of support staff
• Increased efficiencies in work flow

Fig. 10.2 (**a**) Supply and demand relationship in traditional "waitlist and triage" models. (**b**) Supply and demand relationship in the STAT model

It is important to note that community rehabilitation offers significant potential for flexibility in service supply, which could be a contributing factor to the success of this model. However, in a recent study the time from referral to scan was reduced from 12 to 5 days for diagnostic scans using a similar principle of allocating a specific number of scan slots to the treating team based on typical referral numbers, allowing the team to book their patients into those slots at their discretion rather than going through a central triage process (Elloy et al. 2009). The intervention was successful despite the fact that this was a diagnostic service conducted on a single occasion using a piece of equipment with fixed capacity, suggesting that flexibility in service provision may not be essential, but that the key element of the intervention is the direct link between the service provider and the triage decision.

The results of this controlled before and after trial demonstrate that the principle of linking supply and demand can enable new patients to be accommodated at the rate of arrival, thus preventing the formation of a waiting list. However, similar to other access models such as the Advanced Access approach for general practice clinics (Murray and Berwick 2003) it is important to stress that supply and demand must be reasonably well matched for this strategy to be sustainable. Services that have a waiting list that is continually growing do not have a match of supply and demand: for every patient who is treated, there is more than one patient placed on the end of the queue. Such a service will never "catch up," and patients will wait longer and longer until the waiting time becomes limited only by patients giving up, recovering or dying before they get treated. Conversely, waiting lists that are stable and not growing longer over time suggest a balance between supply and demand. Therefore, if the waiting list can be stabilized through strategies addressing supply, demand or a combination of both, there is then potential for a system such as STAT to have an impact.

Another important consideration is that no additional resources were used in the introduction of STAT in this setting, although the introduction of the intervention was timed to coincide with a regular seasonal lull in referrals following the Christmas holiday period. Qualitative data collected through interviewing staff suggested that there was significant anxiety during the start-up period as staff were faced with the task of absorbing those patients still on the waiting list. A single injection of resources may be useful in managing the existing backlog, at which point STAT may be effective in preventing the waiting list from forming again.

Unlike triage systems that are isolated from service delivery, STAT is based on the principle that the best person to make priority decisions is a clinician with expertise in the field, having a full understanding of both the patient's needs (through prompt, face-to-face assessment) and the context in which priority is being given (the existing caseload). STAT assumes that the treating clinician has this knowledge and expertise, and can be trusted to use sound clinical judgment to make appropriate priority decisions without the need for complex protocols.

It could be argued that since prioritization does not occur until the first face-to-face visit, very urgent patients who would otherwise have been seen within 2–3 days may be disadvantaged under STAT even if mean time to initial assessment is reasonable. However, results of the study provided no suggestion of any increase in adverse events or concerns to this effect raised by those who used the system. In addition, reliability of allocation of patients to category 1 was also found to be very low under the old system, suggesting that the original triage system did not necessarily identify these patients well.

This model of triage maintains the rationale for implementation of triage systems, while eliminating some of the pitfalls discussed earlier in this chapter. STAT still ensures that urgent patients are seen quickly, by providing clinicians with early and direct information about their needs so that they can make informed decisions about the relative importance of each new patient, weighed against the needs of all the other patients under their care. The model also still allows service managers working within higher level policy guidelines to retain influence over services that are provided, for example through funding choices and setting of eligibility criteria. However, STAT allows clinicians to have the autonomy to prioritize their own time and patient needs within this framework. As a result, two important principles identified in patient flow literature that have been associated with reductions in waiting times can be accomplished with STAT: triage is conducted at the point of care and can be combined with initial management; and the potentially wasteful processes of development and administration of triage systems and monitoring waiting lists can be eliminated from the system. Furthermore, as STAT does not put people into categories based on predetermined criteria, high levels of reliability and validity are not required. Finally, STAT turns attention to maintaining throughput and making decisions in response to demand, rather than defaulting to a waiting list that becomes an expectation of the service and potentially leads to long term complacency.

5 Conclusions

Triage systems originated in the field of emergency medicine, but are now widely applied in nonemergency settings in an attempt to ensure urgent patients are seen quickly, to aid clinical decision making, to translate organizational priorities to frontline care, or to measure case complexity. While they may be important and worthwhile systems in some settings, they can also be unreliable, subject to overuse and contribute to inefficiencies in others.

There is no doubt that clinicians and service providers will always need to make decisions about which patients should take priority over others. Formal, protocol driven triage systems are one way to achieve this, but other alternatives are emerging that have the potential to achieve the same end without adversely impacting on patient flow. The STAT model discussed in this chapter is one example of a model for undertaking the important task of evaluating the relative priority of individual patients, but takes a fundamentally different approach to how and when these decisions are made compared to traditional models that use waitlists with triage systems.

The application of triage systems to nonemergency services may appear at first glance to be a logical strategy to manage waiting lists when services are limited and demand is high. This chapter highlights some of the problems surrounding the use of triage systems in nonemergency services, and the risk of implementing protocol-driven triage systems that have the potential to inhibit patient flow while failing to achieve their purpose. Keeping triage processes simple, integrated with initial management, and conducted by clinicians at the forefront of service delivery may be a better approach for both prioritizing patients and maintaining patient flow in nonemergency services.

References

Bond, K., Ospina, M., Blitz, S., Friesen, C., Innes, G., Yoon, P., et al. (2006). *Interventions to reduce overcrowding in emergency departments*. Ottawa: Canadian Agency for Drugs and Technologies in Health.

Bowen, D. E., & Youngdahl, W. E. (1998). "Lean" service: In defense of a production line approach. *International Journal of Service Industry Management, 9*(3), 207–225.

Brown, A. M., & Pirotta, M. (2011). Determining priority of access to physiotherapy at Victorian community health services. *Australian Health Review, 35*(2), 178–184.

Christie, H. J., Gobert, A. D., Matthew, E., Rousseau, D. C., & Webber, S. C. (1997). Waiting list management strategies for outpatient orthopaedic physical therapy. *Physiotherapy Canada, 49*(3), 191–196. 205-196.

Cooke, M., Fisher, J., Dale, J., McLeod, E., Szczepura, A., Walley, P., et al. (2004). *Reducing attendances and waits in emergency departments: A systematic review of present innovations*. UK: Warwick.

Creaton, A., Liew, D., Knott, J., & Wright, M. (2008). Interrater reliability of the Australasian Triage Scale for mental health patients. *Emergency Medicine Australasia, 20*(6), 468–474.

Cunningham, C., Horgan, F., & O'Neill, D. (2000). Clinical assessment of rehabilitation potential of the older patient: A pilot study. *Clinical Rehabilitation, 14*(2), 205–207.

Dennett, E. R., & Parry, B. R. (1998). Generic surgical priority criteria scoring system: The clinical reality. *New Zealand Medical Journal, 111*(1065), 163–166.

Elloy, M., Jarvis, S., & Davis, A. (2009). A strategy to overcome the radiology lottery in the staging of head and neck cancer: An aid to attaining the 50-day rule. *Annals of the Royal College of Surgeons of England, 91*(1), 74–76.

Feldman, D. E., Swaine, B., Gosselin, J., Meshefedjian, G., & Grilli, L. (2008). Is waiting for rehabilitation services associated with changes in function and quality of life in children with physical disabilities? *Physical and Occupational Therapy in Pediatrics, 28*(4), 291–304.

Fernandez, S., & Rainey, H. (2006). Managing successful organizational change in the public sector. *Public Administration Review, 66*(2), 168–176.

Gauthier, R., Straathof, T., & Wright, S. (2006). The Ottawa hospital occupational therapy prioritization guidelines. *Occupational Therapy Now, 8*(6), 10–12.

Gimbel, H. V., & Dardzhikova, A. A. (2011). Consequences of waiting for cataract surgery. *Current Opinion in Ophthalmology, 22*(1), 28–30.

Gravel, J., Gouin, S., Bailey, B., Roy, M., Bergeron, S., & Amre, D. (2007). reliability of a computerized version of the Pediatric Canadian Triage and Acuity Scale. *Academic Emergency Medicine, 14*(10), 864–869.

Harding, K. E., Leggat, S., Bowers, B., Stafford, M., & Taylor, N. F. (2013a). Perspectives of clinicians and patients following introduction of a new model of triage that reduced waiting time: A qualitative analysis. *Australian Health Review, 37*(3), 324–330.

Harding, K. E., Leggat, S., Bowers, B., Stafford, M., & Taylor, N. F. (2013b). Reducing waiting time for community rehabilitation services: A controlled before and after trial. *Archives of Physical Medicine and Rehabilitation, 94*(3), 23–31.

Harding, K. E., Taylor, N. F., & Leggat, S. (2011). Do triage systems in healthcare improve patient flow? A systematic review of the literature. *Australian Health Review, 35*(3), 371–383.

Harding, K. E., Taylor, N. F., Leggat, S., & Shaw-Stuart, L. (2009). Triaging patients for Allied Health services: A systematic review of the literature. *British Journal of Occupational Therapy, 72*(4), 153–162.

Harding, K. E., Taylor, N. F., Leggat, S. G., & Stafford, M. (2012). The effect of triage on waiting time for community rehabilitation services: A prospective cohort study. *Archives of Physical Medicine and Rehabilitation, 93*(3), 441–445.

Harding, K., Taylor, N., Leggat, S., & Wise, V. (2010a). Prioritising patient for community rehabilitation services: Do clinicians agree on triage decisions? *Clinical Rehabilitation, 24*, 928–934.

Harding, K. H., Taylor, N. F., Leggat, S. G., & Wise, V. (2010b). A training programme did not increase agreement between allied health clinicians prioritising patients for Community Rehabilitation. *Clinical Rehabilitation, 25*(7), 599–606.

Hardy, J. A., Weatherford, R. D., Locke, B. D., DePalma, N. H., & D'Iuso, N. T. (2011). Meeting the demand for college student concerns in college counseling centers: Evaluating a clinical triage system. *Journal of College Student Psychotherapy, 25*(3), 220–240.

Hobbs, J. A., Boysen, J. F., McGarry, K. A., Thompson, J. M., & Nordrum, J. T. (2011). Development of a unique triage system for acute care physical therapy and occupational therapy services: An administrative case report. *Physical Therapy, 90*(10), 1519–1529.

Hodge, W., Horsley, T., Albiani, D., Baryla, J., Belliveau, M., Buhrmann, R., et al. (2007). The consequences of waiting for cataract surgery: A systematic review. *Canadian Medical Association Journal, 176*(9), 1285–1290.

Inglis, G., & Baggaley, M. (2005). Triage in mental health—A new model for acute in-patient psychiatry. *Psychiatric Bulletin, 29*(7), 255–258.

Irion, G. (1997). Prioritization/triage in acute care physical therapy departments. *Acute Care Perspectives, 5*(3), 1–3.

Iserson, K., & Moskop, J. (2007). Triage in medicine, part I: Concept, history and types. *Annals of Emergency Medicine, 49*(3), 275–281.

Isouard, G. (1999). A quality management intervention to improve clinical laboratory use in myocardial infarction. *Medical Journal of Australia, 170*(1), 11–14.

Jones, E., Lucy, C., & Wadland, L. (2000). Triage: A waiting list initiative in a child mental health service. *Psychiatric Bulletin, 24*, 57–59.

Kenis, P. (2006). Waiting lists in Dutch healthcare: An analysis from an organization theoretical perspective. *Journal of Health Organization and Management, 20*(4), 294–308.

King, D. L., Ben-Tovim, D. I., & Bassham, J. (2006). Redesigning emergency department patient flows: Application of lean thinking to health care. *Emergency Medicine Australasia, 18*(4), 391–397.

Kreindler, S. A. (2008). Watching your wait: Evidence-informed strategies for reducing health care wait times. *Quality Management in Health Care, 17*(2), 128–135.

Kreindler, S. A. (2010). Policy strategies to reduce waits for elective care: A synthesis of international evidence. *British Medical Bulletin, 95*, 7–32.

Leonard, C. (1993). An evaluation of the prioritisation of referrals by Leeds social services senior occupational therapists. *British Journal of Occupational Therapy, 56*(12), 448–450.

Lowe, D., & Barber, L. (2005). Implementing a priority and waiting list system in an acute hospital setting. *International Journal of Therapy and Rehabilitation, 12*(7), 294–298.

Lynch, M. E., Campbell, F., Clark, A. J., Dunbar, M. J., Goldstein, D., Peng, P., et al. (2008). A systematic review of the effect of waiting for treatment for chronic pain. *Pain, 136*(1–2), 97–116.

Lynch, G., & Hedderman, E. (2006). Tackling a long waiting list in a child and adolescent mental health service. *Irish Journal of Psychological Medicine, 23*(3), 103–106.

Maddison, P., Jones, J., Breslin, A., Barton, C., Fleur, J., Lewis, R., et al. (2004). Improved access and targeting of musculoskeletal services in northwest Wales: Targeted early access to musculoskeletal services (TEAMS) programme. *British Medical Journal (Clinical Research Ed.), 329*(7478), 1325–1327.

Miller, A. R., Armstrong, R. W., Masse, L. C., Klassen, A. F., Shen, J., & O'Donnell, M. E. (2008). Waiting for child developmental and rehabilitation services: An overview of issues and needs. *Developmental Medicine and Child Neurology, 50*(11), 815–821.

Molyneux, E., Ahmad, S., & Robertson, A. (2006). Improved triage and emergency care for children reduces inpatient mortality in a resource-constrained setting. *Bulletin of the World Health Organization, 84*(4), 314–319.

Murray, M., & Berwick, D. M. (2003). Advanced access: Reducing waiting and delays in primary care. *Journal of the American Medical Association, 289*(8), 1035–1040.

O'Cathain, A., Webber, E., Nicholl, J., Munro, J., & Knowles, E. (2003). NHS direct: Consistency of triage outcomes. *Emergency Medicine Journal, 20*(3), 289–292.

Oredsson, S., Jonsson, H., Rognes, J., Lind, L., Goransson, K. E., Ehrenberg, A., et al. (2011). A systematic review of triage-related interventions to improve patient flow in emergency departments. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 19*(1), 43.

Oudhoff, J. P., Timmermans, D. R., Knol, D. L., Bijnen, A. B., & van der Wal, G. (2007). Waiting for elective general surgery: Impact on health related quality of life and psychosocial consequences. *BMC Public Health, 7*(164).

Parkin, A., Frake, C., & Davison, I. (2003). A triage clinic in a child and adolescent mental health service. *Child and Adolescent Mental Health, 8*(4), 177–183.

Passalent, L. A., Landry, M. D., & Cott, C. A. (2010). Exploring wait list prioritization and management strategies for publicly funded ambulatory rehabilitation services in Ontario, Canada: Further evidence of barriers to access for people with chronic disease. *Healthcare Policy, 5*(4), e139–e156.

Pizer, S. D., & Prentice, J. C. (2011). What are the consequences of waiting for health care in the veteran population? *Journal of General Internal Medicine, 26*(Suppl 2), 676–682.

Porter, J., & Jamieson, R. (2012). Triaging in dietetics: Do we prioritise the right patients? *Nutrition and Dietetics, 70*, 21–26.

Rastall, M., & Fashanu, B. (2001). Hospital physiotherapy outpatient department waiting lists: A survey. *Physiotherapy, 87*(11), 563–572.

Rotstein, D., & Alter, D. (2006). Where does the waiting list begin? A short review of the dynamics and organization of modern waiting lists. *Social Science and Medicine, 62*(12), 3157–3160.

Russell, K. L., Holloway, T. M., Brum, M., Caruso, V., Chessex, C., & Grace, S. L. (2011). Cardiac rehabilitation wait times: Effect on enrolment. *Journal of Cardiopulmonary Rehabilitation and Prevention, 31*(6), 373–377.

Sethuraman, U., Kannikeswaran, N., et al. (2011). Effect of a rapid assessment program on total length of stay in a pediatric emergency department. *Pediatric Emergency Care, 27*, 295–300. Oxford: United Kingdom.

Streiner, D., & Norman, G. (2003). *Health measurement scales: A practical guide to their development and use* (3rd ed.). Oxford University Press.

Tanabe, P., Gimbel, R., Yarnold, P. R., Adams, J. G., Tanabe, P., Gimbel, R., et al. (2004). The Emergency Severity Index (version 3) 5-level triage system scores predict ED resource consumption. *Journal of Emergency Nursing, 30*, 22–29.

Tideman, R., Pitts, M., & Fairley, C. (2003). Effect of a change from an appointment service to a walk-in triage service at a sexual health centre. *International Journal of STD and AIDS, 14*, 793–795.

Vissers, J., & Beech, R. (Eds.). (2005). *Health operations management* (1st ed.). London: Routlege.

Walters, E. H., & Dawson, D. J. (2009). Whole-of-hospital response to admission access block: The need for a clinical revolution. *Medical Journal of Australia, 191*(10), 561–563.

Williams, I., Robinson, S., & Dickinson, H. (2012). *Rationing in health care: The theory and practice of priority setting*. Chicago: The Policy Press.

Woodhouse, A. E. (2006). Reducing waiting times: Using an opt-in system and changing prioritisation criteria. *Child and Adolescent Mental Health, 11*(2), 94–97.

Wright, C., & Ritson, E. (2001). An investigation into occupational therapy referral priorities within Kensington and Chelsea social services. *British Journal of Occupational Therapy, 64*(8), 393–397.

Part IV
Modeling and Analysis Tools

Chapter 11
Personnel Staffing and Scheduling

Michael Warner

Abstract Not only do personnel make up approximately 2/3 of the cost of hospital health care, but they also account for an even higher percentage of the quality of care delivered, and patient safety. Additionally, they, with physicians, are the major determinate of how quickly a patient moves through a hospital. The key factor for throughput, quality, and safety, is that the correct number of personnel, with the correct qualifications and correct motivation, be present at the right times and places of the patient's stay. This is the goal of the personnel staffing and scheduling processes and systems in hospitals. This chapter first briefly reviews the history of the use of modeling for more efficient and effective scheduling and staffing. Then it presents in some detail modeling work to move staff management decisions from the present, where intervention options are severely limited, into the near future (several days) where intervention options are numerous. This significantly improves not only throughput but also quality, safety, and staff satisfaction. This new effort involves the following: (1) Forecasting demand for staff into the near future, (2) Predicting no-shows of prescheduled personnel, and (3) A robust decision support system to include professional judgment and "best practices" allowing the hospital to be able to take advantage of the significant increase in intervention options for staffing (for example) 4 days ahead rather than 4 h ahead. How this effort fits into the larger aspects of staff management is discussed, along with possible future opportunities to use modeling to improve staff management in hospitals.

Keywords Scheduling • Staffing • Forecasting

The concepts, techniques, procedures, and computer systems for (1) moving staffing decisions into the future and (2) Event Driven Workload have Patent Pending status owned by AtStaff Inc. of Durham, NC.

M. Warner (✉)
AtStaff, Inc., 3000 Croasdaile Drive, Suite 100, Durham, NC 27705, USA
e-mail: Warnermich@aol.com

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_11, © Springer Science+Business Media New York 2013

1 Introduction

A critical element in a patient's moving through the health care system as quickly as possible is that the correct number of caregivers, with the correct qualifications, and with the correct attitudes and motivations are available at each phase of the patient's stay. Caregivers are the assembly line managers of health care. Institutions and physicians set protocols to follow, procedures to be performed, etc., but the minute-by-minute execution of health care is primarily performed and managed by caregivers.

Thus, if there is a theoretical "shortest time" that a patient A must stay in the system, achieving such time is largely a function of the environment the patient is in, a set of things that must be done (correctly and at the right time), plus a set of things that must not be done (that can cause delays). Having the right number of motivated caregivers who set the environment, know what to do (and how to do it) and what not to do is critical to minimizing delay.

In addition to throughput, correct matching of the supply of caregivers to demand also affects patient care quality, patient safety, and staff satisfaction (Litvak et al. 2005; Needleman et al. 2002; Kovner and Gergen 1998). These three critical outcomes in turn affect throughput.

Having the optimal number and skill of caregivers available at the right time and place is the goal of good scheduling and staffing. While shift-by-shift staffing is where the real action is, there are a number of steps that must take place to set the stage. The better these earlier decisions are done, the better the chance that optimal shift-by-shift staffing can occur.

2 Definitions

Before introducing the several aspects of Personnel Management, it would be useful to define certain terms as they are used in this chapter.

- "Nurse" will be used as a stand-in for all types of caregivers and other personnel in hospitals, such as pharmacists, radiologists, transport people, etc. Nurses are indeed the most numerous caregivers, plus they are the most difficult to schedule and staff.
- "Skill" will refer to the caregiving abilities of the nurse. At its most basic, it refers to what they are licensed for—registered nurse, nursing aide, senior technician, etc. But for our purposes, when possible it will also cover qualifications, experience, attitudes, motivation, etc.
- "Unit" will refer to a logical unit or cost center of the hospital, where a group of personnel call "home". Examples are the ICU, the ER, the OR, a medical nursing unit, the pharmacy, etc. Again, most examples will be nursing units, but the methodologies discussed apply to any unit.

- "Shift" will refer to an interval of time within the day. Most typical for nursing units are 8-h shifts, or 12-h shifts, but 10-h, 4-h, or other length of time might make up a "shift." Personnel may work a combination of shifts within a unit (such as 12-h and 8-h), and different personnel may be working different shifts at the same time (some on 4-h, some on 8-h, etc.). When measuring demand for care, time slices as short as an hour may be needed (e.g., the ER), or variable time slices defined by events (see the end of chapter). All such "time slices," whether variable or fixed, will be referred to as "shift" in this chapter.
- "Demand" for personnel refers not only to the *number* of people needed, but also the *mix* of skills, qualifications, attitudes, and experience that together is needed to give optimal care to a patient or a set of patients.
- "Supply" has the same dimensions as demand, but will usually include the actual personnel who make up the team that will provide care.

Additional definitions will be provided prior to the discussions in later sections of this chapter.

3 The Stages Leading to Staff Scheduling

There are several stages involved in determining the correct number of qualified caregivers being present to care for a particular set of patients on a particular shift.

3.1 Determining a Measure of Demand for Staff

The first step is to establish a method of measuring the demand for caregivers based on the number and type of patients present to be cared for. The most basic measure is Nursing Hours per Patient Day ("NH/PD"), different at least by the type of Unit (Medical, ICU, Peds, etc.) where the patient is staying. (In radiology, it would be minutes of each skill or type of personnel required for each type of procedure). NH/PD may then be divided within the day by percentages per shift to create Nursing Hours per Patient Shift (NH/PS).

A step up in refinement on NH/PD and NH/PS is to classify patients into several (typically between three and seven) patient classes, where Class 1 requires X amount of time during the day shift (for example), Class 2 requires Y ($Y > X$), etc. For long-term (seasonal, annual) decision making, a typical mix of patients by class is then forecasted for a unit and shift to move to the next step (immediately below). (An even more granular method for measuring demand called "Event Driven Workload" is discussed at the end of the chapter).

3.2 Determination of Core Staffing ("Core Staffing")

Core staffing for a unit (such as a nursing unit) is the number of personnel by skill needed to provide optimal care to the average number of patients at an average acuity or need level, or an average mix of patients by class. For example, it may be determined that on nursing unit 6-West, on the day shift weekdays in the winter, on average we need a head nurse, three RNs, an LPN, three nurse aids, and a unit secretary. (Core staff will be different on the other shifts, probably different on weekends, and possibly different in other seasons). Of course, core staffing is to meet average demand: adjusting to a known level of demand is addressed below.

3.3 Determining Positions that Provide Core Staffing ("Position Control")

In order to have staff available to meet core staffing 24 h a day and 7 days a week, a certain number of positions (by skill and unit) must be budgeted. This number includes the fact that coverage is 24/7, full time staff work 40 h a week, some will be part time, vacation, orientation, professional development, sick time, turnover, etc. The result of this phase is a list of positions needed, by skill and unit, into which personnel are hired.

3.4 Recruiting and Hiring

Staff must be recruited and hired to fill those positions (a not insignificant task!).

3.5 Long Range Scheduling ("Scheduling")

Approximately 3 weeks before it is to start, the long range (4–8 weeks) schedule is determined, considering a host of factors such as equal weekend time, shift rotation, number of days worked in a row, special requests, scheduled vacation, etc. Typically, making the long range schedule manually can take a week, and it is typically published 2 weeks before it starts, so the best scheduling can do to try to have the right number and skill of staff is to schedule to fixed core staffing, knowing that actual demand will be different once that future day and shift arrives. Most typically for nursing, a new long range schedule must be rebuilt every 4 weeks (or the length of the schedule), although some institutions use "cyclical" schedules that cycle repeat over some cycle (multiple of length of schedule). For non-nursing, cyclical or fixed schedules are common.

3.6 Shift Staffing ("Shift Staffing")

Typically, several hours before a shift begins, a nurse manager attempts to determine the need for staff for that shift (or perhaps for the next two or three Shifts). She compares this with who is scheduled (from the long rage schedule) to come in, and makes decisions on how to adjust supply to meet demand. Recall that a shift may be of any length of time, so shift staffing takes place on whatever schedule the decision maker must use to ensure that demand is met by how shifts are defined for that unit.

This decision process—shift staffing—is the main topic of this chapter. After a brief history of the modeling of the scheduling problem and automation of staffing, we will focus on the difficulties of determining demand in the near future, and the difficulties of adjusting supply to meet demand (calling extra nurses in, moving a nurse from one unit to another, or arranging for a nurse to *not* come in).

4 Modeling the Scheduling and Staffing Decisions

Viewed from a modeling perspective, the scheduling and staffing decisions are quite different.

4.1 Scheduling

For scheduling, demand is well defined: it is core staffing, and has been previously determined. Supply is also known, defined as the 30 or so individuals that will be available to work these 4 weeks, their workloads (full time or part time), and their special rules such as only work on day shift or evening shift, no more than 40 % on evenings, work a maximum of 4 days in a row ("work stretch"), every other weekend off, etc. This makes scheduling a difficult problem to fit into an optimization solution technique, but does make it possible to formulate for optimization.

Figure 11.1 shows a 4-week schedule generated by software that "solves" the scheduling problem. Variables are Nurse N (by name) working on Shift B on Day D, and constraints are (1) a minimum and maximum number of nurses of Skill S for each shift and day, (2) that Nurse N works exactly her workload (e.g., five times a week), (3) that she works only on *her* shifts (e.g., day and evening), (4) no more than 4 days in a row, and (5) a host of other work constraints. The objective function is a combination of minimizing how far off actual staffing is to core staffing, and maximizing the "quality" of the nurses' individual schedules as defined in terms of things such as work stretch, rotation (working on a shift other than her "home" shift), requests, etc. Formulated correctly, the objective function contains many nonlinear items (e.g., a shortage of two nurses is much greater than twice the shortage of one, a work stretch of 8 days in a row is much worse than twice 4 days in a row, etc.), variables are integer, and thus the problem does not

Four Week Schedule for 8/1/2005 - 8/28/2005

Emp	Skill	8/1	8/2	8/3	8/4	8/5	8/6	8/7	8/8	8/9	8/10	8/11	8/12	8/13	8/14	8/15	8/16	8/17	8/18	8/19	8/20	8/21	8/22	8/23	8/24	8/25	8/26	8/27	8/28
		Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su
Arch, Kerry	LPN	7a-3p LPN	7a-3p LPN						7a-3p LPN	7a-3p LPN						7a-3p LPN	7a-3p LPN						7a-3p LPN	7a-3p LPN					
Arn, Kelly	RN		7a-3p RN	7a-3p RN	3p-11p	3p-11p Charge	11p-7a Charge		11p-7a Charge	7a-3p RN	11p-7a	3p-11p Charge			3p-11p			11a-3p RN	3p-11p Charge	7a-3p Charge	3p-11p Charge	3p-11p	11p-7a Charge	7a-3p Charge	11p-7a Charge	11a-3p RN		3p-11p	
Cambell, Chris	NA		7a-3p NA	7a-3p NA	7a-3p NA	7a-3p NA	7a-3p NA		7a-3p NA	7a-3p NA	7a-3p NA	7a-3p NA	7a-3p NA			7a-3p NA	7a-3p NA	7a-3p NA	7a-3p NA	7a-3p NA			7a-3p NA	7a-3p NA	7a-3p NA	7a-3p NA	7a-3p NA		
Entz, Mark	NA	7a-3p NA			7a-3p NA				7a-3p NA	7a-3p NA						7a-3p NA	7a-3p NA						7a-3p NA	7a-3p NA					
Favro, Jill	LPN	7a-3p LPN		7a-3p LPN	3p-11p LPN	7a-3p LPN			3p-11p LPN	3p-11p LPN	3p-11p LPN	7a-3p LPN						3p-11p LPN	7a-3p LPN				3p-11p LPN		7a-3p LPN		3p-11p LPN		
Gonzalez, Marbella	NA		11p-7a NA			3p-11p NA		7a-3p NA		11p-7a NA	11p-7a NA			7a-3p NA		3p-11p NA			7a-3p NA	11p-7a NA	7a-3p NA	3p-11p NA	3p-11p NA	11p-7a NA		7a-3p NA	11p-7a NA	3p-11p NA	
Hansen, Jane	LPN	7a-3p LPN	7a-3p LPN	7a-3p LPN				3p-11p LPN						7a-3p LPN	7a-3p LPN		3p-11p LPN			3p-11p LPN		3p-11p LPN						7a-3p LPN	
Hiebert, Maurene	NA	11p-7a NA			3p-11p US	3p-11p US	7a-3p US	7a-3p US		3p-11p NA	3p-11p NA	3p-11p NA		7a-3p US	7a-3p US	3p-11p US	3p-11p US	3p-11p US			7a-3p US	7a-3p US		3p-11p US	3p-11p US	3p-11p US		7a-3p US	7a-3p US
Higgins, Rebbeca	RN	3p-11p			11p-7a	11p-7a	7a-3p RN	7a-3p RN						Off Off	Off Off	11p-7a	11p-7a				7a-3p RN	7a-3p RN			3p-11p	11p-7a		Off Off	Off Off
Jaderborg, Joy	RN	11p-7a Charge	3p-11p	3p-11p			7a-3p RN	7a-3p RN		3p-11p	7a-3p RN	3p-11p	3p-11p	Off Off	Off Off	11p-7a Charge	11p-7a Charge	11p-7a			7a-3p RN	7a-3p RN		11p-7a Charge	3p-11p	7a-3p RN		11p-7a Charge	11p-7a
Jankord, Anna	RN		3p-11p	3p-11p		11p-7a	Off Off	Off Off	11p-7a		11p-7a			7a-3p RN	7a-3p RN			11p-7a	3p-11p		Off Off	Off Off	11p-7a			11p-7a	3p-11p	7a-3p RN	7a-3p RN
Jansen, Silvia	RN		3p-11p	11p-7a	11p-7a		Off Off	Off Off	11p-7a	11p-7a	11p-7a			7a-3p RN	7a-3p RN			11p-7a		3p-11p	Off Off	Off Off	11p-7a	3p-11p			11p-7a	7a-3p RN	7a-3p RN

Fig. 11.1 Four-week schedule showing when each employee (*left column*) is scheduled to work. This is the first page of a three-page schedule, for approximately 30 employees on a particular unit

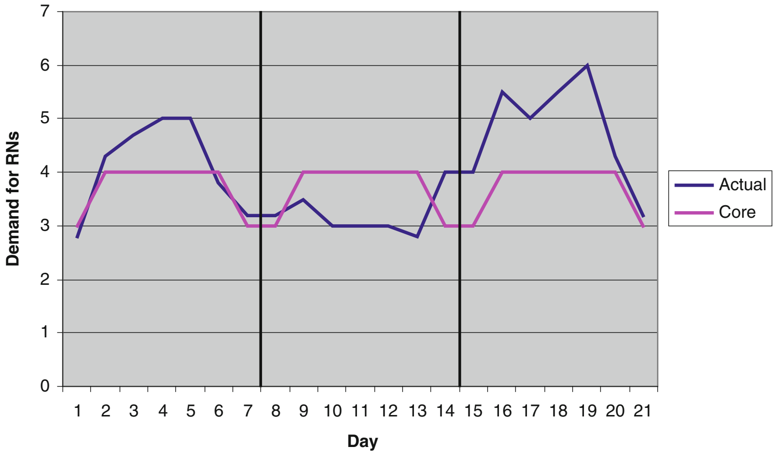


Fig. 11.2 Demand for RNs for the day shift over a 3-week period. Core staffing follows a "fixed" pattern, dipping on the weekends. Actual demand follows the number of patients actually on the unit and for each a measure of how much RN time they require

readily fit an optimization solution technique. Most often an individualized heuristic solution technique (usually of the "branch and bound" type) is designed to find several "good" solutions, if not the optimal to the scheduling problem.

4.2 Shift Staffing

The shift staffing decision, on the other hand, is much fuzzier, and is stochastic rather than deterministic. Here we're dealing with many aspects that are unknown, the three most difficult being (1) what actual demand will be the next shift or two or three, (2) who is able (and willing) to be called in, and (3) who among those Scheduled to come in will not "show" (for whatever reason). The most difficult of these is what demand will be next shift (or two or three shifts).

Figure 11.2 shows the *actual* behavior of demand, versus core staffing, for RNs for the day shift over a 3 week period for a typical nursing unit in a hospital. (This is actually a "moderately" variable example such as a medical or surgical unit: an ICU would show much more variation and a psych unit much less). Clearly core staffing, while useful for approximating demand for recruiting, position control, and scheduling, is not going to closely match actual demand for the shift staffing decision.

The shift staffing decision in hospitals today is typically done for the next shift or two (sometimes a weekend), using best professional estimates of what demand will be, what supply will be, and calling in nurses from previously established pools of employees or free agents to fill expected holes in staffing. What results is a fair amount of chaos, and unhappy employees being called in at the last minute. More seriously, compromises in staffing must be endured because of the lack of choices of nurses to come in, based on lack of knowledge of demand beyond this shift.

The modeling of the shift staffing decision will require dealing with the uncertainty of demand and supply, will be limited to the "near future" (0–8 days ahead), and will require building in professional judgment to supplement what will always be estimates (projections and predictions) of supply and demand.

5 History of Automated Scheduling and Staffing

The scheduling problem was initially modeled and solved in the late 1970s (Warner, 1976; Warner et al. 1991), and by the early 1980s two companies were offering automated staff scheduling systems on personal PCs that included position control, a scheduler that "generated" a good quality 4 week schedule (saving many hours of unpopular work by nurse managers), and a *framework* for shift staffing. This framework was a huge help by providing accurate data, accurate calculations, and numerous "roster" type reports of who will work when and where. The weakness remained the reliance of a mildly retrospective (4–12 h old) measurement of demand, and no support for pushing the decision more than a shift or two into the future.

The automation of scheduling and support for staffing was a large success—by the 1990s, some 60–70 % of hospitals over 100 beds were using some sort of computer system. Scheduling took much less time, was accurate and more "fair" to those scheduled, and very useful reports were produced. Staffing support was also a hit, increasing accuracy and communications with easily printed rosters and reports.

But until 2005 none of the commercially available systems had approached the shift staffing problem as a probabilistic model, required for pushing the decisions into the near future. Even with a helpful (deterministic) framework, decision makers were stuck in the present.

New computer technology has of course played a significant role in automated staff scheduling systems. The first systems were on PCs with dual floppies, then hard disks (with 10 MB of storage!). In the mid to late 1980s, systems became networked. Then in the late 1990s the Internet added new value by allowing access to individual employees, managers, and executives wherever they happened to be. One current system is designed in a "job centric" mode, where employees "see" the system on the Internet in a way that is relevant to them, managers "see" another view, and there are separate views or "dashboards" for the staffing office and for executives.

6 Moving Shift Staffing into the Near Future

The remainder of this chapter focuses on a significant improvement to the shift staffing decision—moving shift staffing from its present 4–8 *h* ahead to 4–8 *days* ahead. The benefits of this redefinition of the problem are:

1. Much greater flexibility in deciding who will be called in: there may be only one or two possible staff to call in with a 4-*h* notice, where there may be 20 or 30 who

can be called in with a 4-*day* notice. This flexibility allows the decision maker to use criteria such as skills, experience (on this unit or total experience), cost, qualifications, attitudes, willingness to come in, how often a person was called before, etc.

2. Higher satisfaction of staff that are called in: they now get notice in time to arrange work time with personal time, which is difficult with short notice. It gives staff much more of a say in their work time—a major concern of 24/7 employees.

This redefinition will involve three elements that will be discussed in turn:

1. Forecasting demand into the near future
2. Predicting "no shows" (staff scheduled to come in but for some reason will not show up)
3. A decision support system that supports the probabilistic nature of the environment of the decision, allowing professional judgment and "best practices" to balance the accuracy of the forecasts and predictions.

7 Forecasting Demand: Projection and Prediction

The model uses a "best information available" approach to forecasting demand, combining *projecting* the behavior of need by current patients until they leave the system, and as they leave backfilling with new admissions based on available future information (where available) and *predicted* census. The result is a simulation of what is expected to happen in terms of demand for caregivers.

7.1 Projecting Demand

Projecting demand is the use of information about what "phase" each patient is in within their typical "care pattern of need over time" within the hospital (their "care pattern"), and then projecting that patient through the remaining phases of their care pattern. For example, a total hip replacement might have a typical "caregiver need" care pattern defined in terms of need for RN care, as in Fig. 11.3.

At any point in time from Fig. 11.3, the *remaining* care pattern will project the care needed for this patient type until this patient is discharged or transferred out. By projecting each patient by their predetermined care pattern, and summing over a time slice (such as a shift or an hour on the second day in the future), the total demand for RNs can be projected for that time slice for all present patients who have yet to leave the system.

The predetermined care patterns are established by:

1. Defining discrete "types" of patients (for example, total hip replacement for patients under 70 years old).

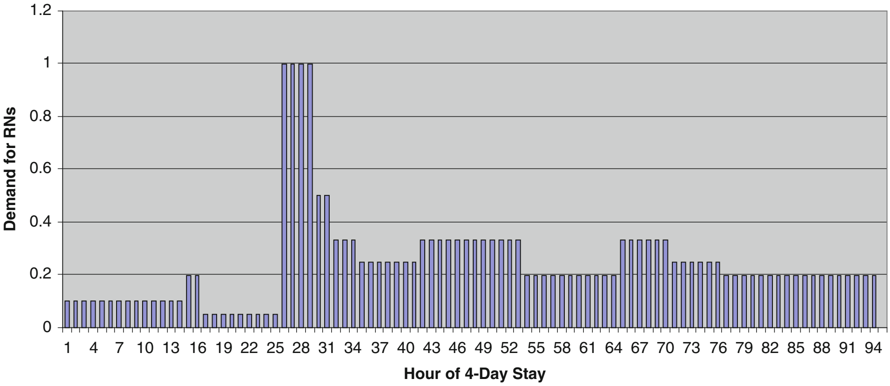


Fig. 11.3 "Total Hip" demand care pattern. Demand for RNs of a typical total hip replacement, by hour of a 4-day stay. Hours 1–14 (Phase 1) have low care need as patient is admitted in PM for surgery next day. Hours 15–16 (Phase 2) are OR prep, then patient is out of unit between hours 17 and 25. Hours 26–29 are immediately after arriving back on unit, etc.

2. Using a suitable sample of patients of this type to define the phases they go through, how much caregiver time is needed at each phase, and building a "typical" care pattern for patients of this type.

As many patient types as desired can be defined, with accuracy improving (with diminishing returns) with more types. For a surgical unit, between 6 and 12 types may return sufficient accuracy for the application described herein.

Each phase in the care pattern of need will be in terms of caregiver demand measured (from less to more granular) by:

- Nursing hours per patient day for that phase (corrected for the length of each state—8 h, 4 h, etc.)
- What patient "Class" this type of patient can be expected to be in this phase
- A measurement based on *events* that are expected to happen to this type of patient over time, where changing from one need level to another is an event, as well as certain procedures, etc. (See Event Driven Workload at the end of the chapter).

The above measure of demand will give *direct* care. To direct care must be added caregiver time on the unit not devoted to any one patient. This is typically done as a percentage of time, or a fixed number of minutes per person.

(The "care pattern" type of projection will give the best results in terms of projecting demand, but a less precise "length of stay" model may be used where care patterns cannot be established. This model will again look at all patients currently on a unit, but instead of a demand care pattern, only an estimate of the remaining length of stay is made. This is used to project the census of current patients until they leave, and is blended with the census prediction model below to forecast total census. Then a NH/PD or NH/PS or other method can be applied to transform census into demand).

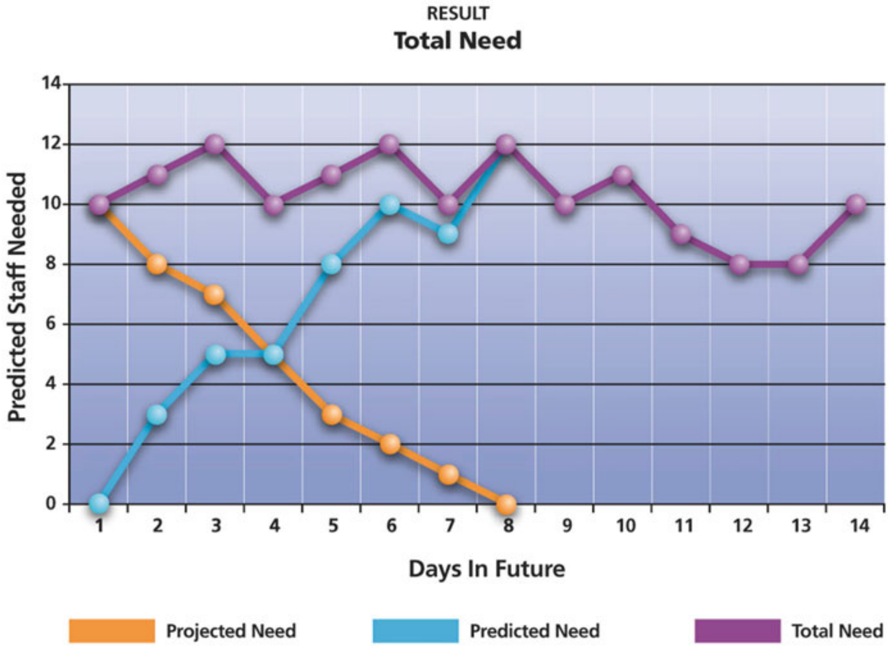


Fig. 11.4 Contribution of projection of current patients and prediction of new admissions over forecasting horizon. In Days 1–3, most of the forecast is from current patients. After Day 4, when many have been discharged, more of the forecast is from predicted new admits (following their need behavior). At some point (Day 8), all current patients have left the system

7.2 Predicting Census and Admissions

Using the projection of a given number of patients starting at time zero, the unit will "lose" patients as they are discharged or transferred out. These will be "replaced" in the simulation model with a certain number of new admissions (and in-transfers) each shift. The number of new admissions and in-transfers is determined by either:

- Using prescheduled admissions data (including the OR schedule and other schedules—see discussion below), or
- Predicting census (Fig. 11.4) for that unit and shift, and comparing predicted census to projected census (total patients still in some phase of their care pattern this shift), and base predicted admissions on the difference. For example, if predicted census is 33 and projected census is 30, the model "admits" three more patients. These new admits have their own care pattern definition ("new admit"), and the model then projects (simulates) these patients' care pattern as it does for the other patients.

Even if prescheduled admission data is available, emergency admissions will need to be predicted to add to scheduled admissions. The model for predicting emergency admissions is more accurate of:

- Correlating emergency admissions to projected (simulated) census, subject to minimum and maximum emergency admissions based on historical data, or
- Predicting total census directly (see below) and subtracting from it projected census (including the pre scheduled admissions).

The census prediction model is a bit more complex than projection, as there is a host of factors that influence census. Considering that the model will be predicting census for the near future of a few days, many variables that are typically used to predict census for longer periods (growth, changes in length of stay, seasonality, weather, etc.) were deemed to not be appropriate. Instead, the census prediction model uses pattern and autocorrelation type methods.

No one predictor method fits each unit (think of the difference between ICU and psych), or even how far out the prediction is being made (autocorrelation methods are superior in the very near future, while pattern type predictors—such as day of week correctors—are superior after a few days.) The "best" census prediction model for Unit U on Shift S on a future Day D is determined by:

1. Establishing a measure of "better". In this case, a weighted "score" of how often a particular method generates demand from its census prediction that is within 0.5 caregivers, is within 1.5 caregivers, is within 2.5 caregivers, or is off by more than 2.5 caregivers is calculated. The weights placed on the errors to make up a "score" are established using professional judgment of the "cost" of a projection being off that much. This cost is a balance of making decisions too soon (possibly wasting money) versus making them too late (and not having as many choices).
2. Establishing a "pool" of "candidate" predictors that an evaluation program can use as variables. For example, average census is a "candidate," as is average census corrected for day of week. So is the prior day's census, the average of the last *x* days (*x* varying between 2 and, say 7), and a weighted average of the last *x* days, where the weights themselves define a new variable.
3. Using a special-purpose branch-and-bound search algorithm that intelligently "tries" thousands of possible *combinations* of candidate predictors on 1 or 2 years of past data for this unit, and picks the one that would have produced the best score on this data using the criteria of one.

More specifically, candidate engines to be evaluated for a Unit u for a particular Shift s in the future are in the form of:

$$F_{u,s} = A_{u,s}{}^{*}OA_{u}{}^{*}DOW_{u,s} + B_{u,s}{}^{*}\left(Z_{i=LT}W_{u,s,i}C_{u,s,i}/DOW_{u,i}\right)^{*}DOW_{u,s}$$

Where:

$F_{u,s}$ = the forecasted census for a Shift s on Unit u in a future forecasting horizon

Subject to:

$F_{u,s} < MAX_{u,s}$, where $MAX_{u,s}$ is the maximum census should be for Shift s on Unit u
$F_{u,s} > MIN_{u,s}$, where $MIN_{u,s}$ is the minimum census should be for Shift s on Unit u

A and B are weights that balance the day-of-week ("DOW") effect with the autocorrelation effect, such that:

A and B are between 0 and 1, and $A + B = 1$
OA_u is the overall average census for the Unit for this "season"
$DOW_{u,s}$ is a day of week factor which "corrects" OA_u for DOW on which s falls
$i = 1,T$ is a shift in the immediate past, and T the number of such shifts looking "backward" from today on which an autocorrelation estimate is to be based
$W_{u,s,i}$ is a weight placed on the DOW corrected census for past Shift i, $i = 1,T$
$ZW_{u,s,i=1,T} = 1$
$C_{u,s,i}$ is the census for that past Shift i
$DOW_{u,i}$ is the DOW "corrector" for the DOW of past Shift i

OA and DOW are constants based on a year or so of data from this unit, and the C is available to the search algorithm from past data. A candidate engine is defined by different values for A, B, and $W_{i,i=1,T}$. Thus, there are an infinite number of candidate engines. The branch-and-bound search engine "intelligently" evaluates different combinations of A, B, and W, and picks the one that produces the best "score" (defined above). The search engine looks "intelligently" at a finite number of candidates for evaluation.

The result is a census predictor "engine" for each unit, and shift of each day in the forecasting horizon (i.e., a 3-day predictor engine for the ICU is different than the 2-day predictor engine). Accuracies of these engines vary widely with the type of unit (not too good for ICU, great for psych, very useful for units in between like medical units and surgical units). In all cases, the predictor is significantly superior to average census, on which core staffing is based.

7.3 Simulation: Combining Projection and Prediction

Thus, forecasting (or simulating) demand for a unit for several days ahead is performed by

1. Projecting all *present* patients out one shift on their care patterns
2. If known, adding prescheduled admissions for this shift with their expected care patterns
3. Predict emergency (or otherwise unknown) admissions by either:
   - Predicting census for that shift, comparing it with projected census, and admitting the difference (subject to minimum and maximum admissions by day of week), or
   - Predicting admissions directly

4. Place each new admission into the care pattern that best represents the admission
5. Sum across all patient care patterns (projected, prescheduled admissions and emergency admissions) to get total demand
6. Add in indirect time, if appropriate
7. Project all patients (the Projected plus the new admits) *another* shift
8. Repeat steps 2–7 until the end of horizon.

7.4 Predicting Admissions Directly from Other Sources

As indicated above, a more accurate prediction of admissions may be available from other sources within the hospital rather than predicting census and subtracting projected census. Fortunately, much of this information is easily obtainable from the hospital's HL7 interface system. For example the Admissions Discharge Transfer (ADT) system may have future information on prescheduled admissions. Scheduled admissions to certain units (elective surgery, etc.) may be a good source for admission predictions, as well as the OR schedule. As mentioned above, to the prescheduled admissions must be added a prediction piece for emergency admissions. Many units mostly serve emergencies to start with, so that scheduled admissions will be less help. The key here is to use the best information available, and to be able to fit the model to the data available.

8 Predicting No-Shows

No-shows are predicted by applying the probability that each scheduled individual does not show up for that particular day of week and shift. This probability is calculated from the known number of times a person did not show in the past, divided by the times they were supposed to show. If the sample size by DOW and shift is not large enough to get an accurate probability for an individual, the fall back is the individual by DOW. The next fall back is the overall probability of a no show for this individual, followed by the average no show by DOW and shift for nurses of her skill on the unit. Such data are readily available from staff scheduling systems, and often from personnel or time and attendance systems.

9 Proactive Protocols

Each decision to place a nurse on duty involves hundreds of dollars, and serious implications on quality, safety, and throughput. So in a probabilistic environment like this one, professional judgment and "best practices" must be incorporated as part of the decision support system.

In this case, *proactive protocols* are developed for each unit, shift, skill, shortage and day in the forecasting horizon which spells out the "optimal" action to be taken. For example, when looking 4 days out on the medical unit, the night shift is forecasted to be short three RNs. (This is from the projection plus prediction of Sect. 7.4, added to the prediction of no shows of Sect. 8.) The proactive protocol for this situation (which is displayed for the decision maker) might be:

1. Call one RN in for the shift.
2. Develop a list of four candidates that could come in if you need another
3. Flag this shift and check again tomorrow

Tomorrow, the forecast will cover 3 days, with higher expected accuracy, and only two RNs may be projected to be needed. In that case, the Protocol may say secure the second one, or wait. The Protocols are constructed by the expert staffers in the hospital to balance the cost of calling in too many with the costs of waiting until the last minute. By publishing such Protocols for the decision maker, "best practices" will evolve over time as lessons are learned.

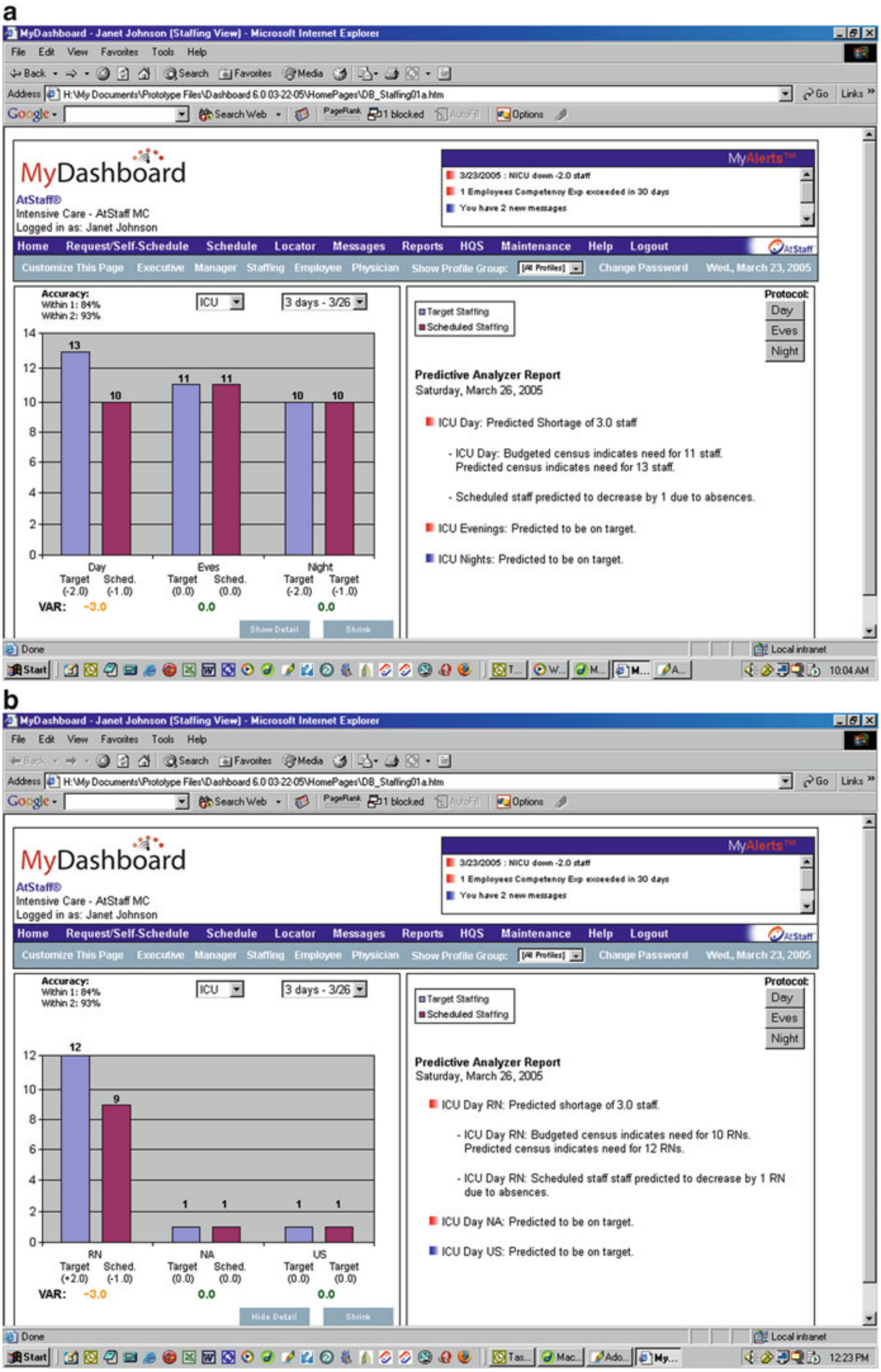
10 The Staffing System

Recall that the traditional staff scheduling system has available most of the information needed to make the shift staffing decision:

1. Who is scheduled to come in each day and shift.
2. Their skills, qualifications, etc.
3. Minima and maxima of how many to staff by skill
4. A framework for adding to or adjusting rosters on screen
5. Lists of staff by skill, availability, with telephone numbers, etc.
6. Other personnel data to make the decisions

What is added here is the demand forecasting and no show predictions, along with the proactive protocols to enhance the framework and allow it to be "pushed" into the near future.

Figure 11.5 shows projected staffing need (incorporating demand forecasting and no show predictions) for the ICU unit for 3 days ahead. The decision maker can look at all three shifts (Fig. 11.5a), or drill down to one shift by skill (Fig. 11.5b), then bring up the proactive protocol for this situation (Fig. 11.5c). From this screen certain actions may be taken such as pulling up lists of nurses by criteria (availability, qualifications, etc.), and posting the need to the internet dashboards of certain staff members. Call-ins are made by phone or email (and/or through the system's communication system), and adjustments made to staff on the screen. Any adjustments immediately update the situation, bringing up a new proactive protocol for the new situation.


a
MyDashboard - Janet Johnson (Staffing View) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address H:\My Documents\Prototype Files\Dashboard 6.0 03-22-05\HomePages\DB_Staffing01a.htm
Go Links
Google Search Web PageRank 1 blocked AutoFill Options
MyDashboard
AtStaff®
Intensive Care - AtStaff MC
Logged in as: Janet Johnson
MyAlerts™
Home Request/Self-Schedule Schedule Locator Messages Reports HQS Maintenance Help Logout
Customize This Page Executive Manager Staffing Employee Physician Show Profile Group: [All Profiles] Change Password Wed., March 23, 2005
Accuracy:
Within 1: 84%
Within 2: 93%
ICU
3 days - 3/26
14
12
10
8
6
4
2
0
13
10
11
11
10
10
Day
Target (-2.0)
Sched. (-1.0)
Eves
Target (0.0)
Sched. (0.0)
Night
Target (-2.0)
Target (-1.0)
VAR: -3.0
0.0
0.0
Show Detail
Shrink
■ Target Staffing
■ Scheduled Staffing
Protocol:
Day
Eves
Night
Done
Local intranet
Start
10:04 AM

- 3/23/2005 : NICU down -2.0 staff
- 1 Employees Competency Exp exceeded in 30 days
- You have 2 new messages

Predictive Analyzer Report
Saturday, March 26, 2005

■ ICU Day: Predicted Shortage of 3.0 staff

- ICU Day: Budgeted census indicates need for 11 staff. Predicted census indicates need for 13 staff.

- Scheduled staff predicted to decrease by 1 due to absences.

■ ICU Evenings: Predicted to be on target.

■ ICU Nights: Predicted to be on target.

b
MyDashboard - Janet Johnson (Staffing View) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address H:\My Documents\Prototype Files\Dashboard 6.0 03-22-05\HomePages\DB_Staffing01a.htm
Go Links
Google Search Web PageRank 1 blocked AutoFill Options
MyDashboard
AtStaff®
Intensive Care - AtStaff MC
Logged in as: Janet Johnson
MyAlerts™
Home Request/Self-Schedule Schedule Locator Messages Reports HQS Maintenance Help Logout
Customize This Page Executive Manager Staffing Employee Physician Show Profile Group: [All Profiles] Change Password Wed., March 23, 2005
Accuracy:
Within 1: 84%
Within 2: 93%
ICU
3 days - 3/26
12
10
8
6
4
2
0
12
9
1
1
1
1
RN
Target (+2.0)
Sched. (-1.0)
NA
Target (0.0)
Sched. (0.0)
US
Target (0.0)
Target (0.0)
VAR: -3.0
0.0
0.0
Hide Detail
Shrink
■ Target Staffing
■ Scheduled Staffing
Protocol:
Day
Eves
Night
Done
Local intranet
Start
12:23 PM

- 3/23/2005 : NICU down -2.0 staff
- 1 Employees Competency Exp exceeded in 30 days
- You have 2 new messages

Predictive Analyzer Report
Saturday, March 26, 2005

■ ICU Day RN: Predicted shortage of 3.0 staff.

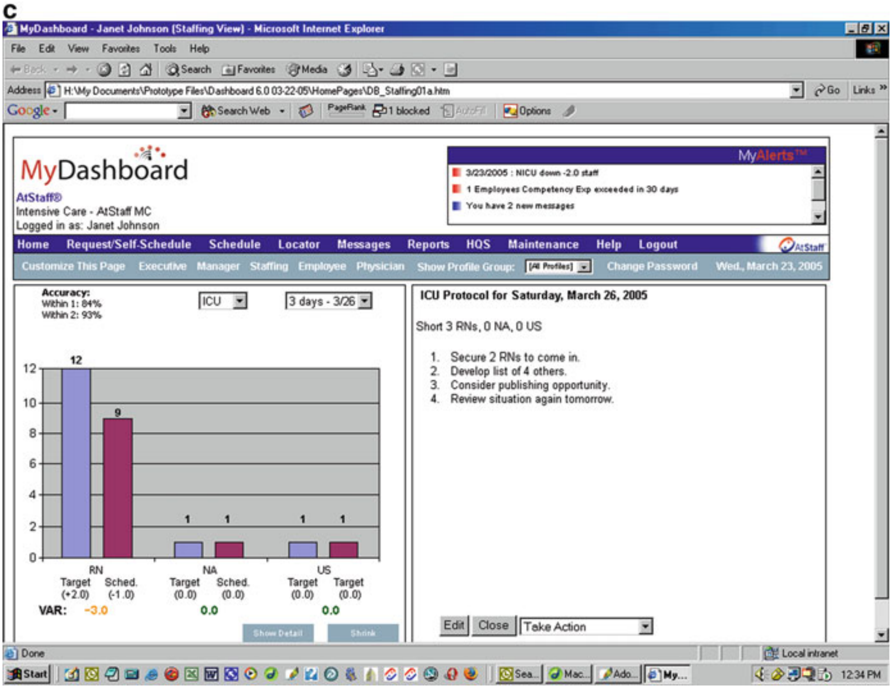
- ICU Day RN: Budgeted census indicates need for 10 RNs. Predicted census indicates need for 12 RNs.

- ICU Day RN: Scheduled staff staff predicted to decrease by 1 RN due to absences.

■ ICU Day NA: Predicted to be on target.

■ ICU Day US: Predicted to be on target.

Fig. 11.5 (**a**) Forecasted staffing 3 days ahead for the ICU. To left is a comparison, by Shift, of forecasted need (*in blue*), versus what is predicted to come in (*red*). For the day shift, the system is predicting being short three nurses. (**b**) Forecasted demand for the day shift "drilled down" to skill. The three forecasted needed nurses are all RNs. (**c**) Proactive Protocol for this situation pulled up (*right of screen*) (color figure online)


c
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go Links
Google Search Web PageRank 1 blocked AutoFill Options
MyAlerts™
Home Request/Self-Schedule Schedule Locator Messages Reports HQS Maintenance Help Logout
AtStaff
Customize This Page Executive Manager Staffing Employee Physician Show Profile Group: [All Profiles] Change Password
12
10
8
6
4
2
0
12
9
1 1
1 1
RN
Target (+2.0)
Sched. (-1.0)
NA
Target (0.0)
Sched. (0.0)
US
Target (0.0)
Target (0.0)
VAR: -3.0 0.0 0.0
Show Detail
Shrink
Edit
Close
Take Action
Done
Local intranet
Start
Sea... Mac... Ado... My...
12:34 PM

MyDashboard - Janet Johnson (Staffing View) - Microsoft Internet Explorer

Address: H:\My Documents\Prototype Files\Dashboard 6.0 03-22-05\HomePages\DB_Staffing01a.htm

MyDashboard

AtStaff®
Intensive Care - AtStaff MC
Logged in as: Janet Johnson

3/23/2005 : NICU down -2.0 staff
1 Employees Competency Exp exceeded in 30 days
You have 2 new messages

Wed., March 23, 2005

Accuracy:
Within 1: 84%
Within 2: 93%

ICU

3 days - 3/26

ICU Protocol for Saturday, March 26, 2005

Short 3 RNs, 0 NA, 0 US

1. Secure 2 RNs to come in.
2. Develop list of 4 others.
3. Consider publishing opportunity.
4. Review situation again tomorrow.

Fig. 11.5 (continued)

11 Conclusions and Future Extensions

The above discussion attempts to place staff management as a critical process for not only optimal throughput of patients in hospitals, but also cost, quality, safety, and staff satisfaction.

11.1 Conclusions

Significant improvements can be made to traditional staff management by moving the staffing decision from the present to the near future. Benefits are improved balance between supply and demand in terms of the number of staff, but perhaps more importantly in the quality, qualifications, motivation, and other less quantitative measures of quality and efficiency. Moreover, by placing the decision out 4 days rather than 4 h, staff satisfaction increases. To move the decision into the near future requires a rather complex mixture of

- Projecting (simulating) current patients and their staff needs into the future
- Predicting (or otherwise obtaining) new admissions

- A decision support system that performs the projections and predictions, displays all options and information to the decision maker, and offers a professional "best practices" framework for those decisions that balance the cost of waiting too late with the costs of acting too early.

Two important side effects of moving staffing decisions from the present to the near future are:

- The establishment of a "mindset" within staff management that things and acts calmly in the near future rather than in chaos in the present, and
- Elevating staff management at the executive level by displaying the effects of staffing in the future to them in the form of executive "dashboards" from the decision support system.

11.2 Extensions

One extension to the work discussed above includes a finer measure of demand for staff using event driven workload, where instead of assuming a patient stays in a "class" or "level" of need for an entire shift, demand is established by "time stamped" events (such as a procedure, return from or, accident, change of condition, etc.) That happens *within* the shift.

An event driven approach "sharpens" all levels of staff management, from determination of core staffing straight through to shift staffing. In addition, retrospectively, it allows a feedback loop to continually refine how well the different parts of staff management perform, as a finer measurement of demand improves decisions at every level.

A second "extension" is represented by improved sources of information about prescheduled admissions and events. These may come through the HL7 interface from the ADT system, for OR schedules, bed control, or other sources. The more information of this type that is available to the simulation, the more accurate the forecasts, and the more aggressive the proactive protocols can be, providing more options to improve the matching of supply to demand, and enhancing employee satisfaction.

References

Kovner, C., & Gergen, P. J. (1998). Nurse staffing levels and adverse events following surgery in U.S. Hospitals. *Image: The Journal of Nursing Scholarship, 30*(4), 315–321.

Litvak, E., Buerhaus, P. I., Davidoff, F., Long, M. C., McManus, M. L., & Berwick, D. M. (2005). Managing unnecessary variability in patient demand to reduce nursing stress and improve patient safety. *Journal on Quality and Patient Safety, 31*(6), 330–338.

Needleman, J., Buerhaus, P., Mattle, S., Stewart, M., & Zelevinsky, K. (2002). Nurse-staffing levels and quality of care in hospitals in the United States. *New England Journal of Medicine, 346*, 1715–1722.

Warner, M. (1976). Scheduling nursing personnel according to nursing preference: A mathematical programming approach. *Operations Research, 24*, 842–856.

Warner, M., Keller, B. J., & Martel, S. H. (1991). Automated nurse scheduling. *Journal of the Society for Health Systems, 2*(2), 66–80.

Chapter 12
Discrete-Event Simulation of Health care Systems

Sheldon H. Jacobson, Shane N. Hall, and James R. Swisher

Abstract Over the past 40 years, health care organizations have faced ever-increasing pressures to deliver quality care while facing rising costs, lower reimbursements, and new regulatory demands. Discrete-event simulation has become a popular and effective decision-making tool for the optimal allocation of scarce health care resources to improve patient flow, while minimizing health care delivery costs and increasing patient satisfaction. The proliferation of increasingly sophisticated discrete-event simulation software packages has resulted in a large number of new application opportunities, including more complex implementations. In addition, combined optimization and simulation tools allow decision-makers to quickly determine optimal system configurations, even for complex integrated facilities. This chapter provides an overview of discrete-event simulation modeling applications to health care clinics and integrated health care systems (e.g., hospitals, outpatient clinics, emergency departments, and pharmacies) over the past 40 years.

Keywords Discrete-event simulation • Health care services • Hospitals • Clinics

1 Introduction

Over the past 40 years, escalating health care costs have provided researchers and health care professionals with the impetus to identify new approaches to improve the efficiency of health care operations and to reduce delivery costs. Discrete-event simulation has become a popular tool for health care decision-makers to support

S.H. Jacobson (✉) • S.N. Hall
Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign, 1206 West Green Street (MC-244), Urbana, IL 61801-2906, USA
e-mail: shj@uiuc.edu

J.R. Swisher
Mary Washington Hospital, 1001 Sam Perry Boulevard, Fredericksburg, VA 22401, USA

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_12, © Springer Science+Business Media New York 2013

their efforts in achieving these objectives. Discrete-event simulation is an operations research modeling and analysis methodology that permits end-users (such as hospital administrators or clinic managers) to evaluate the efficiency of existing health care delivery systems, to ask “what if?” questions, and to design new health care delivery system operations. Discrete-event simulation can also be used as a forecasting tool to assess the potential impact of changes on patient flow, to examine asset allocation needs (such as in staffing levels or in physical capacity), and/or to investigate the complex relationships among different system variables (such as the rate of patient arrivals or the rate of patient service delivery). Such information allows health care administrators and analysts to identify management alternatives that can be used to reconfigure existing health care systems, to improve system performance or design, and/or to plan new systems, without altering the existing system.

The application of discrete-event simulation in the analysis of health care systems has become increasingly more accepted by health care decision-makers as a viable tool for improving operations and reducing costs. This is due in part to the large number of successful discrete-event simulation health care studies reported in the literature, as well as ongoing enhancements to simulation software packages that make their application to health care less arduous. This chapter surveys the large body of discrete-event simulation modeling and analysis efforts that have been reported to address health care delivery system problems and provides an up-to-date, comprehensive collection of articles describing these applications.

This chapter focuses on articles that analyze single or multi-facility health care clinics, including outpatient clinics, emergency departments, surgical centers, orthopedics departments, and pharmacies. An extensive taxonomy of the literature over the past 30 years is presented (though some relevant earlier articles are also referenced). Discrete-event simulation studies on wide area or regional health care community planning, ambulance location service, gurney transportation, disease control planning, and studies that do not address some aspect of patient flow are not discussed. For a complete review of the literature in health care prior to the mid-1970s, see England and Roberts (1978) and Valinsky (1975).

Several excellent review articles have appeared that examine conducting a discrete-event simulation study in health care clinics. England and Roberts (1978) provide a thorough and comprehensive survey on the application of discrete-event simulation in 21 health care settings (including laboratory studies, emergency services, and the national health care system). Their detailed survey cites 92 - discrete-event simulation models out of 1,200 models reviewed, including all published models through 1978. Klein et al. (1993) present a bibliography that includes operational decision making, medical decision making, and system dynamics planning models. Smith-Daniels et al. (1988) present a literature review pertaining to acquisition decisions (e.g., facility location, aggregate capacity, and facility sizing) and allocation decisions (e.g., inpatient admissions scheduling, surgical facility scheduling, and ambulatory care scheduling), including several operations research methodologies, such as heuristics, Markov chains, linear programming, and queuing theory, as well as discrete-event simulation. Jun et al. (1999) present a survey of discrete-event simulation applications to clinic design and analysis. Note that this chapter builds upon the Jun et al. survey and provides new updates that have been reported since 1999.

2 Fundamentals of Discrete-Event Simulation in Health Care

An important advantage of using discrete-event simulation, over other modeling techniques like linear programming or Markov chain analysis, when modeling a health care clinic is the capacity to model complex patient flows through health care clinics, and to play "what if" games by changing the patient flow rules and policies. The success or failure of a discrete-event simulation study within a health care environment often depends on a standard sequence of carefully followed steps. Law and Kelton (2000) outline the key steps necessary to undertake a successful simulation study. These steps include the formulation of the problem and the plan of study, the collection of data and the conceptual model design, the validation of the model, the constructions of the computer representation of the model, the verification of the model, the design of experiments needed to address the problem being studied, production runs using the computer model, the statistical analysis of the data obtained from the production runs, and the interpretation of the results with respect to the system under study. Eldabi and Paul (2001) notes that a key issue in the success of health care simulation studies is the careful formulation of the problem statement, and the buy-in of all stakeholders. In manufacturing simulation studies, modeling and data errors may lead to unexpected costs and poor performance. However, in health care simulation studies, such errors can ultimately lead to lives lost and the associated liabilities surrounding such events. Therefore, the tolerable margin for error in the design and application of health care simulation models is significantly more limited. Such restrictions provide obstacles and barriers that can only be overcome through the highest attention to detail and accuracy, as well as fluid communization between all stakeholders.

The additional references that follow address the fundamental principles for performing a discrete-event simulation study of a health care system. These are excellent references that provide detailed methodologies for a successful simulation study in the context of health care. Hence, for brevity, these methodologies are omitted here. Banks and Carson (1987) and Mahachek (1992) provide structured tutorials on the steps that should be followed when conducting a health care system discrete-event simulation study. Mahachek (1992) also provides details of a discrete-event simulation study on hospital patient flow. Kanon (1974) shows how one could use sample data to build a discrete-event simulation model of a simplified problem in a hospital setting. Eldabi and Paul (2001) discuss an iterative approach to modeling health care systems, while Harper (2002) and Vissers (1998) discuss the general framework and process for operational modeling in heath care. Isken et al. (1999) present a general simulation modeling framework for outpatient obstetrical clinics that is applicable to other types of outpatient clinics. Finally, Morrison and Bird (2003) detail a simulation methodology for improving patient care in ambulatory heath care.

Lowery (1996, 1998) and Standridge (1999) discuss issues facing an analyst when using discrete-event simulation to study a health care system, such as what type of problems are appropriate to be addressed using simulation, the degree of model complexity, the definition of input distributions, model documentation and validation, and the interpretation and reporting of findings. Sanchez et al. (2000) discuss other emerging issues that affect a successful health care discrete-event simulation study, such as the expansion of information technology and the combining of information technology with traditional process models. Baldwin et al. (2004) discuss an iterative approach to health care simulation modeling that is meant to increase the understanding of the problem to decision-makers and enhance communication between stakeholders of the health care system that is being modeled. All of these articles provide useful information for practitioners interested in using discrete-event simulation to study health care systems and issues. Moreover, these articles stress the ability of discrete-event simulation to aptly address the unique factors inherent to health care systems.

Discrete-event simulation models that have been used to analyze health care delivery systems have primarily focused in two areas: (1) optimization and analysis of patient flow and (2) allocation of assets to improve the delivery of services. The first area considers patient flow through hospitals and clinics, with the primary objective of identifying efficiencies that can be realized to improve patient throughput, reduce patient waiting times, and improve medical staff utilization. The second area considers the number of beds and staffing requirements necessary to provide efficient and effective health care services. This section reviews the breadth of studies and approaches taken in these two distinct, though related, areas.

2.1 Patient Flow Optimization and Analysis

As hospitals and clinics face ongoing competition for their services, they must be able to provide fast and efficient health care in order to attract new patients and retain their existing patients. High quality, efficient patient flow is a function of high volume patient throughput, low patient waiting times, short total visit times, and low levels of staff overtime coupled with the maintenance of reasonable staff utilization rates and low physician idle times. Three areas that have a significant impact on patient flow are patient scheduling and admissions, patient routing and flow schemes, and appointment scheduling and availability of resources.

2.1.1 Outpatient Scheduling

Outpatient scheduling focus on procedures for setting the timetable to match patients with caregivers, both in terms of when these appointments are set and their length of time. This involves rules or policies that determine when appointments can be made (such as morning versus afternoon) and the length of time (or spacing) between appointments. This may also include the specific types of caregiver who will be responsible for treating patients and the physical space that will be required to deliver the necessary care and treatment. All these issues have a significant impact on how health care personnel and facility resources can be optimally used such that patient flow is maximized without incurring additional costs or excessive patient waiting.

The majority of discrete-event simulation studies that focus on patient scheduling and admissions are focused on outpatient clinics. Guo et al. (2004) outline a simulation framework for analyzing scheduling rules for outpatient clinics. This framework, termed Patient Scheduling Simulation Model (PSSM), addresses four key components of an outpatient clinic scheduling system: demand for appointments, supply of physician time blocks, patient flow, and the scheduling algorithm. The study provides a demonstration of the framework for a pediatric ophthalmology clinic and discusses some challenges for adapting the framework to other settings. Outpatient clinics typically schedule appointments over some future time horizon. A discrete-event simulation study of rolling horizon appointment scheduling is presented by Rohleder and Klassen (2002). This study considers two common management policies: Overload Rules and Rule Delay. The Overload Rules policy considers scheduling methods such as overtime and double booking that are used when demand is high, while the Rule Delay policy determines when to implement Overload Rules. The authors conclude that determining the “best” scheduling policy depends on the measures of performances that are deemed most important by decision-makers.

Fetter and Thompson (1965) present one of the earliest discrete-event simulation studies conducted in the area of individual clinical facility operations for outpatient clinics. They analyze physician utilization rates with respect to patient waiting time by using different input variables (such as patient load, patient early or late arrival patterns, no-show rates, walk-in rates, appointment scheduling intervals, physician service times, interruptions, and physician lunch and coffee breaks). They concluded that if the physician appointment load increases from 60 to 90 % (capacity), the physician idle time decreases by 160 h and patient waiting time increases by 1,600 h (cumulative over a 50 day period). With such a capacity increase, they suggest that the physician’s time would have to be worth ten times the patient’s time to justify such a shift in patient scheduling and admission policies.

Smoothing the distribution of patient demand has been used to improve patient throughput and patient waiting times in outpatient clinics. Smith and Warner (1971) compare patients arriving according to a uniformly scheduled arrival pattern versus patients arriving in a highly variable manner. They show that the uniformly scheduled arrival pattern can decrease the average length of stay at the clinic by over 40 % (from 40.6 to 24 min), due to the more predictable use of resources when patient arrivals are uniformly spaced. Similarly, Rising et al. (1973) show that increasing the number of appointments slots in an outpatient clinic on those days that had the least number of walk-ins smoothed the demand on the physician, resulting in a 13.4 % increase in patient throughput and less clinic overtime. Kho and Johnson (1976) and Kachhal et al. (1981) show that in a radiology department and in an ear, nose, and throat clinic, respectively, performance can be improved when demand for outpatient services is evenly distributed.

In contrast to uniform scheduling, a number of alternative scheduling rules have been studied. Bailey (1952) studies an outpatient clinic scheduling rule (where two patients are scheduled at the beginning of every session, morning or afternoon, with all other patients scheduled at equal intervals) that yields acceptable results for both patients (in terms of waiting times) and staff (in terms of utilization), assuming that all patients have the same service time distributions and that all patients arrive at their designated times. Smith et al. (1979) use a modified-wave scheduling scheme (where more patients are scheduled at the beginning of each hour and less towards the end of the hour, thus allowing the physician to absorb unexpected delays and return back to schedule at the end of each hour) for an outpatient clinic to find the maximum number of patients a physician could see while minimizing patient waiting time. They show that this schedule is superior to the uniform scheduling scheme, in terms of patient flow and patient waiting times. Williams et al. (1967) study the relationship between physician utilization and patient waiting time in an outpatient clinic using a staggered block scheduling system (i.e., 8 patients arriving every half hour) versus the single block scheduling system (i.e., 16 patients arriving simultaneously). The single block system emphasizes the physician's idle time, while the staggered block system emphasizes the patient's waiting time, resulting in a substantial decrease in the patient waiting times (with no decrement in the utilization of the physician).

2.1.2 Inpatient Scheduling and Admissions

Inpatient scheduling and admissions focus on procedures for matching patients with caregivers (e.g., surgeons, infectious disease specialists) within a hospital or similar health care facility. This involves rules or policies that determine when health care providers must see and provide services for inpatients, as well as matching specific types of caregiver with the needs of patients and the necessary treatments.

Surgical (operating room) center scheduling has also been studied using discrete-event simulation. Magerlein and Martin (1976) review the literature on using discrete-event simulation for scheduling surgical centers. Murphy and Sigal (1985) examine surgical block scheduling, where a block of operating room time is reserved for one or more surgeons. Fitzpatrick et al. (1993) study the use of first-come-first-serve, fixed (scheduling the same block of time in the same time slot each day of the week), variable (scheduling under the influence of seasonal fluctuation in demand), and mixed block scheduling (a combination of fixed and variable) for hospital operating rooms. They observed that variable block scheduling is superior to all scheduling policies, in terms of facility utilization, patient throughput, average patient waiting time, and patient queue length.

Klassen and Rohleder (1996) use discrete-event simulation to study the best time to schedule patients with large patient service time means and variances. They analyze several rules and arrive at the best result (which is to schedule such patients towards the end of the appointment session) that minimizes the patient's waiting time and the physician's idle time. Additionally, they analyze the best position for unscheduled appointment slots for potentially urgent calls and found no conclusive scheduling rule. Swisher et al. (1997) consider scheduling more patients with larger mean service time distributions in the morning session, rather than the afternoon, in an outpatient clinic. They found that staff overtime is sharply reduced, with a corresponding reduction in the physician's lunch time period. Steward and Standridge (1996) report the results of an overtime study using a discrete-event simulation model of a veterinary practice. The veterinary domain is very similar to the larger human medical systems domain, since both involve the challenge of varying flow and service rates, as well as resource utilization, staffing, demand, and scheduling. In this study, the performance measure of interest is the average time interval between the closing time of the clinic and the time the last client is discharged after clinic hours, which serves as an indicator of overhead cost and client satisfaction. They report that performance can improve if the clinic disallows the scheduling of appointments less than 90 or 120 min prior to closing, rather than 60 min, as was the current practice.

Hancock and Walter (1979) attempt to use discrete-event simulation to reduce the variance in occupancy levels in a hospital inpatient facility, with the goal of increasing patient throughput and maximizing average occupancies. Unfortunately, they were unsuccessful in achieving their stated objective, since the staff was accustomed to admitting patients on the date of the requests 90 % of the time, and they refused to schedule over 4 weeks in advance. Hancock and Walter (1984) attempt to smooth the daily patient loads of 19 hospital departments by varying the admission days of urgent inpatient and outpatient loads. The variation in average load for each of the departments suggested that no one single admission policy could provide a stable workload for all the departments, since each department had its own unique patient arrival patterns and treatment requirements, including different inpatient and outpatient requirements.

Lim et al. (1975) apply two admission policies (quickcall and maximum queue lengths) to a discrete-event simulation model of an inpatient orthopedics unit. "Quickcall" is defined as a patient willing to enter the hospital on very short notice; whereas, maximum queue lengths is a concept in which the physicians are required to maintain a maximum number of patient requests on a waiting list. Both systems improved system performance, in terms of patient waiting times and staff utilization. Similarly, Groothuis et al. (2001) investigate two patient scheduling procedures (the current procedure where no patient is scheduled after 4:00 PM, versus scheduling a fixed number of patients each day) for a hospital cardiac catheterization lab. Both scheduling procedures were applied to the current configuration and three additional experimental configurations, with patient throughput and working day duration as the measures of performance. A discrete-event simulation was designed using Medmodel and showed that the third experimental configuration under the current scheduling procedure could, on average, accommodate two additional patients with fewer working days that exceed 8 h.

Walter (1973) describes several aspects of a queuing system in a radiology department, using several different appointment schemes. By segregating patients into inpatient and outpatient sessions with a similar examination time distribution, Walter observed that a substantial staff time savings was possible. He also found that the practice of giving multiple bookings for a given appointment time (i.e., overbooking) yields a small increase in staff utilization while substantially increasing the patient waiting time, and that efficiency always improves when the proportion of patients with appointments increases, resulting in a smoothing of the arrival rate. Goitein (1990) obtained similar conclusions using Monte Carlo simulation to examine factors such as physician idle time relative to patient waiting time. He found that if the physician overbooked the schedule (even slightly), patients would experience very long waiting times. His model provides insights into how delays build up as a result of commonly observed statistical fluctuations. Everett (2002) suggests using a simulation model to help match patient needs with hospital availability (in a public hospital system) by scheduling patients waiting for elective surgery.

In conclusion, patient scheduling and admission rules along with patient appointment timing can have significant impacts on physician utilization and patient waiting. In general, studies using discrete-event simulation as discussed here suggest that rules and policies can be employed that will help to balance the trade-off between physician utilization rates and patient waiting times, though the unique features of each clinic environment need to be taken into account to determine the exact extent of these trade-offs. External market factors often dictate how health care facilities must prioritize the trade-off between patient convenience and caregiver utilization. For example, in highly competitive markets, clinics may favor patient convenience over staff utilization in an effort to retain market share. Obviously, the unique factors that determine optimality must be elicited from each decision-maker given his/her environmental factors. The studies presented herein also point to the importance of, when possible, smoothing patient arrival rates and service times. As in most systems, reducing variability facilitates performance improvement.

2.1.3 Emergency Room Simulation Models

Discrete-event simulation models can capture complex patient flows through health care clinics, as well as analyze the effect of new patient flow rules and policies. Such flows are typical in emergency room settings, where patients arrive (nearly always without appointments), and require treatment over a large and varied set of ailments and conditions, ranging from the benign (e.g., mild sports injuries) to the fatal (e.g., heart attacks, gunshot wounds). Although the patient arrival patterns are highly unpredictable, the treatment sequence can be controlled by clinical staff. Therefore, by altering patient routing and flow, it may be possible to minimize patient waiting times and increase staff utilization rates.

Limited access to primary care has led to extreme increases in emergency department usage across the USA. Emergency department overcrowding has been recognized by national health industry groups and regulatory bodies like the American Hospital Association (AHA) and the Joint Commission on the Accreditation of Health care Organizations (JCAHO) as a significant public health issue. All of this has led to a significant increase in the use of discrete-event simulation in modeling emergency departments in the past decade. General guidelines for analyzing an emergency department using discrete-event simulation exist in the literature. Takakuwa and Shiozaki (2004) propose a procedure for planning emergency room operations that minimize patient waiting times. Sinreich and Marmor (2004) develop a general emergency department simulation tool that is "flexible, intuitive, simple to use and contains default values for most of the system's parameters." Miller et al. (2004) describe steps for building a discrete-event simulation tool meant to determine the best emergency room configuration.

A key service metric used by hospital emergency departments is patient waiting time. Garcia et al. (1995) analyze the impact of a fast track queue on reducing waiting times of low priority patients in an emergency room. Emergency room patients are typically prioritized according to patient acuity (the level of sickness), and hence low-acuity patients regularly wait for excessively long periods of time. A fast track queue is used to treat a particular patient acuity level (in this case, nonurgent patients). They found that a fast track lane that uses a minimal amount of resources could result in significantly reduced patient waiting times. A similar study to assess the effect of fast care processing routes for noncritical patients on waiting times in an emergency department is presented by Mahapatra et al. (2003). This study showed that the addition of an alternate care unit (such as a fast track unit) improved average waiting times by at least 10 %. In a discrete-event simulation model of the emergency department at the University of Louisville Hospital, Kraitsik and Bossmeyer (1993) suggest that patient throughput can be improved using a fast track queue and a "stat" lab for processing high volume tests. Kirtland et al. (1995) examine 11 alternatives to improve patient flow in an emergency department and identified 3 alternatives (using a fast track lane in minor care, placing patients in the treatment area instead of sending them back to the waiting room, and the use of point-of-care lab testing) that can save on average 38 min of waiting time per patient. Blake et al. (1996) also analyze an emergency department at the Children's Hospital of Eastern Ontario using discrete-event simulation. Their study led to the implementation of a fast track queue for treating patients with minor injuries.

Another important measure for emergency department efficiency is the overall time a patient spends in the emergency room (i.e., the patient length of stay). McGuire (1994) uses MedModel to determine how to reduce the length of stay for patients in an emergency service department in a SunHealth Alliance hospital. The results from the study resulted in several alternative recommendations, including adding an additional clerk during peak hours, adding a holding area for waiting patients, extending the hours of the fast track queue, and using physicians instead of residents in the fast track area. Miller et al. (2003) use a discrete-event simulation of an emergency department of a large hospital in the southeast USA to show that significant process changes would be required to meet specified goals for patient length of stay. Samaha et al. (2003) describe how discrete-event simulation was used by the Cooper University Hospital to reduce patient length of stays in their emergency department. Their study determined that length of stay was a process related problem rather than resource dependent. For example, the study showed that adding square footage or beds would not shorten the length of stay, which resulted in significant cost avoidance. Another discrete-event simulation study of patient flow to reduce emergency department length of stay is presented by Blasak et al. (2003). This study also simulates an inpatient medical telemetry unit to see how the processes of other units impact the emergency department. El-Darzi et al. (1998) and Martin et al. (2003) present additional patient flow simulation studies that seek to increase patient throughput and decrease patient length of stay. Both studies, however, model a hospital geriatric department.

Ritondo and Freedman (1993) show that changing a procedural policy (of ordering tests while in triage) results in a decrease in patient waiting times in the emergency room and an increase in patient throughput. Edwards et al. (1994) compare the results of simulation studies in two medical clinics that use different queuing systems: serial processing, where patients wait in a single queue, and quasi-parallel processing, where patients are directed to the shortest queue to maintain flow. They show that patient waiting times could be reduced by up to 30 % using quasi-parallel processing. Johnson (1998) uses a MedModel discrete-event simulation model to examine the effect of new legislation (requiring a minimum length of stay) and physician practices on patient flow and census of the maternity unit at Miami Valley Hospital in Dayton, Ohio, USA. The study led to minor changes in the maternity unit configuration that resulted in a 15–20 % increase in patient volume and more balanced utilization of all areas within the unit. Also, the model results supported decisions to construct new facilities, such as a larger perinatal intensive care unit.

2.1.4 Specialist Clinics

Specialists bring their own unique set of issues when scheduling patients and allocating space within health care facilities, Sepúlveda et al. (1999) use discrete-event simulation to evaluate improvement in patient flow at a cancer treatment center under three different scenarios: (1) a change in the layout of the clinic, (2) different patient scheduling options, and (3) a new facility with increased capacity. The simulation of all three scenarios identified key patient flow bottlenecks and provided insights to improve patient flow and utilization. In particular, under the layout scenario, the simulation was used to identify a facility layout that allowed for a 100 % increase in chair capacity. Simulating different patient scheduling options showed a 20 % increase in the number of patients seen per day, without any change in the operating time of the treatment center. Finally, the new facility scenario showed that one of the waiting rooms did not have the capacity to support patient flow.

Ramakrishnan et al. (2004) describe a discrete-event simulation model used to analyze different “what-if” scenarios for the Wilson Memorial Regional Medical Center in Broome County, New York USA. The center recently implemented a digital image archiving system within its radiology services department and with this implementation wanted to identify patient flow changes in the computerized tomography (CT) scan area that would maximize patient throughput and minimize report generation time. Using simulation, the researchers identified changes within the CT scan area that would increase patient throughput by 20 %, while simultaneously reducing report generation time by over 30 %. Likewise, Alexopoulos et al. (2001) describe a discrete-event simulation used by Partnership of Immunization Providers to study “what-if” scenarios for immunization clinics serving the poor. Such scenarios included narrowing/expanding appointment slots and the impact of bilingual versus monolingual staff on patient throughput.

Groothuis et al. (2002) describe a systematic approach for analyzing the effects on patient flow when a hospital department is relocated. This approach is demonstrated with a MedModel simulation model of relocating a hospital phlebotomy department, which assesses the resulting impact on the average patient turnaround time. They observed that this time could be reduced by as much as 50 % (from 12 min down to 8 min).

2.1.5 Physician and Health Care Staff Scheduling

The majority of discrete-event simulation models for scheduling health care clinics are directed at patient scheduling (so as to optimally distribute patient demand to physicians and clinical staff). A number of studies, however, have addressed the reverse problem; namely, scheduling physician and clinical staff to satisfy patient demand, given a collection of patient arrivals. For example, walk-in clinics, which are unable to control the arrival rate of patients, must schedule their staff accordingly. Incorporating this idea, Alessandra et al. (1978) study both the staffing levels and patient arrival rates to ease bottlenecks and to improve patient throughput. Eight alternatives that varied the staffing pattern and the patient scheduling scheme were analyzed. The best alternative identified was to keep the staffing and arrival rate the same, but to distribute the current morning appointment patients to the afternoon shift. Mukherjee (1991) identifies a staffing mix that reduces patient waiting time and increases patient throughput, while controlling resource costs in a pharmacy.

There have also been discrete-event simulation studies that address physician scheduling. Rossetti et al. (1999) use discrete-event simulation to test alternative physician-staffing schedules at the emergency department at the University of Virginia Medical Center. For each staffing alternative, they analyzed the impact on patient throughput and resource utilization. Tan et al. (2002) present a discrete-event simulation study of an urgent care center that simulates the current physician schedule and a proposed schedule to test if the proposed schedule reduces the average total time patients spend at the facility. The simulation showed an 18 % reduction in total visit time using the proposed schedule. Likewise, using a discrete-event simulation, Lach and Vázquez (2004) study a telemedicine program in Mexico that provides medical assistance to those living in extreme poverty. This study analyzes the effect on patient throughput and resource (tele-consult) utilization when an extra physician is scheduled; both tele-consult utilization and patient throughput increase when the extra physician is scheduled. Osidach and Fu (2003) study the staffing of technicians required to perform medical exams on scheduled survey participants in a mobile examination center. Configurations of three, four, and five technicians for a batch arrival of seven survey participants were simulated to minimize the technicians' idle time and average time spent in the system by the survey participant. The best configuration for utilization was a three technician configuration, whereas a five technician configuration was the most time-effective; however, the five technician configuration also resulted in an overly crowded examination center.

Several discrete-event simulation models of nursing staff scheduling in emergency departments have been developed. Emergency room staff scheduling has its own unique challenges, due to the high volume of visits, significant variability in patient arrival patterns, and the urgency of the care required. Draeger (1992) studies nurse workload in an emergency room and its effect on the average number of patients, average time in system, average number of patients waiting, and average patient waiting time. Comparing the current schedule's performance to those of two alternative staffing schedules, the author found an alternative that could reduce both the average patient time in system (by 23 %) and the average patient waiting time (by 57 %), without any increase in costs. Similarly, Evans et al. (1996) reduce a patient's length of stay by finding the optimal number of nurses and technicians that should be on duty during four shifts in an emergency room. Kumar and Kapur (1989) examine ten nurse scheduling policy alternatives, selecting and implementing the policy yielding the highest nurse utilization rate.

Lambo (1983) applies a recursive linear programming and discrete-event simulation methodology to examine staffing problems in a health care center in Nigeria. In the study, the clinic was observed to be at 50 % capacity due to the misallocation of (rather than the inadequacy of) personnel. After making changes to the staffing patterns and other policy changes, capacity increased by 60 % and patient waiting times were reduced by 45 min. In a similar study, Chan et al. (2002) uses integer programming and discrete-event simulation to study a medical records department to determine the optimal staff schedule and understand the workflow of the transcription function.

All of these discrete-event simulation studies suggest that when patient flow patterns cannot be controlled, staffing strategies can be employed to smooth some of the unavoidable variability in the systems. This can result in improved patient throughput, while keeping staff utilization rates and total staffing costs at acceptable levels. It may also act as an important public relations and marketing tool. Health care is unique in that quality is not always readily identifiable by its customers. A patient whose condition improves may have received quality care; however, if services were delivered inefficiently, the patient's perception of quality may be greatly diminished. Efficient patient flow, then, often acts as a surrogate for quality of care in the patient's mind. Given a choice, patients will tend to a health care provider they perceive to offer higher quality services. Moreover, facilities that minimize the obstacles to the provision of care for health care providers (e.g., physicians, nurses) are better able to attract and retain the best and brightest. In short, patient flow is not just important to the bottom line, but it can serve as a major competitive advantage.

2.2 *Health Care Asset Allocation*

Hospital and clinic administrators have approached cost containment within their operations by working to minimize expenditures for health care provisions while simultaneously providing quality health care services. Such situations pervade the health care community as indicated by the large number of papers and studies that analyze the allocation of scarce health care assets. Discrete-event simulation modeling is an attractive method to help make such allocations since it can be used to estimate the operational characteristics of a health care system operation and to observe the impact of changes in planning or policies prior to the implementation of such changes, and thereby mitigate financial risks. The allocation of health care assets can be broken down into three general areas where assets most directly impact health care delivery: bed sizing and planning, room sizing and planning, staff sizing and planning.

2.2.1 Bed Sizing and Planning

The demand for hospital beds can be classified as either routine (e.g., scheduled) or emergency (e.g., unscheduled) admissions. Both of these admission types impact how many beds are needed to meet demand, while maintaining reasonable bed utilization rates. In the literature, most bed planning discrete-event simulation models attempt to overcome bed shortages or policies that lead to patient misplacement, bumping, or rejection. Hospitals are typically faced with the trade-off between having available beds to service patient demand versus keeping bed occupancy (utilization) rates high.

Butler et al. (1992c) use discrete-event simulation to study patient misplacements, where patients are scheduled and assigned to an alternative unit within a hospital due to a shortage of beds in the preferred hospital area. They examine the sensitivity of patient misplacement with respect to a variety of modifications in their bed allocation policy, including patient transfers, bed scheduling, and assignments, and found that reducing a patient's length of stay and reallocating rooms among the different services within a hospital could substantially decrease patient misplacement. Furthermore, the smoothing of routine patient arrivals only marginally reduced patient misplacement. In another study designed to reduce patient misplacement, Butler et al. (1992a) use a two-phase approach involving a quadratic integer programming model and a discrete-event simulation model to evaluate bed configurations and to determine optimal bed allocations across a number of hospital service areas. Vasilakis and El-Darzi (2001) use discrete-event simulation to identify the possible cause for a hospital bed crisis that occurs each winter in UK National Health Services hospitals. Using simulation, they demonstrate a "breakdown in the discharge of dependent patients from the medium stay (or rehabilitative) stream" because during the Christmas and New Year holiday season surgeons are not performing routine operations. However, once surgeons return to work after the holidays, the resulting surge in the number of surgeries scheduled results in an insufficient number of beds for incoming medical emergencies. The study suggests that the UK winter bed crisis is likely due to staff scheduling, the holidays, and ineffective management of non-acute (longer stay) patients. Note that US hospitals typically experience such "winter census" crises, as well.

Lowery (1992, 1993) and Lowery and Martin (1992) consider the use of discrete-event simulation in a hospital's critical care areas (e.g., operating rooms, recovery units, intensive care units, and intermediate care units) to determine critical care bed requirements. Their literature review reveals that most models do not fully consider the interrelationships between different hospital units and few models have been validated using actual hospital performance data. Focusing on these deficiencies, they demonstrate improvements in their methodologies over previous models. Dumas (1984, 1985) also focuses on the interrelationships between several units within a hospital by comparing two bed planning rules (vacancy basing and home basing) for locating a bed within different hospital units when a patient cannot be allocated a bed at the preferred unit. Vacancy basing rules employ a ranked list of alternative misplacement possibilities, while home basing prohibits off-service misplacements, and hence, is more restrictive with respect to patient placement. They show that home basing policies result in better overall performance but lessen patient days and thereby reduce hospital revenues. Note that in the mid-1980s (the time of Dumas' publications) most hospitals were still paid by third party providers based on the patient's overall length of stay, so reducing patient days was seen as a potentially negative outcome. In contrast, modern reimbursement systems tend to favor case rates that encourage shorter lengths of stay, ultimately resulting in an incentive for hospitals to reduce patient days.

Cohen et al. (1980) present a bed planning model of a progressive patient care hospital, where patients are moved between units within a hospital as their condition changes. In this form of demand-matching, hospitals attempt to apply resources commensurate with patients' condition during their stay by "stepping the patient down" as their conditions improve. The authors demonstrate that the probability of inappropriate patient placement is a function of the capacities of all the units, as well as the policies for handling priority patients and bumped patients.

By considering individual units within a hospital, Zilm et al. (1983) use a discrete-event simulation model to analyze a surgical intensive care unit for various bed levels and future demand. They observe that most of the unit's volume consists of weekday cases (routine admissions), and hence, attempts to maintain a high overall average occupancy level would not be possible without straining the entire system. Similarly, Cahill and Render (1999) study proposed changes to the intensive care unit (ICU) at the Cincinnati Veterans Administration Medical Center. Using discrete-event simulation, they show that creating a respiratory care unit and increasing bed levels in other units closely associated with the intensive care unit would resolve the ICU access problems. However, an increase in ICU bed availability increased bed utilization in other units, which in turn increased the overall length of stay. Therefore, modeling in advance helped the hospital identify policy changes to lessen the impact on length of stay. Masterson et al. (2004) discuss the optimization of the military health system for all military health facilities. They present a case study based on a simulation analysis of the intensive care unit at the US Air Force's Wilford Hall Medical Center, to determine the appropriate ICU size, bed mix, and staffing level.

Romanin-Jacur and Facchin (1987) use discrete-event simulation to study the facility dimensioning problem and the sizing of the assistance team in a pediatric semi-intensive care unit. They compare several different priority-based models by using peak admission conditions to find the optimal number of beds and the best choice of the nurse's care assignment. Other bed sizing discrete-event simulation models can be found in Hancock et al. (1978), Wright (1987), Harris (1985), Wiinamaki and Dronzek (2003), and Akkerman and Knip (2004). Harris (1985) compares the difference in the number of surgical suites needed in a surgical center for three physicians under two operating timetable scenarios. Under the first (and current) scenario, each physician scheduled his/her patients independently of the other two physicians, while in the second scenario, the physicians pooled their resources to schedule their patients and consequently reduced the number of surgical suites required by over 20 %. Wiinamaki and Dronzek (2003) show how simulation was used in determining the bed requirements for the new emergency care center at the Sarasota Memorial Hospital in Sarasota, Florida. Akkerman and Knip (2004) show that the number of beds could be reduced in a cardiac surgery center if recovering patients are transferred once they no longer require the center's specialized care services.

Gabaeff and Lennon (1991) use an extensive time-motion study to collect data on the mix of patient types, patient characteristics (such as x-ray requirements), and staffing mix for emergency admissions in an emergency department feasibility study at Stanford University Hospital. Using discrete-event simulation models, they highlight deficiencies in several key areas, including maximum bed utilization exceeding current bed availability (which would cause displacement of minor care patients). Vassilacopoulos (1985) develops a discrete-event simulation model to determine the number of beds with the following constraints: high occupancy rates, immediate (emergency admission) patients, and low length of waiting lists. He shows that by using a waiting list and smoothing the patient demand, it is possible to achieve high occupancy rates. Emergency department bed planning discrete-event simulation models are also discussed by Altinel and Ulas (1996) at the Istanbul University School of Medicine, Freedman (1994) at St. Joseph Hospital and Washington Adventist Hospital in Maryland, USA, Lennon (1992) at the Stanford University Hospital, and Williams (1983) at the University of Pennsylvania Hospital. All these studies suggest that discrete-event simulation modeling and analysis provides a valuable "what if" tool for hospital planners when deciding how many beds are needed to meet demand and maintain profitability. It also assists decision-makers in judiciously allocating precious financial resources. The ever-increasing costs of medical equipment (e.g., CT scanners) mean that health care administrators must preserve capital for technology that historically could have been allocated to brick-and-mortar expansions. Simulation modeling can play an important role in this effort. Moreover, simulation models allow hospital administrators to experiment with different bed allocation rules to help optimally utilize hospital facilities and improve bed occupancy rates.

2.2.2 Room Sizing and Planning

The ongoing movement towards freestanding surgicenters, as well as the shift to deliver health care services away from inpatients facilities and towards outpatient facilities, has put increased pressure upon hospital management to expand their outpatient services and/or to build new facilities to handle these additional patient demands. Discrete-event simulation has become an important tool for the planning of future expansion, integration, and/or construction of new outpatient facilities and health service departments, by significantly enhancing the hospital administration decision-maker's ability to find the most cost-effective and efficient solutions to such problems.

The number and use of operating rooms is often an important resource in maintaining hospital profitability and patient services. Currie et al. (1984) study operating room utilization, vertical transportation needs, radiology staffing, and emergency medical system operations at the West Virginia University Hospital. They use discrete-event simulation to estimate the number of operating rooms and recovery beds needed to handle a 20 % increase in future demand. Kwak et al. (1975) use discrete-event simulation to determine the capacity of a recovery room needed to support an operating room expansion. Similarly, Kuzdrall et al. (1981) use a discrete-event simulation model of an operating and recovery room facility to determine and assess the facility utilization levels and facility needs under different scheduling policies. Olson and Dux (1994) apply discrete-event simulation modeling to study and evaluate the decision to expand the Waukesha Memorial surgicenter from seven to eight operating rooms. Their study reveals that an eighth operating room would only serve to meet the hospital's needs for no more than 2 years, at a cost of $500,000 (USD). However, an analysis of the cross-departmental and administrative needs reveal that an ambulatory surgery center that separates the inpatients and outpatient procedures would better serve the hospital's future health care delivery needs. Similarly, Amladi (1984) uses discrete-event simulation to help size and plan a new outpatient surgical facility, by considering patient wait time (quality) and facility size (resource). Lowery and Davis (1999) developed a discrete-event simulation to assess the impact of a proposed renovation to the surgical suite of Brigham and Women's Hospital in Boston, Massachusetts, USA. Hospital administrators wanted to ensure that the renovations would be sufficient to handle a projected increase in the number of inpatient surgeries. The simulation showed the projected increases could be met with 30 or less operating rooms (32 operating rooms were planned in the renovation) provided scheduled block times were extended to include the addition of a Saturday block.

Ferrin et al. (2004) apply discrete-event simulation to help St. Vincent's Hospital in Birmingham, Alabama, USA, a not-for-profit hospital, determine the value of implementing an incentive program for their operating room environment. The simulation showed that improving the room turnaround process by 20 % would result in a 4 % improvement in the operating rooms case volume and a 5 % increase in utilization of same day surgical rooms. This increase in volume provided enough increase in revenue to justify an incentive to improve the operating room turnaround process. In addition to evaluating the value of incentives, St Vincent's administration used the simulation to determine the required number of operating rooms, number of beds in the Post Anesthesia Care Unit (PACU), and changes in physician scheduling blocks.

Meier et al. (1985) use discrete-event simulation to compare and evaluate 11 scenarios in varying the number of exam rooms and demand shifts of both a hospital ambulatory center and a freestanding surgicenter. They found that existing room capacity is adequate to handle demands over the next 5 years. Iskander and Carter (1991) use discrete-event simulation to show that current facilities were sufficient for future growth in a study of a same day (outpatient) health care unit in an ambulatory care center. However, they suggest a threefold increase in the size of the waiting room. Using discrete-event simulation, Ramis et al. (2001) evaluate a proposed future center for ambulatory surgery by evaluating several process alternatives. The results determined the bed resources and scheduling rule required to maximize daily surgical throughput. Similarly, Stahl et al. (2003) seek to optimize the management and financial performance of ambulatory care clinics used for teaching medical students. Here they use discrete-event simulation to determine that a teaching ambulatory care clinic runs optimally (where optimality is defined as the policy that minimizes patient flow time and wait time while maximizing revenue) when the trainee-to-preceptor ratio is between 3 and 7 to 1.

Kletke and Dooley (1984) use discrete-event simulation to examine the effects on service level and utilization rates in a maternity unit to determine if the current number of labor rooms, delivery rooms, post-partum rooms, nursery, and nurses are able to meet future demands. Their study recommends increasing the number of labor rooms and the number of postpartum rooms, while maintaining four full-time nurses. Levy et al. (1989) use discrete-event simulation to analyze the operational characteristics of an outpatient service center at Anderson Memorial Hospital to determine whether to merge this service with an off-site outpatient diagnostic center. They collected data on the utilization of the servers, the total number of patients in the center, the maximum and average times spent in the center, the maximum and average times spent in each service queue, and the total number of patients in each queue. This information was used to specify staffing and facility sizing requirements. In another facility integration plan, Mahachek and Knabe (1984) use discrete-event simulation to evaluate a proposal to cut costs by combining an obstetrics clinic and a gynecology clinic into a single facility. The analysis found that this proposal would not be successful due to the shortage of exam rooms. All these studies illustrate the value provided by discrete-event simulation modeling and analysis to determine how to set the size of key hospital facilities (such as operating rooms). As the health care industry continues to move more towards outpatient delivery systems, and away from traditional inpatient health care facilities, discrete-event simulation will continue to assist health care decision-makers in leveraging their resources in undertaking such transitions.

2.2.3 Staff Sizing and Planning

The medical community requires highly skilled staff to deliver quality health care services, making staff sizing and planning an important factor in designing health care delivery systems (such as those found in hospitals). Moreover, the trade-off between insufficient clinical staff to meet demand (hence unacceptable patient waiting times) and underutilization of clinical staff can have an enormous impact on the economic viability and sustainability of a medical facility. Discrete-event simulation has played an important role in addressing the issues inherent in this trade-off.

Several discrete-event simulation studies have been conducted to determine the staff size or the number of physicians for emergency departments (e.g., Carter et al. 1992). Badri and Hollingsworth (1993) analyze the impact of different operational scenarios on scheduling, limited staffing, and changing the patient demand patterns in an emergency room of the Rashid Hospital in the United Arab Emirates. These scenarios included using a patient priority rule based on severity of ailment, not serving a category of patient that does not belong in the emergency room, eliminating one or more doctors on each shift, and a hybrid scenario that combines the last two scenarios. The results from this hybrid scenario were accepted and implemented. Klafehn and Owens (1987) and Klafehn et al. (1989) address the problem of determining the relationship between patient flow and the number of staff available in an emergency department. They conclude that moving one nurse from the regular emergency area to a triage position significantly reduces patient waiting lines and waiting times. Furthermore, they found that the addition of a second orthopedic team in the emergency department increases patient throughput, though utilization levels were lower and the average length of stay remains virtually the same (since the number of patients flowing through the orthopedic area was relatively small). Liyanage and Gale (1995) formulate an M/M/n queuing model of the Campbelltown Hospital emergency facility to estimate and develop patient arrival time distributions, patient waiting times, and patient service times. These parameters were then used in a discrete-event simulation model to estimate the expected patient waiting times, the expected physician idle times, and the optimal number of doctors. Baesler et al. (2003) use a discrete-event simulation model to predict a patient's time spent in the emergency room of a private hospital in Chile. These results were then used in a design of experiments to minimize the number of resources (four fulltime physicians and one part-time physician) required to meet patient demand. Similarly, Centeno et al. (2003) combine discrete-event simulation with integer programming to develop an optimal schedule for emergency room staff. The study showed a 28 % improvement over the current method of staffing which offers a potential significant savings in hospital labor costs. Lopez-Valcarcel and Perez (1994) use discrete-event simulation to evaluate the staffing levels, the arrival rates, and the service times of diagnostic equipment (alterable by purchasing better equipment) in an emergency department. They recommend that the arrival rate should not exceed 12 patients per hour. Moreover, they recommend that investments in human resources would be more effective than investments in newer (better) equipment. In contrast, Bodtker et al. (1992) and Godolphin et al. (1992) determine that a reduction in staff by at least one staff member could be achieved if better equipment were purchased.

O'Kane (1981), Klafehn (1987), and Coffin et al. (1993) use discrete-event simulation to analyze staff allocations to improve patient flow in a radiology department. Klafehn and Connolly (1993) model an outpatient hematology laboratory and compare several configurations. They observed that if the staff is cross-trained (and hence, can be more fully utilized), then patient waiting times can be reduced. Vemuri (1984) and Ishimoto et al. (1990) use discrete-event simulation to identify the optimal medical staff size and mix that reduces patient waiting times in a hospital pharmacy. Weng and Houshmand (1999) simulate a general hospital outpatient clinic to find the optimal staff size that maximizes patient throughput and minimizes patient flow time.

Jackson Memorial Hospital (JMH) in Dade County, Florida, USA uses discrete-event simulation to model hospital operations. Centeno et al. (2001) present a simulation study of the labor and delivery rooms at JMH. This study used historical data for all simulation inputs and identified ways to improve physician scheduling and better staffing levels. A simulation study of the radiology department at JMH is discussed in Centeno et al. (2000). Six different scenarios, that vary staff and physical resources, are studied to determine the impact on patient flow and utilization of the department staff and operating rooms. This study determined the most cost-effective staff level for each procedure and identified additional revisions to improve process and service efficiencies.

Hashimoto and Bell (1996) conduct a time-motion study to collect data for a discrete-event simulation model of an outpatient (general practice) clinic. They show that increasing the number of physicians, and consequently the number of patients, without increasing the support staff, would significantly increase patient length of stay. By limiting the number of physicians to four and increasing the number of dischargers to two, they were able decrease the patient's average total time in the system by almost 25 %. Wilt and Goddin (1989) use discrete-event simulation to evaluate patient waiting times to determine appropriate staffing levels in an outpatient clinic. McHugh (1989) uses discrete-event simulation to examine hospital nurse-staffing levels and their impact on cost and utilization. This analysis shows that 55 % of the maximum workload produces the best results based on these measures. Swisher et al. (2001) discuss a discrete-event simulation model of the Queston Physician Practice Network where individual family outpatient clinics are modeled and integrated into a network of clinics that uses a central appointment scheduling center. Performance measures such as patient throughput, patient waiting time, staff utilization, and clinic overtime are analyzed for various numbers of exam rooms and staff mixes. In certain cases, adding support personnel had negligible effects on the performance measures. Swisher and Jacobson (2002) use an object-oriented visual discrete-event simulation to evaluate different staffing options and facility sizes for a two physician family practice health care clinic. They describe a clinic effectiveness measure that is used to evaluate the overall effectiveness of a given clinic configuration. This clinic effectiveness measure integrates clinic profits, patient satisfaction, and medical staff satisfaction into a single performance measure.

Rossetti et al. (1998) use discrete-event simulation to study clinical laboratory and pharmacy delivery processes in a mid-size hospital environment. The study specifically assesses the costs and performance benefit of procuring a fleet of mobile robots to perform delivery functions. The study found that a fleet of six mobile robots improved the turn-around time by 33 % and reduced costs by 56 % compared to the current system of three human couriers. Similarly, Wong et al. (2003) use simulation "to quantify the advantages of an electronic medication ordering, dispensing, and administration process" at an academic acute care center. The automated system had an average turnaround time of 123 min versus the current manual system turnaround time of 256 min. Dean et al. (1999) also study a hospital pharmacy distribution system, where they use simulation to help determine when a pharmacist should visit each nursing unit to minimize the mean time delay between the time when a prescription is filled and its arrival to the ward.

Stafford (1976) and Aggarwal and Stafford (1976) develop a multi-facility discrete-event simulation model of a university health center that incorporates 14 separate stations (e.g., receptionist area, injections, dentist, gynecology, physical therapy, radiology, and pharmacy). Using student population figures and seven performance measures, they were able to estimate the level of demand for services in the clinic. They also show that patient inter-arrival times are distributed negative exponential with the mean changing according to the time of day, and patient service times are distributed Erlang-K. Using these data, they investigate the effects of adding another pharmacist to the pharmacy. A multifactor experimental design was developed to examine the relationships between the controllable system variables and the system performance variables. They show that different calling population sizes and different levels of staffing can impact the performance measures at each station. Additionally, the aggregation of two or more similar facilities can cause an increase in the average number of patients waiting at each of the remaining facilities and the average patient waiting times, though these increases were offset by a significant decrease in the staff idle times and staff costs. These studies suggest that staffing levels and staff distributions have a significant impact on patient throughput and waiting times. As with facility sizing and planning, discrete-event simulation can be an effective tool to study various staffing strategies for a wide variety of health care facilities and systems. As the national shortage of skilled clinicians (particularly nurses) deepens, such studies will become increasingly important to health care organizations as they attempt to optimally deploy scarce human resources.

3 Recent Innovations and Future Directions

There is a growing amount of literature on using discrete-event simulation to study the design and operation of health care delivery systems. Publications based on such studies have steadily increased from 8 in 1973–1977 to 28 in 1993–1997 to over 50 in 1998–2004. This positive trend can be attributed to the increasing demand for cost-cutting in health care coupled with an increase in the ease-of-use and power of discrete-event simulation software packages (especially over the past decade). A growing number of these studies attempt to apply optimization techniques to analyze discrete-event simulation models. Despite the increase in health care simulation studies and the integration of discrete-event simulation and optimization techniques, few studies focus on complex integrated systems. This may be a result of the associated complexity issues and resource requirements required for such studies. Moreover, no matter how complex modeled systems are or what techniques are applied, it will continue to be a challenge to implement the results of such studies. However, recent advances in discrete-event simulation software may help to overcome some of these obstacles.

Most discrete-event simulation models focus on individual units within multifacility clinics or hospitals. Using a macroscopic analysis of multifacility systems, discrete-event simulation can be used to estimate patient demand (directly related to arrival rates), utilization of staff, and overall costs. The estimation of these performance measures may not be possible in a microscopic, single level model, due to the duplication of and overlapping of facilities and services. Discrete-event simulation models that capture the interaction of major service departments and support services in a hospital, and the information that can be gained from analyzing the system as a whole, can be invaluable to hospital planners and administrators.

To remain competitive in today's market, the health care industry is being forced to integrate hospitals and clinics, especially the ever-growing number of ambulatory care facilities, into health maintenance organizations (HMO), multi-hospital, or multi-clinic organizations. This presents a challenging application for discrete-event simulation: to operate these networks of clinics or departments efficiently and cost effectively. Studies in multi-facility simulation models have been conducted by Rising et al. (1973), Aggarwal and Stafford (1976), Hancock and Walters (1984), Swisher et al. (1997, 2001), and Lowery and Martin (1992).

A benefit from simulating integrated systems is the more realistic representation of the system under study, hence greater confidence in the results. Though this may not be significant when analyzing a small system, the consequences of invalid results or the lack of a thorough study may potentially be a costly decision for large multimillion dollar organizations. With this potential benefit, the question that has to be asked is: Why is there a lack of literature in this area? The answer may lie in one or both of two issues: (1) the level of complexity and resulting data requirements of the simulation model and/or (2) the resource requirements, including time and cost.

A widely recognized guideline in discrete-event simulation modeling is to keep the model as simple as possible while capturing the necessary measures of interest. This is reiterated by Dearie et al. (1976) who stress the importance of capturing only relevant performance variables when creating a simple, though not necessarily the most complete model. They suggest that it is best to depict the various subsystems at the lowest level of complexity such that the model is accurate while providing information that is easily interpreted. Moreover, Lowery (1996) suggests using simple analytical models if they can provide the necessary level of detail. However, when analyzing integrated systems, the level of detail that is required far exceeds the complexity and demands of analytical techniques. Therefore, care must be taken when determining the required level of detail (since more detail typically means that more data must be collected). The soft system methodology (SSM), an approach that aids in determining the level of detail, identifying system boundaries, and ascertaining system activities, is suggested by Lehaney and Paul (1994, 1996) and Lehaney and Hlupic (1995). Through increased participation of the users/customers, SSM encourages acceptability of the model, its results, and eventually the model's implementation.

Resource requirements, such as the length of time, the cost, and the skills necessary to complete the project must be fully considered before commencing such a large-scale project. Today's health care delivery environment is rapidly changing and if the process of developing and searching for a solution requires a large investment in time and resources, the system may be outdated before the results from the simulation study can be implemented. Consequently, an adequate amount of resources must be dedicated to the project to ensure completion of the study in a reasonable length of time. For example, the cost of collecting the required data (in terms of time and money), the cost of purchasing a discrete-event simulation software package that would ease the development of complex models, and the cost of skilled consultants or in-house engineers may all be prohibitive.

Given that discrete-event simulation is not an optimization tool, it can only provide estimates of performance measures for various system alternatives. Moreover, discrete-event simulation models typically have several output performance measures upon which to optimize, hence creating a multi-criteria objective function environment. There are several advantages and disadvantages of using either discrete-event simulation methodologies or optimization techniques to model complex systems. Karnon (2003), Davies and Davies (1994) and Stafford (1978) compare discrete-event simulation modeling to several techniques, such as Markov chain analysis, semi-Markov chain analysis, input–output analysis, and queuing analysis of an outpatient clinic. They find that discrete-event simulation is particularly well suited for modeling health care clinics due to the complexity of such systems, whereas many optimization techniques, such as linear programming, have a limited capacity for characterizing the complexities of medical systems. An optimization technique may require too many unrealistic assumptions about the process, hence rendering the solution invalid and unrealistic. For example, optimization models cannot be used to study the details of the day to day operations of a medical clinic, such as appointment scheduling, service routing, and service priorities, which can be easily captured by a discrete-event simulation model. On the other hand, many optimization models require only one experimental run to produce optimal or near optimal solutions, though the complexity of the model may result in an intractable solution; whereas, discrete-event simulation models require a large amount of effort in time, cost, and data collection. For all of these reasons, operations researchers have attempted to combine simulation with deterministic operations research techniques, such as linear programming, to simultaneously exploit the advantages of using both techniques.

Several studies have reported success in combining these techniques to find the best staffing allocations and facility sizes. A common technique when applying an optimization methodology to discrete-event simulation models of health care clinics is a recursive method employed by Carlson et al. (1979), Kropp et al. (1978), and Kropp and Hershey (1979). First, an optimization technique is used to analyze and reduce the number of alternatives of the system at an aggregate level (i.e., the total system level). These results are then used in a more complex and detailed discrete-event simulation model of the same system, which is then used to identify additional information and validate the results. Finally, these additional constraints are passed back into the optimization model and this process is iteratively repeated. Similarly, Butler et al. (1992a, b) employ a two phase approach by first using quadratic integer programming for facility layout and capacity allocation questions, and then a discrete-event simulation model to capture the complexities of alternative scheduling and bed assignment problems. Baesler and Sepúlveda (2001) extend the study of Sepúlveda et al. (1999) by using a simulation model of a new cancer treatment facility as a case study for solving a multi-objective (minimize patient waiting time, maximize chair utilization, minimize closing time, maximize nurse utilization) simulation optimization problem.

All of these studies use a variety of optimization techniques to arrive at parameters for the discrete-event simulation models. In general, recursive simulation optimization techniques can be very difficult, and therefore, costly to implement in the health care sector. However, a growing number of simulation software packages have appeared that provide an optimization add-on (Carson and Maria

1997). Instead of an exhaustive, time-consuming, and indiscriminate search for an optimal alternative, discrete-event simulation software companies are now starting to provide special search algorithms to guide a simulation model to an optimal or near-optimal solution. Examples of these include an add-on to MicroSaint 2.0 called OptQuest, that uses a scatter search technique (based on tabu search) to find the best value for one or multiple objective functions (Glover et al. 1996). Other optimization simulation software includes ProModel's SimRunner Optimization (Benson 1997) and AutoStat for AutoMod (Carson 1996).

Since their introduction, discrete-event simulation software packages have gone through a series of technological leaps and advances. First, the introduction of visually oriented graphical outputs has greatly aided in the verification and validation of models and results (Gipps 1986; Sargent 1992), though this does not necessarily guarantee model correctness (Paul 1989). Moreover, discrete-event simulation model animation is primarily used to present movie-like images of the actual operation of the model and system which, in essence, helps to sell insights into the system under study. Second, the wide use of the object-oriented paradigm (OOP) in discrete-event simulation software design enables analysts to model a system without writing a single line of code (Banks 1997). Numerous companies are developing general-purpose software packages incorporating the latest technologies (Banks 1996), with packages like MedModel (Harrell and Lange 2001; Harrell and Price 2000; Price and Harrell 1999; Heflin and Harrell 1998; Carroll 1996; Keller 1994) and ARENA with a health care template (Drevna and Kasales 1984) specifically aimed to serve the health care industry.

Jones and Hirst (1986) present one of the early articles on using visual simulation, using the discrete-event simulation software package See-Why. The visualization of different policies in the visual simulation of a surgical unit and surrounding resources plays an integral part in assisting managers in identifying the best solutions. Paul and Kuljis (1995) use CLIMSIM, a generic discrete-event simulation package, to illustrate how clinic appointments and operating policies can influence patient waiting time. Evans et al. (1996) use ARENA to model an emergency department using 13 patient categories. They reduce patients' length of stay using alternative scheduling rules for the number of nurses, technicians, and physicians on duty during each particular hour of the simulation run. In addition to these studies, several other visual simulation modeling project of interest have been conducted (including McGuire 1994 and Ritondo and Freedman 1993).

The number of health care organizations and government agencies using advanced discrete-event simulation software packages has grown, with much of their work and the results of their efforts not available in the open literature. Considering the number of easy to use discrete-event simulation software packages available today, it seems unusual to find that such a small number of visually oriented simulation models of health care clinics have been published. This may be attributed to the shifting face of simulation modelers. As discrete-event simulation models have become easier to build with new software packages, the type of users have also changed. Since it is no longer necessary to have an advanced technical degree to use discrete-event simulation software packages (due largely to the drag-and-drop operation of such packages), numerous nontechnical (and typically non-publishing) users have emerged. However, this development does not diminish the importance of contributions from operations research professionals to the health care field, since such individuals will continue to be needed to provide technical expertise when conducting or managing critical or large-scale discrete-event simulation projects.

Discrete-event simulation modeling of health care clinics has been extensively used to assist decision-makers to identify areas of service where efficiencies can be improved. For discrete-event simulation to reach its full potential as the key tool for analyzing health care clinics, the results from such simulation studies must be implemented. Unfortunately, in a survey of 200 papers reporting the results of discrete-event simulation studies in health care, Wilson (1981) found that only 16 projects reported successful implementations. A number of recommendations were given to increase the opportunities for and likelihood of implementation success. These recommendations include: the system being studied is actually in need of a decision, the project must be completed before a deadline, data must be available, and the organizer or the decision-maker must participate in the project.

Lowery (1994, 1996, 1998) addresses some additional implementation barriers, as well as solutions to help overcome the resistance to implementation. Some of the suggestions include animating the simulation model execution to more easily communicate the problem and the solution to the decision-maker, making sure management stays involved throughout the project, and avoiding too many assumptions or making the model too complex. She also suggests that management engineers must simplify the simulation process and improve their sales skills. Marsh (1979) lists three key elements necessary for the successful implementation of simulation results: total commitment and support from the users, credibility of the model, and the analyst must work with the real operations under study rather than any esoteric studies.

Despite the lack of implementation observed in the literature, other benefits can still be gained from conducting a discrete-event simulation study. The procedure and methodology of applying discrete-event simulation requires decision-makers and managers to work closely with the simulation analyst to provide details of the system, often for the first time. As a result, the manager is likely to gain a new perspective on the relationships between the available resources and the quality of health care services offered by the system. Rakich et al. (1991) study the effects of discrete-event simulation in management development. They conclude that conducting a simulation study not only develops a manager's decision-making skills, but also forces them to recognize the implications of system changes. Moreover, as also noted Wilson (1981), in the cases where managers developed their own discrete-event simulation models, implementation occurred much more frequently. Finally, Lowery (1996) notes that there are benefits, such as identifying unexpected problems unrelated to the original problem, which arise even if implementation fails.

4 Summary and Conclusions

In conclusion, this chapter surveys the literature (focusing primarily on the past 30 years) on the application of discrete-event simulation modeling and analysis to understand the operations of health care facilities. A significant amount of research has been conducted in the area of patient flow and asset allocation. The multiple performance measures associated with health care systems makes discrete-event simulation particularly well-suited to tackle problems in these domains. A large number of discrete-event simulation studies reported in the literature have the common theme that they attempt to understand the relationship that may exist between various inputs into a health care delivery system (e.g., patient scheduling and admission rules, patient routing and flow schemes, facility and staff resources) and various output performance measures from the system (e.g., patient throughput, patient waiting times, physician utilization, staff and facility utilization). The breadth and scope of units within hospitals and clinics makes it impossible to undertake a single comprehensive study that simultaneously addresses all of these issues.

The aforementioned observations, together with the dearth of literature in the area of complex integrated multi-facility systems, suggest the need to develop a comprehensive simulation modeling framework for determining clinical performance measures and interdepartmental resource relationships. Furthermore, this survey identified a number of continuing trends in discrete-event simulation software such as; the development of optimization add-ons, increased visualization, and the shift to an object-oriented paradigm. These powerful features will have the greatest impact when educating decision-makers on what changes need to be made and weakening the resistance to implementation. The outlook for discrete-event simulation in health care looks promising. The further development of more powerful high speed processing, distributed simulations (Baezner et al. 1990), and object-oriented simulation, will facilitate the creation of complex, but tractable, models of large integrated systems. Greater decision-maker buy-in will lead to model results being implemented more easily and frequently, providing greater opportunities for success. Twenty-first century health care decision-makers are faced with a complex and challenging environment. Costs are rising, human and fiscal resources are becoming scarcer, consumer expectations are rising, and technology is becoming more complex. It is crucial that health care managers make informed decisions on health care policies and the application of resources. Discrete-event simulation offers perhaps the most powerful and intuitive tool for the analysis and improvement of complex health care systems.

Acknowledgements The authors wish to thank Dr. R.W. Hall for his comments and feedback on an earlier draft of this chapter, resulting in a significantly improved final version. The first author has been supported in part by the National Science Foundation (DMI-0222597, DMI-0457176).

References

Aggarwal, S. C., & Stafford, E. F. (1976). *A simulation study to identify important design parameters of a typical outpatient health system and to analyze measures of its performance. Proceedings of the 1976 summer computer simulation conference* (pp. 544–553). Washington, DC: Simulation Council.

Akkerman, R., & Knip, M. (2004). Reallocation of beds to reduce waiting time for cardiac surgery. *Health Care Management Science, 7*, 119–126.

Alessandra, A. J., Grazman, T. E., Parameswaran, R., & Yavas, U. (1978). Using simulation in hospital planning. *Simulation, 30*(2), 62–67.

Alexopoulos, A., Goldsman, D., Fontanesi, J., Sawyer, M., De Guire, M., Kopald, D., et al. (2001). A discrete-event simulation application for clinics serving the poor. In B. A. Peters, J. S. Smith, D. J. Medeiros, & M. W. Rohrer (Eds.), *Proceedings of the 2001 winter simulation conference* (pp. 1386–1391). Arlington, VA: Institute of Electrical and Electronics Engineers.

Altinel, I. K., & Ulas, E. (1996). Simulation modeling for emergency bed requirement planning. *Annals of Operations Research, 67*, 183–210.

Amladi, P. (1984). Outpatient health care facility planning and sizing via computer simulation. In S. Sheppard, U. W. Pooch, & C. D. Pegden (Eds.), *Proceedings of the 1984 winter simulation conference* (pp. 705–711). Dallas, TX: Institute of Electrical and Electronics Engineers.

Badri, M., & Hollingsworth, J. (1993). A simulation model for scheduling in the emergency room. *International Journal of Operations and Production Management, 13*(3), 13–24.

Baesler, F. F., Jahnsen, H. E., & DaCosta, M. (2003). The use of simulation and design of experiments for estimating maximum capacity in an emergency room. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1903–1906). New Orleans, LA: Institute of Electrical and Electronics Engineers.

Baesler, F. F., & Sepúlveda, J. A. (2001). Multi-objective simulation optimization for a cancer treatment center. In B. A. Peters, J. S. Smith, D. J. Medeiros, & M. W. Rohrer (Eds.), *Proceedings of the 2001 winter simulation conference* (pp. 1405–1411). Arlington, VA: Institute of Electrical and Electronics Engineers.

Baezner, D., Lomow, G., & Unger, B. W. (1990). Sim++™: The transition to distributed simulation. In D. Nicol (Ed.), *Proceedings of the 1990 SCS western multiconference on simulation: Distributed simulation* (pp. 211–218). San Diego, CA: Society for Computer Simulation.

Bailey, N. T. (1952). A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting times. *Journal of the Royal Statistical Society, A14*, 185–199.

Baldwin, L. P., Eldabi, T., & Paul, R. J. (2004). Simulation in healthcare management: A soft approach (MAPIU). *Simulation Modelling Practice and Theory, 12*, 541–557.

Banks, J. (1996). Software for simulation. In J. M. Charnes, D. M. Morrice, D. T. Brunner, & J. J. Swain (Eds.), *Proceedings of the 1996 winter simulation conference* (pp. 31–38). Coronado, CA: Institute of Electrical and Electronics Engineers.

Banks, J. (1997). The future of simulation software: A panel discussion. In S. Andradottir, K. J. Healy, D. E. Withers, & B. L. Nelson (Eds.), *Proceedings of the 1997 winter simulation conference* (pp. 166–173). Atlanta, GA: Institute of Electrical and Electronics Engineers.

Banks, J., & Carson, J. S. (1987). Applying the simulation process. In A. Thesen, H. Grant, & W. D. Kelton (Eds.), *Proceedings of the 1987 winter simulation conference* (pp. 68–71). Atlanta, GA: Institute of Electrical and Electronics Engineers.

Benson, D. (1997). Simulation modeling and optimization using ProModel. In S. Andradottir, K. J. Healy, D. E. Withers, & B. L. Nelson (Eds.), *Proceedings of the 1997 winter simulation conference* (pp. 587–593). Atlanta, GA: Institute of Electrical and Electronics Engineers.

Blake, J. T., Carter, M. W., & Richardson, S. (1996). An analysis of emergency room wait time issues via computer simulation. *Information Systems and Operational Research, 34*(4), 263–273.

Blasak, R. E., Armel, W. S., Starks, D. W., & Hayduk, M. C. (2003). The use of simulation to evaluate hospital operations between the emergency department and a medical telemetry unit. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1887–1893). New Orleans, LA: Institute of Electrical and Electronics Engineers.

Bodtker, K., Wilson, L., & Godolphin, W. (1992). Simulation as an aid to clinical chemistry laboratory planning. In J. Anderson (Ed.), *Proceedings of the 1992 conference on simulation in health care and social services* (pp. 15–18). San Diego, CA: Society for Computer Simulation.

Butler, T. W., Karwan, K. R., & Sweigart, J. R. (1992a). Multi-level strategic evaluation of hospital plans and decisions. *Journal of the Operational Research Society, 43*(7), 665–675.

Butler, T. W., Karwan, K. R., Sweigart, J. R., & Reeves, G. R. (1992b). An integrative model-based approach to hospital layout. *IIE Transactions, 24*(2), 144–152.

Butler, T., Reeves, G., Karwan, K., & Sweigart, J. (1992c). Assessing the impact of patient care policies using simulation analysis. *Journal of the Society for Health Systems, 3*(3), 38–53.

Cahill, W., & Render, M. (1999). Dynamic simulation modeling of ICU bed availability. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 winter simulation conference* (pp. 1573–1576). Phoenix, AR: Institute of Electrical and Electronics Engineers.

Carlson, R. C., Hershey, J. C., & Kropp, D. H. (1979). Use of optimization and simulation models to analyze outpatient health care settings. *Decision Sciences, 10*, 412–433.

Carroll, D. (1996). MedModel-Healthcare simulation software. In J. M. Charnes, D. M. Morrice, D. T. Brunner, & J. J. Swain (Eds.), *Proceedings of the 1996 winter simulation conference* (pp. 441–446). Coronado, CA: Institute of Electrical and Electronics Engineers.

Carson, J. S. (1996). AutoStat output statistical analysis for AutoMod Users. In J. M. Charnes, D. M. Morrice, D. T. Brunner, & J. J. Swain (Eds.), *Proceedings of the 1996 winter simulation conference* (pp. 492–499). Coronado, CA: Institute of Electrical and Electronics Engineers.

Carson, Y., & Maria, A. (1997). Simulation optimization: Methods and applications. In S. Andradottir, K. J. Healy, D. E. Withers, & B. L. Nelson (Eds.), *Proceedings of the 1997 winter simulation conference* (pp. 118–126). Atlanta, GA: Institute of Electrical and Electronics Engineers.

Carter, M. W., O'Brien-Pallas, L. L., Blake, J. T., McGillis, L., & Zhu, S. (1992). Simulation, scheduling, and operating rooms. In J. G. Anderson (Ed.), *Proceedings of the 1992 simulation in health care and social services conference* (pp. 28–30). San Diego, CA: Simulation Council Inc.

Centeno, M. A., Albacete, C., Terzano, D. O., Carrillo, M., & Ogazon, T. (2000). A simulation study of the radiology department at JMH. In J. A. Joines, R. R. Barton, K. Kang, & P. A. Fishwick (Eds.), *Proceedings of the 2000 winter simulation conference* (pp. 1978–1984). Orlando, FL: Institute of Electrical and Electronics Engineers.

Centeno, M. A., Giachetti, R., Linn, R., & Ismail, A. M. (2003). A simulation-ILP based tool for scheduling ER staff. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1930–1938). New Orleans, LA: Institute of Electrical and Electronics Engineers.

Centeno, M. A., Lee, M. A., Lopez, E., Fernandez, H. R., Carrillo, M., & Ogazon, T. (2001). A simulation study of the labor and delivery rooms and JMH. In B. A. Peters, J. S. Smith, D. J. Medeiros, & M. W. Rohrer (Eds.), *Proceedings of the 2001 winter simulation conference* (pp. 1392–1400). Arlington, VA: Institute of Electrical and Electronics Engineers.

Chan, S.-Y. E., Ohlmann, J., Dunbar, S., Dunbar, C., Ryan, S., & Savory, P. (2002). Operations research methods applied to workflow in a medical records department. *Health Care Management Science, 5*, 191–199.

Coffin, M. A., Lassiter, G., Killingsworth, B., & Kleckley, J. W. (1993). A simulation model of an X-ray facility. In J. G. Anderson & M. Katzper (Eds.), *1993 SCS western multiconference on simulation: Simulation in the health sciences and services* (pp. 3–7). La Jolla, CA: Society for Computer Simulation.

Cohen, M. A., Hershey, J. C., & Weiss, E. N. (1980). Analysis of capacity decisions for progressive patient care hospital facilities. *Health Services Research, 15*, 145–160.

Currie, K., Iskander, W., Michael, L., & Coberly, C. (1984). Simulation modeling in health care facilities. In S. Sheppard, U. W. Pooch, & C. D. Pegden (Eds.), *Proceedings of the 1984 winter simulation conference* (pp. 713–717). Dallas, TX: Institute of Electrical and Electronics Engineers.

Davies, R., & Davies, H. (1994). Modeling patient flows and resource provision in health systems. *Omega, 22*(2), 123–131.

Dean, B., van Ackere, A., Gallivan, S., & Barber, N. (1999). When should pharmacists visit their wards? An application of simulation to planning hospital pharmacy services. *Health Care Management Science, 2*, 35–42.

Dearie, D., Gerson, J., & Warfield, T. (1976). *The development and use of a simulation model of an outpatient clinic. Proceedings of the 1976 summer computer simulation conference* (pp. 554–558). Washington, DC: Simulation Council.

Draeger, M. A. (1992). An emergency department simulation model used to evaluate alternative nurse staffing and patient population scenarios. In J. J. Swain, D. Goldsman, R. C. Crain, & J. R. Wilson (Eds.), *Proceedings of the 1992 winter simulation conference* (pp. 1057–1064). Arlington, VA: Institute of Electrical and Electronics Engineers.

Drevna, M. C., & Kasales, C. J. (1984). Introduction to Arena. In S. Sheppard, U. W. Pooch, & C. D. Pegden (Eds.), *Proceedings of the 1984 winter simulation conference* (pp. 431–436). Dallas, TX: Institute of Electrical and Electronics Engineers.

Dumas, M. B. (1984). Simulation modeling for hospital bed planning. *Simulation, 43*, 69–78.

Dumas, M. B. (1985). Hospital bed utilization: An implemented simulation approach to adjusting and maintaining appropriate levels. *Health Services Research, 20*(1), 43–61.

Edwards, R., Clague, J., Barlow, J., Clarke, M., Reed, P., & Rada, R. (1994). Operations research survey and computer simulation of waiting times in two medical outpatient clinic structures. *Health Care Analysis, 2*, 164–169.

Eldabi, T., & Paul, R. J. (2001). A proposed approach for modeling healthcare systems for understanding. In B. A. Peters, J. S. Smith, D. J. Medeiros, & M. W. Rohrer (Eds.), *Proceedings of the 2001 winter simulation conference* (pp. 1412–1420). Arlington, VA: Institute of Electrical and Electronics Engineers.

El-Darzi, J., Vasilakis, C., Chaussalet, T., & Millard, P. H. (1998). A simulation modeling approach to evaluating length of stay, occupancy, emptiness, and bed blocking in a hospital geriatric department. *Health Care Management Science, 1*, 143–149.

England, W., & Roberts, S. (1978). Applications of computer simulation in health care. In H. J. Highland, L. G. Hull, & N. R. Neilsen (Eds.), *Proceedings of the 1978 winter simulation conference* (pp. 665–676). Miami Beach, FL: Institute of Electrical and Electronics Engineers.

Evans, G. W., Gor, T. B., & Unger, E. (1996). A simulation model for evaluating personnel schedules in a hospital emergency department. In J. M. Charnes, D. M. Morrice, D. T. Brunner, & J. J. Swain (Eds.), *Proceedings of the 1996 winter simulation conference* (pp. 1205–1209). Coronado, CA: Institute of Electrical and Electronics Engineers.

Everett, J. E. (2002). A decision support simulation model for the management of an elective surgery waiting system. *Health Care Management Science, 5*, 89–95.

Ferrin, D. M., Miller, M. J., Wininger, S., & Neuendorf, M. S. (2004). Analyzing incentives and scheduling in a major metropolitan hospital operating room through simulation. In R. G. Ingalls, M. D. Rossetti, J. S. Smith, & B. A. Peters (Eds.), *Proceedings of the 2004 winter simulation conference* (pp. 1975–1980). Washington, DC: Institute of Electrical and Electronics Engineers.

Fetter, R. B., & Thompson, J. D. (1965). The simulation of hospital systems. *Operations Research, 13*(5), 689–711.

Fitzpatrick, K. E., Baker, J. R., & Dave, D. S. (1993). An application of computer simulation to improve scheduling of hospital operating room facilities in the United States. *International Journal of Computer Applications in Technology, 6*(4), 215–224.

Freedman, R. W. (1994). Reduction of average length of stay in the emergency room using discrete simulation. In J. G. Anderson & M. Katzper (Eds.), *Proceedings of simulation in the health sciences* (pp. 6–8). Tempe, AR: Society for Computer Simulation.

Gabaeff, S. C., & Lennon, J. (1991). *New computerized technology for the design and planning of emergency departments. Proceedings of the American Hospital Association* (pp. 1–15). Anaheim, CA: American Hospital Association.

Garcia, M. L., Centeno, M. A., Rivera, C., & DeCario, N. (1995). Reducing time in an emergency room via a fast-track. In C. Alexopoulos, K. Kang, W. R. Lilegdon, & D. Goldsman (Eds.), *Proceedings of the 1995 winter simulation conference* (pp. 1048–1053). Washington DC: Institute of Electrical and Electronics Engineers.

Gipps, P. J. (1986). The role of computer graphics in validating simulation models. *Mathematics and Computers in Simulation, 28*(4), 285–289.

Glover, F., Kelly, J. P., & Laguna, M. L. (1996). New advances and applications of combining simulation and optimization. In J. M. Charnes, D. M. Morrice, D. T. Brunner, & J. J. Swain (Eds.), *Proceedings of the 1996 winter simulation conference* (pp. 144–152). Coronado, CA: Institute of Electrical and Electronics Engineers.

Godolphin, W., Bodtker, K., & Wilson, L. (1992). Simulation modeling: A tool to help predict the impact of automation in clinical laboratories. *Laboratory Robotics and Automation, 4*(5), 249–255.

Goitein, M. (1990). Waiting patiently. *The New England Journal of Medicine, 323*(9), 604–608.

Groothuis, S., Goldschmidt, H. M. J., Drupsteen, E. J., de Vries, J. C. M., Hasman, A., & van Merode, G. G. (2002). Application of computer simulation analysis to assess the effects of relocating a hospital phlebotomy department. *Annals of Clinical Biochemistry, 39*, 261–272.

Groothuis, S., van Merode, G. G., & Hasman, A. (2001). Simulation as decision tool for capacity planning. *Computers Methods and Programs in Biomedicine, 66*, 139–151.

Guo, M., Wagner, M., & West, C. (2004). Outpatient clinic scheduling—A simulation approach. In R. G. Ingalls, M. D. Rossetti, J. S. Smith, & B. A. Peters (Eds.), *Proceedings of the 2004 winter simulation conference* (pp. 1981–1987). Washington, DC: Institute of Electrical and Electronics Engineers.

Hancock, W. M., Magerlein, D. B., Storer, R. H., & Martin, J. B. (1978). Parameters affecting hospital occupancy and implications for faculty sizing. *Health Services Research, 13*, 276–289.

Hancock, W., & Walter, P. (1979). The use of computer simulation to develop hospital systems. *Simuletter, 10*(4), 28–32.

Hancock, W., & Walter, P. (1984). The use of admissions simulation to stabilize ancillary workloads. *Simulation, 43*(2), 88–94.

Harper, P. R. (2002). A framework for operational modelling of hospital resources. *Health Care Management Science, 5*, 165–173.

Harrell, C. R., & Lange, V. R. (2001). Healthcare simulation modeling and optimization using MEDMODEL. In B. A. Peters, J. S. Smith, D. J. Medeiros, & M. W. Rohrer (Eds.), *Proceedings of the 2001 winter simulation conference* (pp. 233–238). Arlington, VA: Institute of Electrical and Electronics Engineers.

Harrell, C. R., & Price, R. N. (2000). Healthcare simulation modeling and optimization using MEDMODEL. In J. A. Joines, R. R. Barton, K. Kang, & P. A. Fishwick (Eds.), *Proceedings of the 2000 winter simulation conference* (pp. 203–207). Orlando, FL: Institute of Electrical and Electronics Engineers.

Harris, R. A. (1985). Hospital bed requirements planning. *European Journal of Operational Research, 25*, 121–126.

Hashimoto, F., & Bell, S. (1996). Improving outpatient clinic staffing and scheduling with computer simulation. *Journal of General Internal Medicine, 11*, 182–184.

Heflin, D. L., & Harrell, C. R. (1998). Healthcare simulation modeling and optimization using MEDMODEL. In J. D. Medeiros, E. F. Watson, J. S. Carson, & M. S. Manivannan (Eds.), *Proceedings of the 1998 winter simulation conference* (pp. 185–190). Washington, DC: Institute of Electrical and Electronics Engineers.

Ishimoto, K., Ishimitsu, T., Koshiro, A., & Hirose, S. (1990). Computer simulation of optimum personnel assignment in hospital pharmacy using a work-sampling method. *Medical Informatics, 15*(4), 343–354.

Iskander, W. H., & Carter, D. M. (1991). A simulation model for a same day care facility at a university hospital. In B. L. Nelson, W. D. Kelton, & G. M. Clark (Eds.), *Proceedings of the 1991 winter simulation conference* (pp. 846–853). Phoenix, AR: Institute of Electrical and Electronics Engineers.

Isken, M. W., Ward, T. J., & McKee, T. C. (1999). Simulating outpatient obstetrical clinics. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 winter simulation conference* (pp. 1557–1563). Phoenix, AR: Institute of Electrical and Electronics Engineers.

Johnson, W. C. (1998). Birth of a new maternity process. In J. D. Medeiros, E. F. Watson, J. S. Carson, & M. S. Manivannan (Eds.), *Proceedings of the 1998 winter simulation conference* (pp. 1429–1432). Washington, DC: Institute of Electrical and Electronics Engineers.

Jones, L. M., & Hirst, A. J. (1986). Visual simulation in hospitals: A managerial or a political tool? *European Journal of Operational Research, 29*, 167–177.

Jun, J. B., Jacobson, S. H., & Swisher, J. R. (1999). Applications of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society, 50*(2), 109–123.

Kachhal, S. K., Klutke, G. A., & Daniels, E. B. (1981). Two simulation applications to outpatient clinics. In T. I. Oren, C. M. Delfosse, & C. M. Shubl (Eds.), *Proceedings of the 1981 winter simulation conference* (pp. 657–665). Atlanta, GA: Institute of Electrical and Electronics Engineers.

Kanon, D. (1974). Simulation of waiting line problems in a hospital setting. In J. Anderson & J. M. Forsythe (Eds.), *Proceedings of the Ist world conference on medical information* (pp. 503–507). Stockholm, Netherlands: North-Holland.

Karnon, J. (2003). Alternative decision modelling techniques for the evaluation of health care technologies: Markov processes versus discrete-event simulation. *Health Economics, 12*, 837–848.

Keller, L. F. (1994). MedModel-specialized software for the healthcare industry. In J. D. Tew, S. Manivannan, D. A. Sadowski, & A. F. Seila (Eds.), *Proceedings of the 1994 winter simulation conference* (pp. 533–537). Lake Buena Vista, FL: Institute of Electrical and Electronics Engineers.

Kho, J. W., & Johnson, G. M. (1976). Computer simulation of a hospital health-care delivery system. In R. G. Sargent, H. J. Highland, & T. J. Schriber (Eds.), *Proceedings of the 1976 bicentennial winter simulation conference* (pp. 349–360). Gaithersburg, MD: Institute of Electrical and Electronics Engineers.

Kirtland, A., Lockwood, J., Poisker, K., Stamp, L., & Wolfe, P. (1995). Simulating an emergency department is as much fun as.... In C. Alexopoulos, K. Kang, W. R. Lilegdon, & D. Goldsman (Eds.), *Proceedings of the 1995 winter simulation conference* (pp. 1039–1042). Washington, DC: Institute of Electrical and Electronics Engineers.

Klafehn, K. A. (1987). Impact points in patient flows through a radiology department provided through simulation. In A. Thesen, H. Grant, & W. D. Kelton (Eds.), *Proceedings of the 1987 winter simulation conference* (pp. 914–918). Atlanta, GA: Institute of Electrical and Electronics Engineers.

Klafehn, K. A., & Connolly, M. (1993). The simulation/animation of a new outpatient hematology laboratory. In J. G. Anderson & M. Katzper (Eds.), *1993 SCS western multiconference on* *simulation: Simulation in the health sciences and services* (pp. 12–15). La Jolla, CA: Society for Computer Simulation.

Klafehn, K. A., & Owens, D. (1987). A simulation model designed to investigate resource utilization in a hospital emergency room. In W. Stead (Ed.), *Proceedings of the eleventh annual symposium on computer applications in medical care* (pp. 676–679). Washington, DC: Institute of Electrical and Electronics Engineers.

Klafehn, K. A., Owens, D., Felter, R., Vonneman, N., & McKinnon, C. (1989). Evaluating the linkage between emergency medical services and the provision of scarce resources through simulation. In L. Kingsland III (Ed.), *Proceedings of the 13th annual symposium on computer applications in medical care* (pp. 335–339). Washington, DC: Institute of Electrical and Electronics Engineers.

Klassen, K. J., & Rohleder, T. R. (1996). Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management, 14*, 83–101.

Klein, R. W., Dittos, R. S., Roberts, S. D., & Wilson, J. R. (1993). Simulation modeling and health-care decision making. *Medical Decision Making, 13*(4), 347–354.

Kletke, M. G., & Dooley, T. E. (1984). A simulation model of a maternity ward: A key planning tool for hospital administrators, *Proceedings of the Conference on Simulation in Health Care Delivery System* (pp. 35–39), San Diego, California, USA, 2–4 February.

Kraitsik, M. J., & Bossmeyer, A. (1993). Simulation applied to planning an emergency department expansion. In J. G. Anderson & M. Katzper (Eds.), *1993 SCS western multiconference on simulation: Simulation in the health sciences and services* (pp. 19–27). La Jolla, CA: Society for Computer Simulation.

Kropp, D., Carlson, R. C., & Jucker, J. V. (1978). Use of both optimization and simulation models to analyze complex systems. In H. J. Highland, L. G. Hull, & N. R. Neilsen (Eds.), *Proceedings of the 1978 winter simulation conference* (pp. 195–201). Miami Beach, FL: Institute of Electrical and Electronics Engineers.

Kropp, D., & Hershey, J. (1979). Recursive optimization-simulation modeling for the analysis of ambulatory health care facilities. *Simuletter, 10*(4), 43–46.

Kumar, A. P., & Kapur, R. (1989). Discrete simulation application-scheduling staff for the emergency room. In E. A. MacNair, K. J. Musselman, & P. Heidelberger (Eds.), *Proceedings of the 1989 winter simulation conference* (pp. 1112–1120). Washington, DC: Institute of Electrical and Electronics Engineers.

Kuzdrall, P. J., Kwak, N. K., & Schmitz, H. H. (1981). Simulating space requirements and scheduling policies in a hospital surgical suite. *Simulation, 27*, 163–171.

Kwak, N., Kuzdrall, P., & Schmitz, H. (1975). Simulating the use of space in a hospital surgical suite. *Simulation, 24*(5), 147–152.

Lach, J. M., & Vázquez, R. M. (2004). Simulation model of the telemedicine program. In R. G. Ingalls, M. D. Rossetti, J. S. Smith, & B. A. Peters (Eds.), *Proceedings of the 2004 winter simulation conference* (pp. 2012–2017). Washington, DC: Institute of Electrical and Electronics Engineers.

Lambo, D. (1983). An optimization-simulation model of a rural health center in Nigeria. *Interfaces, 13*(3), 29–35.

Law, A. M., & Kelton, W. D. (2000). *Simulation modelling and analysis* (3rd ed.). New York, NY: McGraw-Hill.

Lehaney, B., & Hlupic, V. (1995). Simulation modeling for resource allocation and planning in the health sector. *Journal of the Royal Society of Health, 115*(6), 382–385.

Lehaney, B., & Paul, R. J. (1994). Using soft systems methodology to develop a simulation of out-patient services. *Journal of the Royal Society for Health, 114*, 248–251.

Lehaney, B., & Paul, R. J. (1996). The use of soft systems methodology in the development of a simulation of outpatient services at Watford General Hospital. *Journal of the Operational Research Society, 47*, 864–870.

Lennon, J. (1992). *Simulation in the design and planning of emergency departments. 1992 SCS western multiconference: Simulation in education for business, management, and MIS* (pp. 93–97). Newport Beach, CA: Society for Computer Simulation.

Levy, J. L., Watford, B. A., & Owen, V. T. (1989). Simulation analysis of an outpatient services facility. *Journal of the Society for Health Systems, 1*(2), 35–49.

Lim, T., Uyeno, D., & Vertinsky, I. (1975). Hospital admissions systems: A simulation approach. *Simulation and Games, 6*(2), 188–201.

Liyanage, L., & Gale, M. (1995). *Quality improvement for the Campbelltown hospital emergency service. 1995 I.E. international conference on systems, man, and cybernetics* (pp. 1997–2002). Vancouver, BC: Institute of Electrical and Electronics Engineers.

Lopez-Valcarcel, B. G., & Perez, P. B. (1994). Evaluation of alternative functional designs in an emergency department by means of simulation. *Simulation, 63*(1), 20–28.

Lowery, J. C. (1992). Simulation of a hospital's surgical suite and critical care area. In J. J. Swain, D. Goldsman, R. C. Crain, & J. R. Wilson (Eds.), *Proceedings of the 1992 winter simulation conference* (pp. 1071–1078). Arlington, VA: Institute of Electrical and Electronics Engineers.

Lowery, J. C. (1993). Multi-hospital validation of critical care simulation model. In G. W. Evans, M. Mollaghasemi, E. C. Russell, & W. E. Biles (Eds.), *Proceedings of the 1993 winter simulation conference* (pp. 1207–1215). Los Angeles, CA: Institute of Electrical and Electronics Engineers.

Lowery, J. C. (1994). Barriers to implementing simulation in health care. In J. D. Tew, S. Manivannan, D. A. Sadowski, & A. F. Seila (Eds.), *Proceedings of the 1994 winter simulation conference* (pp. 868–875). Lake Buena Vista, FL: Institute of Electrical and Electronics Engineers.

Lowery, J. C. (1996). Introduction to simulation in health care. In J. M. Charnes, D. M. Morrice, D. T. Brunner, & J. J. Swain (Eds.), *Proceedings of the 1996 winter simulation conference* (pp. 78–84). Coronado, CA: Institute of Electrical and Electronics Engineers.

Lowery, J. C. (1998). Getting started in simulation healthcare. In J. D. Medeiros, E. F. Watson, J. S. Carson, & M. S. Manivannan (Eds.), *Proceedings of the 1998 winter simulation conference* (pp. 31–35). Washington, DC: Institute of Electrical and Electronics Engineers.

Lowery, J. C., & Davis, J. A. (1999). Determination of operating room requirements using simulation. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 winter simulation conference* (pp. 1568–1572). Phoenix, AR: Institute of Electrical and Electronics Engineers.

Lowery, J. C., & Martin, J. B. (1992). Design and validation of a critical care simulation model. *Journal of the Society for Health Systems, 3*(3), 15–36.

Magerlein, J. M., & Martin, J. B. (1976). Surgical demand scheduling: A review. *Health Services Research, 11*, 53–68.

Mahachek, A. (1992). An introduction to patient flow simulation for health-care managers. *Journal of the Society for Health Systems, 3*(3), 73–81.

Mahachek, A. R., & Knabe, T. L. (1984). Computer simulation of patient flow in obstetrical/gynecology clinics. *Simulation, 30*, 95–101.

Mahapatra, S., Koelling, C. P., Patvivatsiri, L., Fraticelli, B., Eitel, D., & Grove, L. (2003). Pairing emergency severity index5-level triage data with computer aided system design to improve emergency department access and throughput. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1917–1925). New Orleans, LA: Institute of Electrical and Electronics Engineers.

Marsh, J. (1979). Simulation as a decision aid: A management perspective. *Simuletter, 10*(5), 47–49.

Martin, E., Grønhaug, R., & Haugene, K. (2003). Proposals to reduce over-crowding, lengthy stays and improve patient care: Study on the geriatric department In Norway's largest hospital. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1876–1881). New Orleans, LA: Institute of Electrical and Electronics Engineers.

Masterson, B. J., Mihara, T. G., Miller, G., Randolph, S. C., Forkner, M. E., & Crouter, A. L. (2004). Using models and data to support optimization of the military health system: A case study in an intensive care unit. *Health Care Management Science, 7*, 217–224.

McGuire, F. (1994). Using simulation to reduce length of stay in emergency departments. In J. D. Tew, S. Manivannan, D. A. Sadowski, & A. F. Seila (Eds.), *Proceedings of the 1994 winter simulation conference* (pp. 861–867). Lake Buena Vista, FL: Institute of Electrical and Electronics Engineers.

McHugh, M. L. (1989). Computer simulation as a method for selecting nurse staffing levels in hospitals. In E. A. MacNair, K. J. Musselman, & P. Heidelberger (Eds.), *Proceedings of the 1989 winter simulation conference* (pp. 1121–1129). Washington, DC: Institute of Electrical and Electronics Engineers.

Meier, L., Sigal, E., & Vitale, F. R. (1985). The use of simulation model for planning ambulatory surgery. In D. T. Gantz, G. C. Blais, & S. L. Solomon (Eds.), *Proceedings of the 1985 winter simulation conference* (pp. 558–563). San Francisco, CA: Institute of Electrical and Electronics Engineers.

Miller, M. J., Ferrin, D. M., & Messer, M. G. (2004). Fixing the emergency department: A transformational journey with EDsim. In R. G. Ingalls, M. D. Rossetti, J. S. Smith, & B. A. Peters (Eds.), *Proceedings of the 2004 winter simulation conference* (pp. 1988–1993). Washington, DC: Institute of Electrical and Electronics Engineers.

Miller, M. J., Ferrin, D. M., & Szymanski, J. M. (2003). Simulating six sigma improvement ideas for a hospital emergency department. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1926–1929). New Orleans, LA: Institute of Electrical and Electronics Engineers.

Morrison, B. P., & Bird, B. C. (2003). A methodology for modeling front office and patient care processes in ambulatory health care. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1882–1886). New Orleans, LA: Institute of Electrical and Electronics Engineers.

Mukherjee, A. K. (1991). A simulation model for management of operations in the pharmacy of a hospital. *Simulation, 56*(2), 91–103.

Murphy, D. R., & Sigal, E. (1985). Evaluating surgical block scheduling using computer simulation. In D. T. Gantz, G. C. Blais, & S. L. Solomon (Eds.), *Proceedings of the 1985 winter simulation conference* (pp. 551–557). San Francisco, CA: Institute of Electrical and Electronics Engineers.

O'Kane, P. C. (1981). A simulation model of a diagnostic radiology department. *European Journal of Operational Research, 6*, 38–45.

Olson, E., & Dux, L. E. (1994). Computer model targets best route for expanding hospital surgicenter. *Industrial Engineering, 26*(9), 24–26.

Osidach, V. Z., & Fu, M. C. (2003). Computer simulation of a mobile examination center. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1868–1875). New Orleans, LA: Institute of Electrical and Electronics Engineers.

Paul, R. J. (1989). Visual simulation: Seeing is believing. *Impacts of Recent Computer Advances on Operations Research, 9*, 422–432.

Paul, R. J., & Kuljis, J. (1995). A generic simulation package for organizing outpatient clinics. In C. Alexopoulos, K. Kang, W. R. Lilegdon, & D. Goldsman (Eds.), *Proceedings of the 1995 winter simulation conference* (pp. 1043–1047). Washington, DC: Institute of Electrical and Electronics Engineers.

Price, R. N., & Harrell, C. R. (1999). Healthcare simulation modeling and optimization using MEDMODEL. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 winter simulation conference* (pp. 215–219). Phoenix, AR: Institute of Electrical and Electronics Engineers.

Rakich, J. S., Kuzdrall, P. J., Klafehn, K. A., & Krigline, A. G. (1991). Simulation in the hospital setting: Implications for managerial decision making and management development. *Journal of Management Development, 10*(4), 31–37.

Ramakrishnan, S., Nagarkar, K., DeGennaro, M., Srihari, K., Courtney, A. K., & Emick, F. (2004). A study of the CT scan area of a healthcare provider. In R. G. Ingalls, M. D. Rossetti, J. S. Smith, & B. A. Peters (Eds.), *Proceedings of the 2004 winter simulation conference* (pp. 2025–2031). Washington, DC: Institute of Electrical and Electronics Engineers.

Ramis, F. J., Palma, J. L., & Baesler, F. F. (2001). The use of simulation for process improvement at an ambulatory surgery center. In B. A. Peters, J. S. Smith, D. J. Medeiros, & M. W. Rohrer (Eds.), *Proceedings of the 2001 winter simulation conference* (pp. 1401–1404). Arlington, VA: Institute of Electrical and Electronics Engineers.

Rising, E. J., Baron, R., & Averill, B. (1973). A systems analysis of a university-health-service outpatient clinic. *Operations Research, 21*(5), 1030–1047.

Ritondo, M., & Freedman, R. W. (1993). The effects of procedure scheduling on emergency room throughput: A simulation study. In J. G. Anderson & M. Katzper (Eds.), *1993 SCS western multiconference on simulation: Simulation in the health sciences and services* (pp. 8–11). La Jolla, CA: Society for Computer Simulation.

Rohleder, T. R., & Klassen, K. J. (2002). Rolling horizon appointment scheduling: A simulation study. *Health Care Management Science, 5*, 201–209.

Romanin-Jacur, G., & Facchin, P. (1987). Optimal planning of a pediatric semi-intensive care unit via simulation. *European Journal of Operational Research, 29*, 192–198.

Rossetti, M. D., Kumar, A., & Felder, R. A. (1998). Mobile robot simulation of clinic laboratory deliveries. In J. D. Medeiros, E. F. Watson, J. S. Carson, & M. S. Manivannan (Eds.), *Proceedings of the 1998 winter simulation conference* (pp. 1415–1421). Washington, DC: Institute of Electrical and Electronics Engineers.

Rossetti, M. D., Trzcinski, G. F., & Syverud, S. A. (1999). Emergency department simulation and determination of optimal attending physician staffing schedules. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 winter simulation conference* (pp. 1532–1540). Phoenix, AR: Institute of Electrical and Electronics Engineers.

Samaha, S., Armel, W. S., & Starks, D. W. (2003). The use of simulation to reduce the length of stay in an emergency department. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1907–1911). New Orleans, LA: Institute of Electrical and Electronics Engineers.

Sanchez, S. M., Ogazon, T., Ferrin, D. M., Scpúlveda, J. A., & Ward, T. J. (2000). Emerging issues in healthcare simulation. In J. A. Joines, R. R. Barton, K. Kang, & P. A. Fishwick (Eds.), *Proceedings of the 2000 winter simulation conference* (pp. 1999–2003). Orlando, FL: Institute of Electrical and Electronics Engineers.

Sargent, R. G. (1992). Validation and verification of simulation models. In J. J. Swain, D. Goldsman, R. C. Crain, & J. R. Wilson (Eds.), *Proceedings of the 1992 winter simulation conference* (pp. 104–114). Arlington, VA: Institute of Electrical and Electronics Engineers.

Sepúlveda, J. A., Thompson, W. J., Baesler, F. F., Alvarez, M. I., & Cahoon, L. E., III. (1999). The use of simulation for process improvement in a cancer treatment center. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 winter simulation conference* (pp. 1541–1548). Phoenix, AR: Institute of Electrical and Electronics Engineers.

Sinreich, D., & Marmor, Y. N. (2004). A simple and intuitive simulation tool for analyzing emergency department operations. In R. G. Ingalls, M. D. Rossetti, J. S. Smith, & B. A. Peters (Eds.), *Proceedings of the 2004 winter simulation conference* (pp. 1994–2002). Washington, DC: Institute of Electrical and Electronics Engineers.

Smith, S. R., Schroer, B. J., & Shannon, R. E. (1979). Scheduling of patients and resources for ambulatory health care. In H. J. Highland, M. G. Spiegel, & R. Shannon (Eds.), *Proceedings of the 1979 winter simulation conference* (pp. 553–562). San Diego, CA: Institute of Electrical and Electronics Engineers.

Smith, E. A., & Warner, H. R. (1971). Simulation of a multiphasic screening procedure for hospital admissions. *Simulation, 17*(2), 57–64.

Smith-Daniels, V. L., Schweikhart, S. B., & Smith-Daniels, D. E. (1988). Capacity management in health care services: Review and future research directions. *Decision Sciences, 19*, 889–918.

Stafford, E. F., Jr. (1978). *Simulation vs. mathematical analysis for systems modeling: A comparison of techniques Proceedings of the 1978 summer computer simulation conference* (pp. 53–159). Los Angeles, CA: AFIPS.

Stafford, E. F., Jr. (1976). *A general simulation model for multifacility outpatient clinics*, M.S. Thesis, Pennsylvania State University, Pennsylvania.

Stahl, J. E., Roberts, M. S., & Gazelle, S. (2003). Optimizing management and financial performance of the teaching ambulatory care clinic. *Journal of General Internal Medicine, 18*, 266–274.

Standridge, C. R. (1999). A tutorial on simulation in health care: Applications and issues. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 winter simulation conference* (pp. 49–55). Phoenix, AR: Institute of Electrical and Electronics Engineers.

Steward, D., & Standridge, C. R. (1996). A veterinary practice simulator based on the integration of expert system and process modeling. *Simulation, 66*, 143–159.

Swisher, J. R., & Jacobson, S. H. (2002). Evaluating the design of a family practice healthcare clinic using discrete-event simulation. *Health Care Management Science, 5*, 75–88.

Swisher, J. R., Jacobson, S. H., Jun, J. B., & Balci, O. (2001). Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers and Operations Research, 28*(2), 105–125.

Swisher, J. R., Jun, B. J., Jacobson, S. H., & Balci, O. (1997). Simulation of the Queston physician network. In S. Andradottir, K. J. Healy, D. E. Withers, & B. L. Nelson (Eds.), *Proceedings of the 1997 winter simulation conference* (pp. 1146–1154). Atlanta, GA: Institute of Electrical and Electronics Engineers.

Takakuwa, S. T., & Shiozaki, H. (2004). Functional analysis for operating emergency department of a general hospital. In R. G. Ingalls, M. D. Rossetti, J. S. Smith, & B. A. Peters (Eds.), *Proceedings of the 2004 winter simulation conference* (pp. 2003–2011). Washington, DC: Institute of Electrical and Electronics Engineers.

Tan, B. A., Gubaras, A., & Phojanamongkolkij, N. (2002). Simulation study of Dreyer urgent care facility. In E. Yücesan, C. H. Chen, J. L. Snowdon, & J. M. Charnes (Eds.), *Proceedings of the 2002 winter simulation conference* (pp. 1922–1927). San Diego, CA: Institute of Electrical and Electronics Engineers.

Valinsky, D. (1975). Simulation. In L. J. Shuman, R. D. Speas, & J. P. Young (Eds.), *Operations research in healthcare: A critical analysis*. Baltimore, MD: The Johns Hopkins University Press.

Vasilakis, C., & El-Darzi, E. (2001). A simulation study of the winter bed crisis. *Health Care Management Science, 4*, 31–36.

Vassilacopoulos, G. (1985). A simulation model for bed allocation to hospital inpatient departments. *Simulation, 45*(5), 233–241.

Vemuri, S. (1984). Simulated analysis of patient waiting time in an outpatient pharmacy. *American Journal of Hospital Pharmacy, 41*, 1127–1130.

Vissers, J. M. H. (1998). Health care management modelling: A process perspective. *Health Care Management Science, 1*, 77–85.

Walter, S. D. (1973). A comparison of appointment schedules in hospital radiology department. *British Journal of Preventive and Social Medicine, 27*, 160–167.

Weng, M. L., & Houshmand, A. A. (1999). Healthcare simulation: A case study at a local clinic. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 winter simulation conference* (pp. 1577–1584). Phoenix, AR: Institute of Electrical and Electronics Engineers.

Wiinamaki, A., & Dronzek, R. (2003). Using simulation in the architectural concept phase of an emergency department design. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1912–1916). New Orleans, LA: Institute of Electrical and Electronics Engineers.

Williams, S. V. (1983). How many intensive care beds are enough? *Critical Care Medicine, 11*(6), 412–416.

Williams, W. J., Covert, R. P., & Steele, J. D. (1967). Simulation modeling of a teaching hospital clinic. *Hospitals, 41*(21), 71–75.

Wilson, J. C. T. (1981). Implementation of computer simulation projects in health care. *Journal of the Operational Research Society, 32*, 825–832.

Wilt, A., & Goddin, D. (1989). Health care case study: Simulating staffing needs and work flow in an outpatient diagnostic center. *Industrial Engineering, 21*(5), 22–26.

Wong, C., Geiger, G., Derman, Y. D., Busby, C. R., & Carter, M. W. (2003). Redesigning the medication ordering, dispensing, and administration process in an acute care academic health sciences centre. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 winter simulation conference* (pp. 1894–1902). New Orleans, LA: Institute of Electrical and Electronics Engineers.

Wright, M. B. (1987). The application of a surgical bed simulation model. *European Journal of Operational Research, 32*, 26–32.

Zilm, F., Arch, D., & Hollis, R. B. (1983). An application of simulation modeling to surgical intensive care bed need analysis in a university hospital. *Hospital & Health Services Administration, 28*(5), 82–101.

Chapter 13
Process Mapping of a Regional Trauma System

David C. Evans, Douglas L. Andrusiek, and Boris Sobolev

Abstract This chapter describes the processes of care of the severely injured patient by using a "swim lanes" diagram to depict the sequence of care steps that govern patient flow, the agents that enact these steps, and the associated managerial and clinical tasks. Our objective was to create a clear and comprehensive narrative that would be useful to health care administrators tasked with operational decision-making in the oversight of a regionally integrated trauma system. As a representative illustration of modern trauma systems, we map the processes of trauma care occurring within Vancouver Coastal Health Authority (VCH) in the province of British Columbia (BC), Canada. We show how these steps influence associated processes, and which measures may best describe optimal system performance overall, and at key junctions.

Keywords Trauma systems • Process mapping • Performance evaluation

1 Introduction

In this chapter, we present process mapping of the flow of patients and information within a regional trauma system as a means to framing an effective whole-system performance improvement strategy. Using the example of organized trauma care in British Columbia (BC), Canada, we use an expanded swim lane diagram to illustrate how process mapping could help to clarify system goals and support the implementation of meaningful process measures useful for streamlining system performance.

D.C. Evans • B. Sobolev (✉)
University of British Columbia, Vancouver, BC, Canada
e-mail: david.evans@vch.ca; boris.sobolev@ubc.ca; sobolev@interchange.ubc.ca

D.L. Andrusiek
Edith Cowan University, Joondalup, WA, Australia
e-mail: d.andrusiek@ecu.edu.au

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_13, © Springer Science+Business Media New York 2013

Regional systems of advanced trauma care have been well developed and thoroughly described over the past three decades (Evans 2007). In its accreditation guidelines, the Trauma Association of Canada (TAC) offers the following broadly accepted definition of a trauma system:

> A fully comprehensive and inclusive trauma system is a preplanned, organized, and coordinated injury control effort in a defined geographic area (province or region) which: has an identifiable lead agency; is publicly administered, funded and accountable; engages in comprehensive injury surveillance, reporting and prevention programs; delivers the full spectrum of trauma care from the time of injury to recovery…; engages in research, training and performance improvement; [and] establishes linkages with an all-hazards emergency preparedness program (Trauma Association of Canada/Association canadienne de traumatologie 2007, p. 8).

More recent conceptualizations of trauma systems advocate a public health framework for injury management (US Department of Health and Human Services 2006), shifting the emphasis from optimizing outcomes after injury at the patient level towards minimizing burden of injury at the population level. Because multiple factors influence the societal burden of injury (e.g., regulation and legislation, injury prevention programs, education, emergency preparedness planning), designing and implementing system changes that confer true value is difficult. Indeed, how trauma systems fit *operationally* within a public health model of illness and injury management is challenging to specify and has not been well characterized. Figure 13.1 illustrates our conceptualization of how injury management could be viewed within a public health framework.

Recognizing that the integration of multiple parallel activities is critical to optimal system performance, modern health administrators seek to ensure that desired outcomes are achieved. Given that trauma systems generally consist of the coordinated effort of multiple independent agents and organizations working more toward organization-specific objectives rather than larger shared downstream goals, this effort is challenging.

Modern trauma systems remain principally focused on patient care processes aimed at delivering the right patient to the right place at the right time, and these processes command the major share of health care resources targeting injury management. As such, we believe that mapping the processes of trauma care provides administrators with critical knowledge about the influence care processes bring to bear on the public health objectives of minimizing burden of injury at a regional, state/provincial, or national level. In delineating care steps and overlaying appropriate indices that capture optimal performance, it becomes possible to gain a better understanding of how specific processes might influence desired outcomes on a larger scale. More importantly, the value of investment in key trauma system processes, whether implicit processes of care or explicit processes of population management, become more readily appreciable to decision-makers responsible for system design and development.

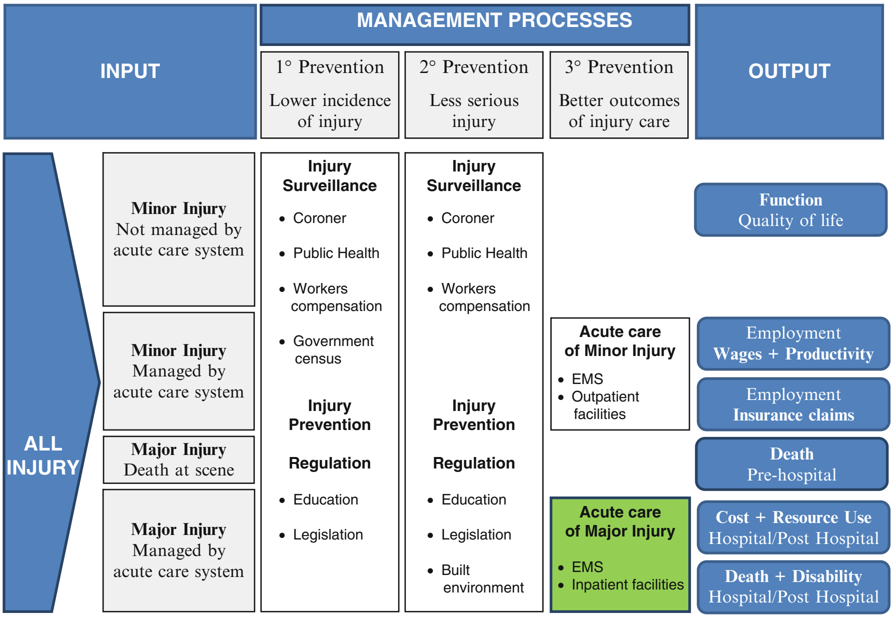


Fig. 13.1 Processes of injury management: the public health approach

2 Processes of Care for Major Trauma

Most health care services in BC are publicly administered and universally accessible through a public health insurance plan, as per the *Canada Health Act* governing hospital-based health care throughout Canada. Similar to the other Canadian provinces and territories, the BC Ministry of Health works with the province's five regional health authorities to organize, regulate, and oversee the delivery of all major health services, including pre-hospital emergency medical services and post-hospital rehabilitation and recovery care. In BC, emergency medical services (EMS) and patient transport and communications systems are single, integrated, province-wide organizations. Additionally, a provincial advisory body oversees trauma services with the support of a provincial trauma registry.

Clarifying how system components are organized to deliver system functions is essential to developing a usable process map that can drive effective performance improvement. Figure 13.2 depicts a full mapping of major trauma care, which we have developed through the continuum of prehospital, hospital, and post-acute care for BC. A key first step in mapping processes of care is the identification of the agents active in the sequence of steps that govern patient flow through the system and the processes—both managerial and clinical—attributable to them (Table 13.1). The sequential ordering of actions ascribed to these agents, organized into "swim lanes," creates the process map and its narrative.

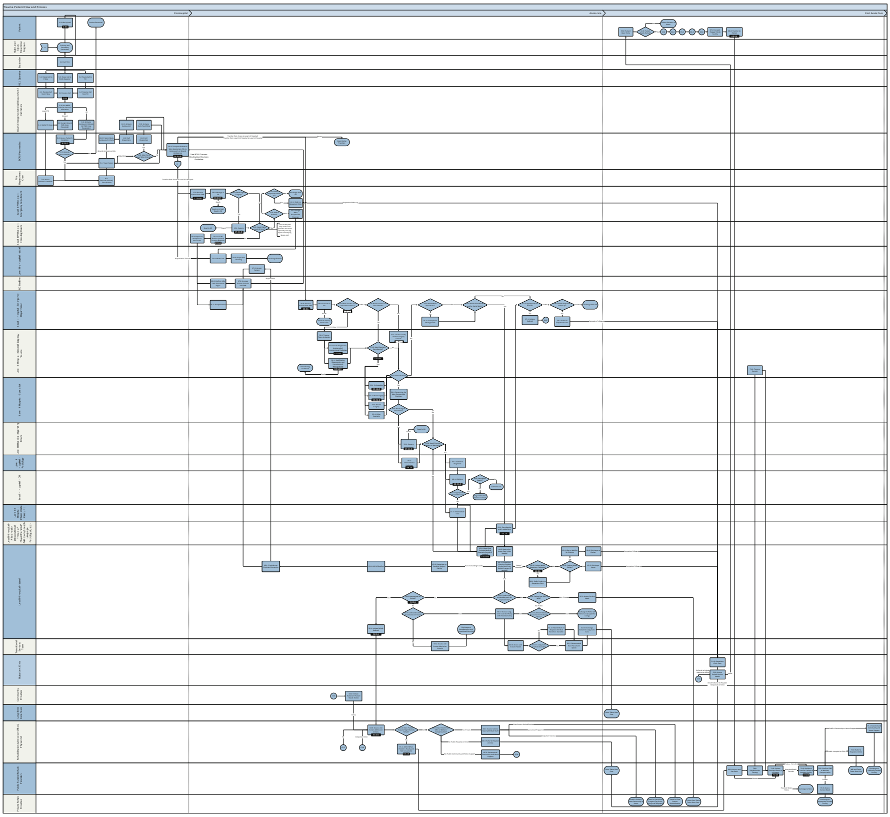

Trauma Patient Flow and Process
Pre-Hospital
Acute care
Post-Acute Care
Patient
High Level Injury Prevention Program
Bystander
911 Operator
BCAS Emergency Medical Dispatchers/ Calltakers
BCAS Paramedics
See BCAS Trauma Destination Decision Guideline

Fig. 13.2 Regional trauma system: integrated process map of patient flow

2.1 *Pre-hospital Care*

2.1.1 Patient Management

Figure 13.3 shows the key processes of pre-hospital trauma care, starting with notification of the occurrence of major injury anywhere within the province, which initiates the patient trajectory through the trauma system. Most commonly, the system is activated by a 911 bystander call to the provincial emergency communications system. Using basic information extracted from this initial contact, the 911 operator routes the call forward to police, fire, and/or EMS as appropriate. In cases of major trauma, the subsequently contacted EMS dispatcher assesses the level of medical response required using a series of scripted questions. The *Medical Priority Dispatch System* (*MPDS*) (International Academies of Emergency Dispatch 2013) then generates a triage priority code for the call and available resources are allocated accordingly. These may include nearby fire department personnel able to provide basic life support (BLS), ground ambulance offering either BLS or advanced life support (ALS), and helicopter or fixed-wing air medical

Table 13.1 Service providers active in the care of acute major trauma

Lane	Agent	Principal action
Pre-hospital phase		
1	Patient	Sustains major injury
2	Bystander	Activates trauma care system
3	911 Operator	Configures combined emergency response
4	PTN call taker	Configures emergency medical response
5	EMS dispatch	Prioritizes and arranges medical transport
6	EMS providers	Field stabilization, triage, transport
7	Fire/police	Secures accident scene/extricates patient
Hospital phase		
8	Level III–V Emergency department	Provides initial medical care/prioritizes needs
9	Level III–V Operating room	Emergent or uncomplicated surgery
10	Level III–V Ward	Acute recovery and discharge preparation
11	Patient transfer network	Interfacility transfer of urgent/complex cases
12	Level I–II Emergency department	Reassessment, stabilization, triage
13	Level I–II Trauma service	Stabilization, prioritization, emergent care
14	Level I–II Specialty consultants	Urgent or complex specialized care
15	Level I–II Operating room	Stabilizing and/or reconstructive surgery
16	Level I–II Radiology	Diagnostic imaging and intervention
17	Level I–II Intensive care unit	Physiologic support of critically injured
18	Level I–II Intermediate care unit	Monitoring/high dependency care
19	Level I–II Allied health services	Multidisciplinary supportive care
20	Level I–II Ward	Acute recovery and discharge preparation
21	Level I–II Transitional services team	Post-hospital needs assessment and planning
Post-hospital phase		
22	Level I–II Outpatient clinics	Specialty-oriented follow-up
23	Community providers	Multidisciplinary care external to hospital
24	Long term care team	Extended care support external to hospital
25	Rehabilitation medicine	Assessment/focused rehabilitation plan
26	Rehabilitation services (public)	Basic rehabilitation support
27	Rehabilitation services (private)	Adjunctive rehabilitation services

evacuation offering the highest level of support through the accompanying critical care paramedics (CCP).

While largely predictable and amenable to algorithmic protocolization, medical and logistical considerations may require that triage and transport be tailored. A "layered" approach of overlapping services is common to avert unnecessary delays. This usually involves first medical response by BLS-trained fire crews followed by simultaneously activated ambulance.

The *British Columbia Ambulance Service* (*BCAS*) has also developed an *Autolaunch* scene response for certain regions. This is a dispatch protocol that, based on basic but critical information provided directly by bystanders at the scene (e.g., an accident with an unconscious injured person), simultaneously activates a dedicated air ambulance helicopter and a paired ground ambulance. Depending on logistics, these modalities may rendezvous at the scene, outside the closest medical facility, or somewhere in between. In remote regions where helicopter transport is

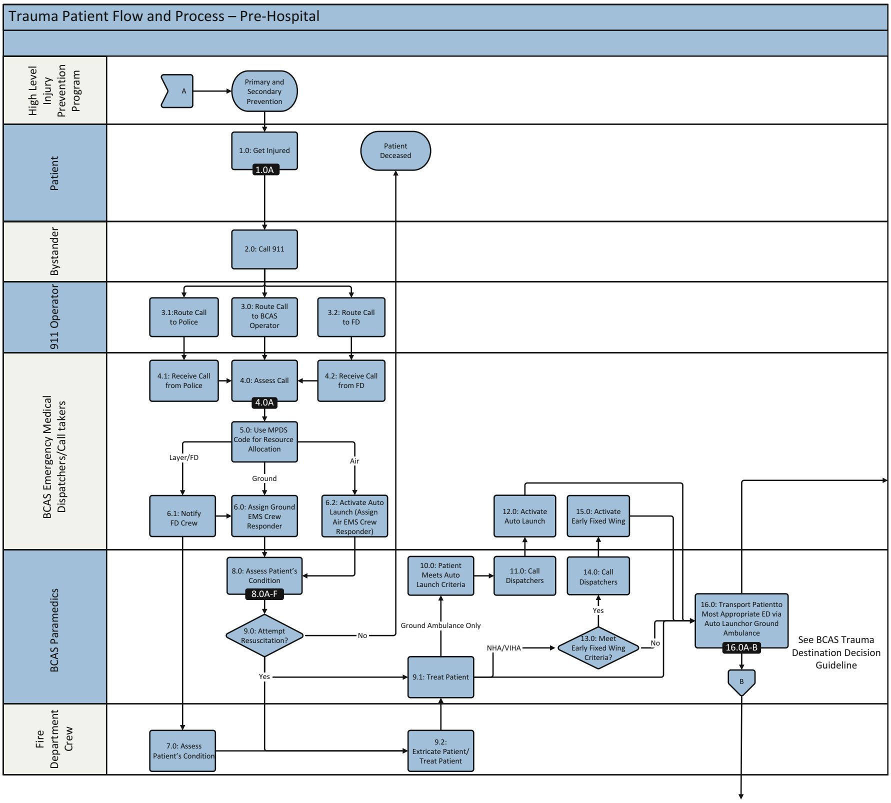

Trauma Patient Flow and Process – Pre-Hospital
High Level Injury Prevention Program
Patient
Bystander
911 Operator
BCAS Emergency Medical Dispatchers/Call takers
BCAS Paramedics
Fire Department Crew
A
Primary and Secondary Prevention
1.0: Get Injured
1.0A
Patient Deceased
2.0: Call 911
3.1:Route Call to Police
3.0: Route Call to BCAS Operator
3.2: Route Call to FD
4.1: Receive Call from Police
4.0: Assess Call
4.0A
4.2: Receive Call from FD
5.0: Use MPDS Code for Resource Allocation
Layer/FD
Ground
Air
6.1: Notify FD Crew
6.0: Assign Ground EMS Crew Responder
6.2: Activate Auto Launch (Assign Air EMS Crew Responder)
12.0: Activate Auto Launch
15.0: Activate Early Fixed Wing
8.0: Assess Patient's Condition
8.0A-F
10.0: Patient Meets Auto Launch Criteria
11.0: Call Dispatchers
14.0: Call Dispatchers
Yes
16.0: Transport Patientto Most Appropriate ED via Auto Launchor Ground Ambulance
16.0A-B
9.0: Attempt Resuscitation?
No
Ground Ambulance Only
NHA/VIHA
13.0: Meet Early Fixed Wing Criteria?
No
See BCAS Trauma Destination Decision Guideline
Yes
9.1: Treat Patient
B
7.0: Assess Patient's Condition
9.2: Extricate Patient/ Treat Patient

Fig. 13.3 Process map of patient flow: pre-hospital care

not possible (e.g., northern British Columbia and remote areas of Vancouver Island), a similar *Early Fixed-Wing Activation Protocol* may launch a medevac aircraft at the request of first response ground ambulance crews. Additionally, based on expanded medical criteria immediately discernible by medical staff, the arrival of a severely injured patient at the emergency department (ED) of a rural or remote facility who is clearly in need of transport to a higher level of care will trigger an *Expedited Transport* response. EMS dispatch activates the most appropriate transport plan before basic details of the case are assembled and communicated. If subsequent assessment finds that expedited transport is not required, the activation is terminated and the patient transfer conducted in standard fashion.

The most common method of directing severely injured patients requiring transport into the trauma system is via direct communication between a sending emergency physician and a receiving trauma team leader (usually a general surgeon with trauma subspecialty training), following initial assessment and stabilization at a designated trauma referral center. This dialogue is facilitated by a provincial *Patient Transfer Network* (*PTN*) that oversees all patient transports and transfers within the province and is accessed by a single contact number. The PTN conferences system call-takers, sending and receiving physicians, and a supervisory transport nurse, to prioritize and plan the patient transport. As required, EMS dispatch and/or an EMS transport advisor, as well as any required additional on-call medical specialists, may be brought in. To limit unnecessary discussion, standing inter-facility agreements dictate the destination of certain critical injuries. A *Life, Limb and Threatened Organ* (*LLTO*) policy invokes a no-refusal agreement for the emergent transfer of highest priority cases to designated trauma receiving facilities. Urgent but non-emergent cases requiring the full spectrum of expertise available at a higher level center may be designated *Higher Level of Care* (*HLOC*); these patients are also transferred according to standing inter-facility agreements. In this streamlined manner, triage and transport decision-making is optimized not only for individual patients but also for the system as a whole.

The processes of care applied by EMS responders follow basic algorithms driven by clinical signs and symptoms. While too nuanced to map in detail, these processes center on rudimentary airway, breathing, and circulatory support. In reality, a limited array of basic interventions is available to EMS crews. These include:

1. Maintenance of an open airway (using orotracheal intubation, if appropriate), application of supplemental oxygen.
2. Establishment of effective breathing (using bag-valve-mask assistance if necessary).
3. Control of hemorrhage using direct pressure or tourniquets.
4. Replacement of lost blood volume by establishing intravenous access and administering crystalloid solution.
5. Immobilization of the spinal column.
6. Splinting of extremity fractures.
7. Pharmacologic treatment of pain and agitation.
8. Prevention of hypothermia.

2.1.2 Contingencies and Variants

As not all remote regions have one-number public access to the emergency response system, alternative methods of system activation are sometimes necessary. In these instances, the local police or fire departments (FD) are usually involved initially and notify the EMS as required.

Air transport resources may not be directly available, so private air transport providers or military agencies may be called upon to initiate movement of the patient into the system. In instances where the needed response will be significantly delayed, the dispatcher may assist the patient indirectly through instructions guiding a third party to perform basic first aid care.

2.1.3 Information Flow

With activation of the 911 system, a flow of information begins that serves two purposes. The first is to document, communicate, and conserve pertinent medical information about actively managed patients for care providers. The second, for administrative purposes, is to inform the ongoing assessment of system performance through a series of targeted quality improvement strategies or sanctioned research.

The 911 emergency communications system, the PTN, and the EMS systems collect and store initial case-specific information. These systems collect both clinical data on patients gathered by paramedics (using handwritten forms in duplicate), and electronically gathered administrative data on transport logistics provided by call-takers and dispatch officers. Telephone conversations are recorded and may be transcribed for quality assurance purposes. Key EMS performance indicators include time to scene response, on-scene time, and completed transport time. *Autolaunch, Early Fixed-Wing Activation* and *Expedited Transport* transfers are noted for specific evaluation by EMS given the significant associated resource use implications. In addition to other administrative measures, the PTN similarly catalogs and reviews all LLTO transfers. Digital radiologic imaging is shared as part of the clinical record when images are mounted on a provincial grid connecting the *picture archiving and communication systems* (*PACS*) of enabled acute care facilities.

An electronic provincial database of drug prescriptions (*PharmaNet*) is maintained with individual patient records available to licensed treating physicians at any stage of care. This augments clarity around the chronic health issues of injured patients presenting initially to ED, and greatly enhances patient safety.

When patients are transported from the accident scene to a local facility for stabilization and then transferred on to definitive care at a trauma referral center, the clinical information documented by nurses and physicians is handwritten on standardized forms, copied, and forwarded with transferring patients. There is thus great variability in the quantity and quality of information transmitted and available for *post hoc* performance improvement work.

2.1.4 Performance Evaluation

The primary quality indicators for pre-hospital systems are scene response time and total transport time. The most valuable indicator is the rate of under-triage expressed as the proportion of patients transported to an inappropriate low level trauma facility. Although under-triage can be inferred using trauma registry data, which incorporates basic EMS transport information, it may be best gauged through prospective adjudication by receiving physicians at the time of EMS transport. The appropriateness of expedited (*Autolaunch* and *Early Fixed-Wing Activation*) and protocol-driven (LLTO and HLOC) transports using data from EMS, communications systems, and trauma registries is an important measure of efficient stewardship of public resources. The elaboration of trauma systems is predicated on the “golden hour” concept of getting the right patient to the right place at the right time (American College of Surgeons Committee on Trauma 2006). Surprisingly, there has been little hard evidence that optimized time to care, a quality measure used

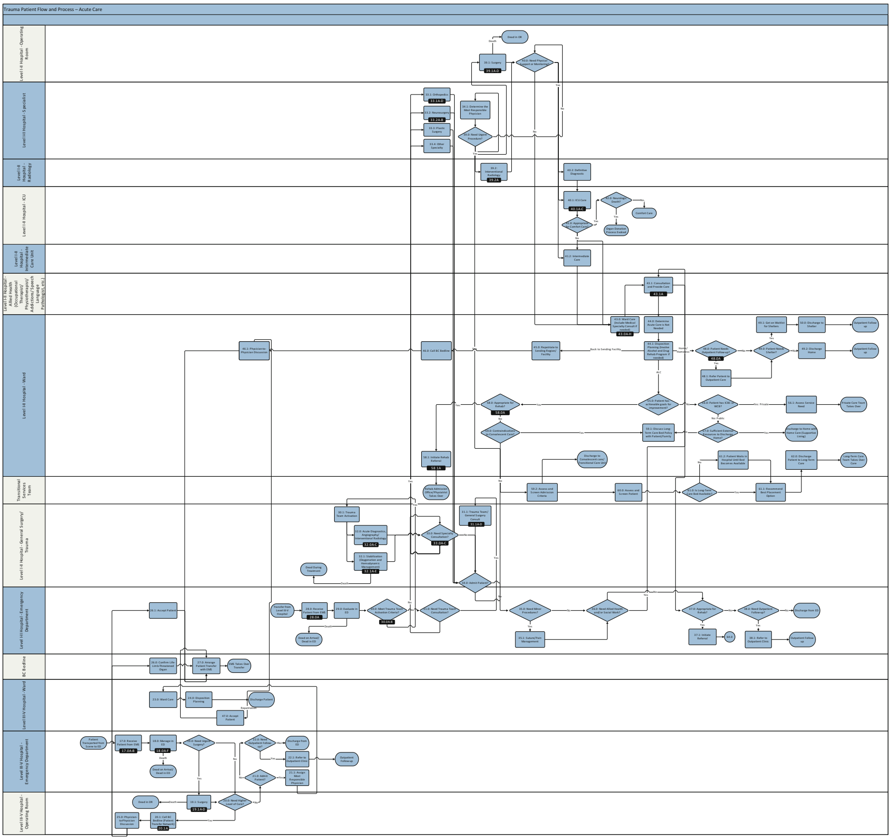

Trauma Patient Flow and Process – Acute Care
Level I-II Hospital - Operating Room
Level I-II Hospital - Specialist
Level I-II Hospital - Radiology
Level I-II Hospital - ICU
Level I-II Hospital - Intermediate Care Unit
Level I-II Hospital - Allied Health (Occupational Therapist/ Physiotherapist/ Addictions/ Speech Language Pathologist, etc.)
Level I-II Hospital - Ward
Transitional Services Team
Level I-II Hospital - General Surgery/ Trauma
Level I-II Hospital - Emergency Department
BC Bedline
Level III-V Hospital - Ward
Level III-V Hospital - Emergency Department
Level III-V Hospital - Operating Room

Fig. 13.4 Process map of patient flow: hospital care

extensively in trauma system performance improvement programs, is associated with improved outcomes (Stelfox et al. 2010; Di Bartolomeo et al. 2008).

2.2 *Hospital Care*

2.2.1 Patient Management

Hospital-based acute care of the severely injured is necessarily less standardized than pre-hospital EMS care. Nonetheless, as depicted in Fig. 13.4, we have reduced the principal acute care processes to a series of steps through consultation with key informants.

Hospitals within a regional system are designated to receive and manage trauma according to capacity. As per 2010 *Trauma Association of Canada* (*TAC*) accreditation guidelines (Trauma Association of Canada/Association canadienne de traumatologie 2007), Level I (university-affiliated) and Level II (non-university affiliated) centers offer comprehensive and definitive trauma care at the highest level. Level I centers also usually provide supra-specialized care for entities such as spinal cord injury, major burns, pediatric trauma, and complex orthopedics and plastic surgery (i.e., limb reimplantation). Level III centers are large general community hospitals that provide a full spectrum of basic trauma care and serve to offload routine cases from Level I centers. Level IV and V centers may manage minor trauma but have no capacity to handle major trauma beyond life-saving stabilization and referral. Verification that trauma receiving facilities meet appropriate standards is carried out by a recognized authority—in Canada, the *Trauma Association of Canada* in partnership with *Accreditation Canada*—through the periodic process of accreditation. In an inclusive trauma system, such as that in BC, all acute care facilities have a designated role in the provincial trauma system. In exclusive systems, selected centers are bypassed in favor of direct transport to a designated trauma-receiving center.

As up to two-thirds of referrals to Level I and II regional trauma centers are indirect transfers through Level III, IV, and V centers, it is appropriate to diagram care in these facilities. In all centers, the first step is assessment of transported trauma patients by a triage nurse using a tool such as the *Canadian Triage and Acuity Scale* (*CTAS*) (Bullard et al. 2008) to assign priority status. Patents are then registered into the ED information system and seen by a physician with appropriate urgency. Major trauma is assessed as CTAS 1 and considered an emergent priority.

Level IV and V Trauma Centers

The medical management of major trauma in North America generally follows clinical guidelines set out by the *Advanced Trauma Life Support* (*ATLS*) course of the *American College of Surgeons Committee on Trauma* (2012), which promulgates a standardized, priority-based approach to the diagnosis and treatment of clinical findings in the severely injured patient. As the details of this care are complex and not appropriately described at the higher altitude of our system-level process map, we have labeled these processes in aggregate as *ED management*. The critical decision-making that emerges from these care processes determines the next steps in the patient flow.

Level V centers are generally remote clinic-type facilities that regularly receive and rapidly transfer major trauma. Level IV centers are larger urban centers with active EDs that may handle a large volume of minor trauma but are bypassed by major trauma triaged to nearby Level I or II centers, except when they are the closest facility for a major trauma patient *in extremis*.

Patient flow is governed by simple disposition-oriented decision-making. Firstly, appropriate recognition that patients are deceased (not always straight forward with brain death), or highly unlikely to survive despite best efforts, terminates the patient trajectory immediately, which is essential if valuable resources are not to be wasted on the protracted provision of hopeless care. In most cases, a decision must be made about stability for transport to a higher level of care. While unnecessary procedures delaying definitive care are discouraged in lower level facilities, some patients do periodically require emergent intervention prior to transport onward from the first facility. This challenging decision is often shared by sending and receiving physicians, in conference through the PTN. Usual examples include procedural intervention in the ED (e.g., intubation and mechanical ventilation, circulatory support with blood product replacement, pharmacologic management, or other physiologic stabilization). Basic life or limb-saving surgical procedures within the capabilities of local surgeons may also be appropriate. These include damage control surgery for abdominal bleeding, decompression of critical traumatic brain hemorrhage, and stabilization of limb-threatening fractures, all of which can be carried out by well-trained local general surgeons. Both Level IV and V hospitals make referrals to outpatient clinics, assisted living, home care, convalescent care, long-term care, or rehabilitation facilities as needed.

Level III Trauma Centers

Level III trauma centers triage, register, assess, and stabilize presenting trauma in similar fashion to Level IV centers. Because Level III centers are designated to provide definitive care for major multiple trauma of moderate complexity, the key initial decision-making focuses on whether to retain or transfer patients onward.

Transfer to a Level I or II regional center is indicated when local resource capacity is exceeded. Patients may be transferred emergently if assigned LLTO status, or urgently (but non-emergently) if assigned HLOC status. Transfer is organized according to *trauma destination decision guidelines* modelled after World Health Organization (WHO) recommendations (Sasser et al. 2012), by contacting the PTN which brings in necessary medical and paramedical personnel as previously described. If transfers occur within 48 h of injury, patients are transported by appropriate level EMS staff directly to the ED of the receiving hospital, bypassing normal requirements that a bed in an appropriate unit (usually intensive care) first be identified and available. Emergent life and limb-saving ED interventions or formal operative procedures may be necessary to stabilize patents prior to transfer. Patients deemed appropriate for local management undergo a complete diagnostic evaluation in the ED and receive required care from required consultant specialists as prioritized by the TTL. Occasional patients found to have only minor injuries are discharged home if possible. If the patient requires admission, a most responsible physician (MRP) is assigned and the patient is admitted to an appropriate unit (intensive care, intermediate care, ward). As patients improve, they progress to lower levels of care, ultimately needing only ward care. On the ward, staff provides care and undertakes disposition planning that may require referrals to assisted living, home care, convalescent care, long-term care, or rehabilitation. When sufficiently recovered, as judged by a range of medical and allied health staff (e.g., physiotherapy, occupational therapy, social work, pain specialists), the most responsible physician discharges the patient to an appropriate disposition and outpatient follow-up appointments with required specialists are arranged, if required.

Level I–II Trauma Centers

Emergency Department (ED)

Trauma patients are transported to Level I and II centers by air or by ground EMS crews, either directly from the event scene (roughly 65 %), or indirectly from a lower level center after sending and receiving physicians agree on the transfer of care responsibility (35 %) (Vancouver Coastal Health 2010). Patients are triaged, registered, and evaluated in the ED by a designated ED physician. If the patient is found to meet set criteria on the basis of physiologic (e.g., respiratory distress, shock, or altered level of consciousness), anatomic (e.g., multiple long bone fractures, unstable pelvis, flail chest), mechanistic (e.g., high-speed motor vehicle collision, fall from >20 ft, stab or gunshot wound), or administrative (e.g., multiple casualties, ED staff unable to leave other patients) criteria, a Trauma Team Activation (TTA) is called. The operating room (OR) and blood bank are alerted and an in-house team of physicians, nurses, and support staff, directed by an assigned Trauma Team Leader (TTL), assemble immediately to assume care of the patient. When patients *in extremis* meet pre-activation criteria in the field, EMS crews alert the receiving center to activate the Trauma Team prior to the arrival of the patient.

Once it is determined that the patient is not dead on arrival (DOA), the initial focus of the trauma team is physiologic resuscitation (adequate oxygenation and hemodynamic stability). This occurs in parallel with prioritized diagnostic evaluation to further confirm suspected injuries. Level I and II centers provide specialized radiologic techniques including advanced computed tomography (CT) imaging, diagnostic and therapeutic interventional radiology (e.g., angioembolization to control surgically inaccessible hemorrhage), and magnetic resonance imaging (MRI) to assess complex neurologic trauma.

If patients do not meet TTA criteria, the ED physician completes an appropriate evaluation and either discharges the patient or refers to a consultant specialist, usually a general surgeon specialized in trauma care for polytrauma patients or a surgical subspecialist in the case of single system injury. A nonsurgical physician, often a hospitalist, may be requested to assess and, if indicated, admit stable patients with problematic nonmedical impediments to discharge.

From the ED, a major trauma patient will be directed to one of (1) an OR for urgent surgical management; (2) an intensive care or intermediate care unit for monitoring, physiologic support, further treatment and diagnostics; or (3) a ward bed.

Occasionally, patients will die in the ED during management (the designation DIE is applied), either before stabilization is possible, or, in hopeless situations, after a deliberate decision is made with family or others to palliate.

Operating Room (OR)

Transfer to the OR from the ED, or later from ICU, is common for major trauma. Level I and II centers generally maintain an OR on standby for emergent (stat) trauma cases in need of immediate operative intervention. A specially outfitted operating theater up-sized to facilitate multiple teams simultaneously, close to the central nursing station and with direct tube access to blood bank, is common. A strategy for ensuring immediate access to surgery with anesthesia and nursing teams promptly available is essential. Several operative procedures completed in priority sequence over days to weeks by a variety of surgical specialists may be required. Notably, these include general surgery, orthopedic and spine surgery, neurosurgery, and plastic surgery. Stabilizing procedures are performed early on, followed by definitive repair procedures.

Intensive Care Unit (ICU)

The majority of major trauma patients will require a period of mechanical ventilation and other invasive physiologic support in an ICU. Need for acute brain injury care, stabilization after massive volume replacement, and renal replacement therapy (dialysis or hemofiltration) are other common reasons for ICU admission. A dedicated critical care physician will direct management in collaboration with numerous surgical and medical specialists while critical injuries and their complications continue to be diagnosed and treated. Because ICU services are in constant high demand in the tertiary facilities that serve as Level I and II trauma centers, a no-refusal policy for ICU access for trauma is critical for unimpeded flow of patients and effective function of the trauma system.

Much of the care in ICU is standardized based on best practices established by evidence from clinical research, and locally adapted protocols guide many aspects of the medical care. Most patients ultimately stabilize, wean from invasive support, and discharge to an intermediate care unit or ward where responsibility for care is transferred to a designated physician, usually a general surgeon. Non-survivors may die of their injuries or related complications in ICU, or be switched to comfort care and permitted to die on the basis of poor prognosis for meaningful recovery. Confirmed brain death may result in organ donation after assessment and intervention by a transplant coordinator and organ retrieval team.

Intermediate Care Unit

Patients not needing full ICU support, but still needing continuous monitoring for high acuity (e.g., blood pressure, oxygenation, altered level of consciousness, neurovascular status) or high-dependency care for tenuous breathing or complex wounds, are admitted to an intermediate care unit. Medical oversight is provided by the MRP. Disposition of surviving patients is usually to a clinical ward, but may also be repatriation to the initial sending facility, or occasionally discharge to home or a convalescent facility.

Hospital Ward

Once stabilized and no longer requiring high-acuity or high-dependency care, patients are transferred to a clinical unit, usually a general surgery, orthopedic, spine, or neurosurgical ward. Medical direction is provided by the MRP in collaboration with consultant specialists and allied health staff (occupational and physical therapy, speech-language pathology therapy, social services, and nutritionists). Basic care, early rehabilitation therapy, patient education, and discharge planning are the main ward activities. Drug and alcohol addictions are flagged during disposition planning, with referral to drug and alcohol rehabilitation support programs as appropriate.

If patients are not repatriated at this juncture to the initial sending facility, through the PTN, in accordance with existing agreements (48 h for routine repatriation), the primary goal is for them to return home, with or without assistance for activities of daily living. Transfer to an inpatient rehabilitation facility is appropriate for directable patients with recoverable deficits, once assessed and approved by a consulting physiatrist. Many patients will discharge to a convalescent facility, either in preparation for return home, or in anticipation of eligibility for inpatient rehabilitation. Outpatient clinic appointments and rehabilitation services are arranged prior to discharge. Sometimes, patients will be ineligible for disposition in any of these categories and will need to remain in an acute care setting as alternate level of care (ALC) patients until a solution can be innovated, a process which can take months. Frequently, these are unrehabilitatable brain-injured patients, or the elderly whose injuries have created a transitional state of decompensation.

2.2.2 Contingencies and Variants

Despite a formal no-refusal policy for ICU admission implemented uniquely for major trauma, physical space may not be available when stabilized patients are ready for transfer from the ED. In these instances, patients are managed temporarily in a resuscitation bed in the ED by intensive care physicians aided by ED nursing staff. Similarly, patients who have undergone urgent surgery may need to be maintained in recovery room awaiting ICU bed availability; here again care will be directed by ICU staff "off-site" in transfer from the treating anesthesiologist.

Lack of bed space, usually a consequence of nursing staff limitations resulting from budgetary or manpower constraints, also frequently impedes patient flow to an intermediate care unit or clinical ward. Beyond this, bed availability may similarly limit flow in discharging patients to a repatriating hospital, convalescent facility, or inpatient rehabilitation unit.

A full range of allied health team assessments are required prior to movement of patients, particularly at discharge, to ensure that wound care, mobilization, toileting, pain control, nutrition, and other basic functions can be safely managed at the next level. Many of the required personnel are not available after hours and on weekends, further constraining the efficiency of patient flow. Dedicated patient care managers (PCMs) and a transitional services team (TST) work to coordinate and streamline the required care needed for effective discharge in an effort to optimize patient flow.

2.2.3 Information Flow

The principal repository of information is the medical record (hospital chart), which follows patients as they move through different phases of care in a given hospital. Most hospitals are transitioning in stages to fully electronic medical records to enhance information clarity and flow. At discharge, the patient record is verified and maintained electronically in the Canadian Institute for Health Information *Discharge Abstract Database* (*DAD*). A "tertiary survey" is completed by trauma service physicians to verify and document diagnoses and treatments prior to discharge, adding layers of verification by knowledgeable clinicians that greatly enhances the administrative and clinical utility of compiled data. Formatted transfer notes and discharge summaries communicate this information to subsequent caregivers.

Trauma registries are established in Level I, II and III hospitals to abstract information on the majority of acute trauma patients; these data are used for performance improvement purposes. Criteria for entry into the regional trauma registry includes death in hospital, admission for more than 2 days, or an Injury Severity Score (ISS) (Baker et al. 1974) greater than 16. These criteria vary somewhat between registries, and patients treated at more than one hospital will be represented by multiple registry entries. Data abstraction and cleaning by registry staff is resource intensive, but essential for measuring outcomes of interest and being able to risk-adjust outcomes using standard procedures. The principal outcome of interest is hospital mortality, risk-adjusted using probability of death based on the constellation of major injuries represented by the ISS. Cleaned facility-level registry data are directly available to submitting centers for quality assurance. It is also aggregated at the provincial level to describe major injury burden, and this is, in turn, is submitted as a comprehensive dataset to DAD.

Patients assessed in ED but not admitted to hospital will not be recorded in the DAD or trauma registry unless they die. An abundance of lesser injured patients able to discharge home are thus not captured. The *National Ambulatory Care Reporting System* (*NACRS*) captures information on all ED patients from the ED record, but is not linked to the DAD or trauma registry. Similarly, most ICUs and

Table 13.2 Key acute trauma care process metrics and data sources

Measure	Data source
Pre-hospital phase	
Scene response time	EMS record
Scene time	EMS record
Failed airway management rate	EMS record
Transport time	EMS record
Presence of EMS record in hospital chart	Hospital record
Under-triage rate	Trauma registry
Hospital phase	
Trauma team activation rate	Hospital record/trauma registry
Trauma team leader response time	Hospital record/trauma registry
Time to CT scan for major head injury	Hospital record/trauma registry
Time to surgery for control of hemorrhage	Hospital record/trauma registry
Time in ED	Hospital record/trauma registry
Venothromboembolism prophylaxis rate	Hospital record/trauma registry
Complication rate	Hospital record/trauma registry
Completion of tertiary trauma survey	Hospital record/trauma registry
Hospital length of stay	Hospital record/trauma registry
Hospital mortality (ISS adjusted)	Hospital record/trauma registry
Preventable death	Mortality review proceedings
Post-hospital phase	
Head injury referral to rehabilitation	Hospital record
Wait time to referral for rehabilitation	Hospital record

some specialty services (e.g., those treating spinal cord injury, major burns, and hip fractures) maintain dedicated registries of clinical and administrative information that is separate from, and unlinked to, searchable hospital data.

As BC has universal publicly funded health care, costs for all insured services are captured and available for system-level evaluation.

2.2.4 Performance Evaluation

The most common measure of performance at the hospital level is ISS-adjusted hospital survival. As partially listed in Table 13.2, a large array of quality indicators have been developed for performance improvement programs, many of which are endorsed by verification/accreditation authorities. There exists, however, surprisingly little evidence that validates any of these as useful measures of processes of care leading to improved survival or other desired system-level outcomes.

Trauma registries represent a considerable investment of resources. They exist largely to support basic evaluation of the design and function of systems of care for major trauma, predicated on maintaining rapid flow of patients through systems of care. Registries are heavily focused on providing accurate data for calculating risk-adjusted survival, and also document complications of care (not easily distinguishable from complications of injury), although benchmarks permitting actionable interpretation of findings are still needed.

2.3 Post-hospital Care

2.3.1 Patient Management

Rehabilitation Network

When hospitalized patients have resolved active medical issues, a physiatrist is consulted to assess whether they are sufficiently directable to participate in therapies to improve functional disabilities. If considered eligible for rehabilitation at an inpatient unit, an inter-facility transfer is organized. Physical, emotional, social, and spiritual skills needed for independent reintegration into community living are addressed. For patients admitted for inpatient rehabilitation, a streamlined high-intensity rehabilitative track may be pursued for uncomplicated musculoskeletal injuries, while specialized programs exist for acute brain injury and limb amputees. In all cases, the goal is to achieve sufficient physical, communicative, and cognitive function to enable continued rehabilitation using outpatient resources. When patients cannot reach targeted performance goals, discharge home with support or to an extended care facility is arranged. Patients may be readmitted to an acute care facility if new active medical issues arise.

In the community, publicly supported services such as clinic, community, or home support are available to all appropriate patients for lower intensity therapy when patients don't have access to insured services compensation. A slow stream rehabilitation course is pursued for patients with limited goals (e.g., the elderly). This course focuses on community and home support rather than outpatient assistance, prioritizing activities of daily living over building higher levels of functional performance.

When a patient's injury is related to employment or involvement of a motor vehicle, insurance is available through public agencies—*WorkSafe BC* and the *Insurance Corporation of British Columbia* (*ICBC*), respectively—to further assist rehabilitation and compensate material losses. The additional resources afforded by these means often enable patients to return home with non-publicly funded, and otherwise unavailable, support including care attendants, adaptive equipment, and supplemental therapies. Patients with independent financial resources may privately secure similar resources to facilitate recovery and hasten autonomous living.

Other Dispositions

Patients not transferring to inpatient rehabilitation discharge from acute care to a variety of dispositions, including home (with limited assistance if needed), convalescent hospitals, assisted living facilities, or extended care facilities that provide support for basic self-care for the severely disabled. Convalescent care is a short-term residential service, available at minimal cost, which requires identification of achievable goals (e.g., mobilization, pain control) for patients needing additional recovery time, typically ranging from 1 to 7 weeks. Homeless patients are waitlisted for shelter placement and discharged once space is available.

Outpatient Care

Discharged patients are referred to external clinics for follow-up by surgical and other specialists who provided care during hospitalization. Outpatient rehabilitation is variably organized. Limited treatment by physical and occupational therapists, dieticians, social workers, etc. occurs both in the home setting and in external clinics as needed. Patients discharged to distant regions of the health authority, or to another health authority, have most follow-up arranged locally, but do return, often travelling great distances, for essential specialized care and assessment. Most patients have a family physician who takes over responsibility for orchestrating care in the patient's local community. Patients without a family physician are scheduled to be seen in public clinics where ownership of the post-hospital processes of care is more problematic.

2.3.2 Contingencies and Variants

Effective discharge planning can be complex and time-consuming, requiring multidisciplinary input from many services. At times, patients who have reached a stable state but remain debilitated, are not dischargeable from acute care. They are classified as ALC patients with no active medical issues and are seen in limited fashion by hospital physicians while awaiting discharge to an appropriate setting. Family capacity to provide for patients' needs is a major factor. These patients can remain in an acute care bed for many months. Patients injured while visiting from out-of-province or foreign countries require repatriation, which is greatly assisted when applicable insurance coverage applies. Hospitals typically have to absorb the full cost of repatriation for uninsured patients who are unable to pay.

2.3.3 Information Flow

Patients discharged within the health authority are generally seen in clinics directly linked to a trauma center, making information flow efficient and enabling access to imaging. There is limited electronic information linkage between BC's five regional health authorities, posing challenges for post-hospital management. Functional assessment of resolving disability is most comprehensively documented by rehabilitation specialists, particularly in the inpatient setting, using standardized clinical instruments such as the *Functional Independence Measure* (*FIM*) (Keith et al. 1987), and various other metrics. Limited data on the relatively small proportion of patients receiving inpatient rehabilitation, notably brain and spinal cord injuries, are fed to the National Rehabilitation Reporting System (NRS), predominantly for research purposes.

Surgical specialists convey practical information to insurers and employers on physical suitability for ongoing rehabilitation from external clinics and private offices. Information on return-to-work capacity and suitably staged resumption of employment activities is generally directed by the family physician, with input from specialists and therapists. Extensive data are collected by public insurers, namely the provincial workers' compensation agency and automobile insurance bureau. These data inform compensation programs and direct provincial level planning to reduce injury-related death, disability, and associated burden of injury costs.

2.3.4 Performance Evaluation

Quality assurance activity in the post-hospital phase is limited to administrative monitoring of resource utilization (primarily inpatient rehabilitation length of stay), and patient disposition. There are no established performance thresholds and no known firmly validated standards by which to judge processes of care or outcomes. Time to initiation of rehabilitation programs known to be effective in improving functional recovery after injury is a commonly cited metric, but this applies only to the relatively small subset of patients with major injury who are eligible for institutional rehabilitation.

As yet, there is little direct linkage between the system providing clinical care and the nonclinical agencies that likely have important influence over high-level drivers of injury control. For instance, extensive data are collected on injury related to employment (*WorkSafe BC*) and automobile collision (*ICBC*), yet system-level performance evaluation within these domains is largely limited to cost of injury, gauged by claims paid and care-related expenses for eligible patients. Disability, quality of life, employment status, and economic consequences of injury are not comprehensively measured outcomes in the post-acute phase of care. Routine measurement of these system outputs would seem important to gauging the effectiveness and value of injury management processes.

3 Practical Considerations and Limitations

While the practical activities of care for severely injured patients lend themselves to process mapping, there are practical considerations that must be acknowledged. The *functional* method we use to describe care processes in this process map organizes a predetermined set of clinical and managerial tasks sequentially such that inputs are transformed into outputs for achieving clearly defined outcomes. As our mapping illustrates, the principally reported outcome of the processes of care for major trauma is severity-adjusted hospital survival. Any patient that discharges alive from acute care is therefore considered an optimal outcome, and this may not truly be the case. From a public health perspective, reduced incidence of injury, improved functional outcomes, quality of life, or diminished aggregate cost of injury (societal burden of injury), may be more appropriate outcomes toward which to tailor system processes. Stakeholder consensus prioritizing clearly defined system objectives is essential to developing a useful system analysis tool such as process mapping.

Additionally, while the functional approach helps to identify key processes, define desired outcomes, and describe the structures required to support care delivery, it may not optimally account for the reactive nature of emergency care as applies in trauma systems (Sobolev et al. 2012). Specifically, the inherent variability in patients and providers, the competition for resources from elsewhere within the health care system, and the practical difficulties characterizing (and measuring) processes of trauma care all limit the predictability of outcomes of care. As a complement to this functional approach, a *behavioral* approach to system analysis would focus on the conditions and events that trigger activities and the transitions between them. This approach might reflect more directly the hierarchy of activities, as well as the interactions of concurrent events, that occur in complex systems such as those developed for injury management. A behavioral paradigm might, for example, place more emphasis on processes outside of the care domain that lead to regulatory change or built environment modification aimed at reducing injury incidence and/or severity (*secondary* injury prevention). A growing interest in the application of a public health framework, emphasizing population-based injury management over patient-based trauma care, supports the need for a broader and more nuanced strategy of system management. In the future, it is conceivable that processes that lead to injury prevention through regulation, legislation, education, or research may become better defined and governable by health system administrators. Expanding the mapping of injury management processes into these less well-delineated domains would be useful. The end result of this more expansive mapping might reveal that the societal burden of injury is more effectively impacted by investment in secondary prevention, rather than further investment in improvements to processes of care.

Figure 13.1 shows where the acute care processes for the management of major trauma (green box) act within a larger system of injury care, control and prevention. It seems likely that a combined functional and behavioral approach will ultimately afford the most realistic means of characterizing trauma systems such that decision-makers are better enabled to devise and implement strategies that achieve desired outcomes.

4 Summary

Modern trauma systems have been organized to focus almost exclusively on acute care processes. By illustrating a conventional trauma system's operations through the care continuum, process mapping provides an effective tool that can enable decision-makers to more readily comprehend the system for which they are responsible, and the hierarchy of supportable processes at work therein. Once developed, such a map and accompanying narrative clarifies information flow and provides an objective starting point for developing practical process and outcome measures. These measures—if reliable, feasible, and practicable—will support the testing of actionable hypotheses about where the system, as a series of processes, succeeds and fails in achieving desired outcomes. With a clear set of agreed upon system objectives and priorities, the more granular understanding provided by process mapping supports improved understanding of delivered health services for trauma, and thereby supports both smarter system design and management, and more accountable and efficient use of public funds.

Acknowledgements The authors would like to acknowledge Hugh Anton, MD, Division of Physical and Rehabilitation Medicine, University of British Columbia; Julie Wei, Manager, Quality Analytics, BC Emergency Health Services; and Ellen Randall, Research Project Manager, Vancouver General Hospital Trauma Program, for significant contributions to the work presented here.

References

American College of Surgeons Committee on Trauma. (2006). *Resources for optimal care of the injured patient, 2006*. Chicago, IL: American College of Surgeons.

American College of Surgeons Committee on Trauma. (2012). *ATLS® for doctors student manual*. Chicago, IL: American College of Surgeons.

Baker, S. P., Long, W. B., Haddon, W., & O'Neill, B. (1974). The injury severity score: A method for describing patients with multiple injuries and evaluating emergency care. *Journal of Trauma, 14*(3), 187–196.

Bullard, M. J., Unger, B., Spence, J., Grafstein, E., & Group CNW. (2008). Revisions to the Canadian Emergency Department Triage and Acuity Scale (CTAS) adult guidelines. *Canadian Journal of Emergency Medicine, 10*(2), 136–151.

Di Bartolomeo, S., Valent, F., Sanson, G., Nardi, G., Gambale, G., & Barbone, F. (2008). Are the ASCOT filters associated with outcome? Examining morbidity and mortality in a European setting. *Injury, 39*(9), 1001–1006. doi:10.1016/j.injury.2008.04.009.

Evans, D. C. (2007). From trauma care to injury control: A people's history of the evolution of trauma systems in Canada. *Canadian Journal of Surgery, 50*(5), 364–369.

International Academies of Emergency Dispatch. (2013). *The emergency priority dispatch systems*. Retrieved July 31, 2013, from http://www.emergencydispatch.org/ResourcesEDS

Keith, R. A., Granger, C. V., Hamilton, B. B., & Sherwin, F. S. (1987). The functional independence measure: A new tool for rehabilitation. *Advances in Clinical Rehabilitation, 1*, 6–18.

Sasser, S. M., Hunt, R. C., Sullivent, E. E., Wald, M. M., Mitchko, J., Jurkovich, G. J., et al. (2012). Guidelines for field triage of injured patients. Recommendations of the National Expert Panel on Field Triage, 2011. *Morbidity and Mortality Weekly Report, 61*(RR01), 1–20.

Sobolev, B., Sánchez, V., & Kuramoto, L. (2012). *Health care evaluation using computer: Simulation concepts, methods, and applications* (p. 417). New York: Springer.

Stelfox, H. T., Bobranska-Artiuch, B., Nathens, A., & Straus, S. E. (2010). Quality indicators for evaluating trauma care: A scoping review. *Archives of Surgery, 145*(3), 286–295.

Trauma Association of Canada/Association canadienne de traumatologie. (2007). *Trauma system accreditation guidelines,* Fourth Revision. Retrieved July 31, 2013, from http://www.traumacanada.ca/accreditation_committee/Accreditation_Guidelines_2011.pdf

U.S. Department of Health and Human Services. (2006). *Model trauma system planning and evaluation*. U.S. Department of Health and Human Services—Health Resources and Services Administration. Retrieved July 31, 2013, from http://www.ncdhhs.gov/dhsr/ems/trauma/pdf/hrsatraumamodel.pdf

Vancouver Coastal Health. (2010). *Regional trauma annual report 2009–2010*. Retrieved August 1, 2013, from http://www.vch.ca/media/RegionalTraumaAnnualReport09-10Final.pdf

Chapter 14
Forecasting Demand for Regional Health Care

Peter Congdon

Abstract Trends in developed nations point to increased demand for acute inpatient care across most age groups, though especially at older ages. Demand growth has also been differentiated by specialty, with evidence of a major rise in demand for medical rather than surgical hospital care. This growth has occurred despite new emphasis on siting appropriate care in primary and community settings. Building on existing work relating to spatial perspectives on changing health demand, the present analysis develops a Bayesian approach to modelling the generation of health demand in a region and its allocation to health providers. At the first stage projections of acute demand are made by specialty, patient age and area of residence, with the allocation to providers then determined by a gravity model. A case study considers trends in health demand during the 1990s as a basis for forecasting during 2000–2010. The study region comprises North East London and South Essex. In this application, projections of specialty referral rates (usage rates) by age are based on national data and are applied to regional population projections to give a forecast health demand. The allocation of this demand to hospitals then takes account of projected changes in the configuration of hospital beds in the region.

Keywords Health demand • Gravity model

1 Introduction

There is growing awareness of the impact of changing age structures on demand for health care. In terms of demand for hospital care (i.e. the acute health sector) there is a pronounced age gradient of usage in many health specialties. With increasing

P. Congdon (✉)
Department of Geography, Queen Mary and Westfield College, Mile End Road,
London El 4NS, UK
e-mail: p.congdon@qmul.ac.uk

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_14, © Springer Science+Business Media New York 2013

proportions of the very old in the total population as life expectancy extends, this raises issues about growing morbidity and hence health demand from this age group (Kane 1994; Metz 1999).

Combined with population ageing are trends in the demand for health care. National evidence in the UK and other developed nations points to increased demand for acute inpatient care (via inpatient referrals) across most age groups, though especially at older ages. Demand growth for hospital care has also been differentiated by specialty and method of admission, with evidence of a major rise in demand for medical rather than surgical care, in part through a rise in emergency admissions (Hull et al. 1997; Puig-Junoy et al. 1998). Growth in acute demand has occurred despite the potential for compressing morbidity, namely, reducing cumulative lifetime morbidity through success in lifestyle risk prevention, so postponing the age of onset of morbidity (Fries et al. 1998).

Demand for hospital care will also be affected by the performance of health care systems and success or otherwise of strategies regarding the appropriate balance of care between community and acute settings. The Department of Health in the UK undertook a Hospital Beds Inquiry to investigate future bed needs in acute hospitals (Department of Health 2000). This inquiry included in its remit the impact of demographic change and policy initiatives on demand as well as reviewing changing patterns of bed usage in hospitals in terms of reduced lengths of stay and increased day case rates. Day cases may involve temporary use of hospital beds but not overnight stays and have been increasing as a proportion of all hospital admissions (Sibbritt 1992).

Policy initiatives to manage demand and plan capacity are aimed at containing the increase in hospital admissions. They include targets to increase day case rates and provide community-based rehabilitation and intermediate care in residential and nursing home settings (so acting against “bed blocking” by certain patient categories) and efforts to boost primary care in order to reduce avoidable admissions, especially those classified as emergencies. Strategies within hospital include improved bed management and reductions in length of inpatient stay. These initiatives are set against longer term strategic objectives to raise health expectancies by preventive measures and encouragement of healthier lifestyles.

Against such a background of demographic and service provision change, the present analysis considers projections of demand for hospital care in 2005–2006 in a study region covering two million people in North East London and parts of the adjacent county of Essex. The study region includes socially deprived parts of inner East London (the East London and City Health Authority), suburban and rural South Essex and four suburban boroughs in outer North East London (Redbridge, Waltham Forest, Barking and Havering). Altogether there are 13 local authority areas considered, with the small City of London area merged with the London Borough of Tower Hamlets. There are ten hospitals (“providers”) considered in the analysis, accounting for over 90 % of the regions’ hospital referrals (Fig. 14.1). The hospital sites considered are general or district hospitals including accident and emergency facilities (i.e. special hospitals, such as psychiatric hospitals, are not considered).

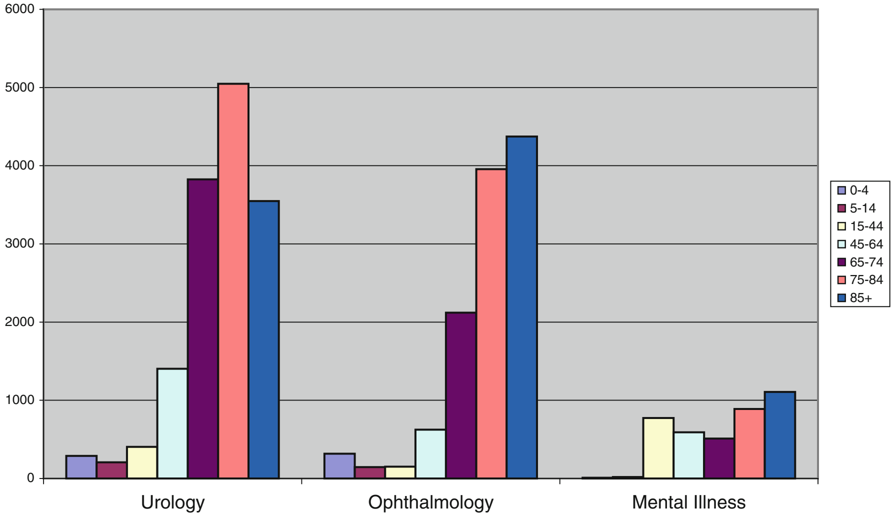


Fig. 14.1 Referral rates in selected specialties

2 Modelling Issues

The analysis focuses especially on the impact of changing population totals and age compositions and of changes in referral rates (i.e. health demand) on levels of acute referrals by specialty and provider within the context of the publicly funded UK National Health Service.

Spatial perspectives on changing health demand have been discussed before (Clarke and Wilson 1984; Lowe and Sen 1996). The modelling of future demand and its impact on acute care provision is complicated by the interplay of factors involved. These include demographic change in different residential areas (e.g. changing age and ethnic structure), changes in demand for different categories of acute care (e.g. specialties) and blurred interface between acute and community care for many chronic conditions (for example, hospital admissions may be avoidable given suitable community care and case management), interaction between residence area and hospital location (e.g. in terms of travel mode and choice between providers) and questions of hospital capacity management to meet demand. Models will tend to be at best partial and to involve simplifying assumptions. They also tend to be either statistical models aiming at best fit to observed data with forecasts based on extrapolating to the future or simulation models which generate future scenarios based on realistic input values without necessarily considering issues of fit.

Much statistical work has focussed on the less predictable emergency care element of hospital workload, on particular care areas (e.g. geriatric care) or on making demand estimates for single providers. This work may have a regional focus but often concentrates on simple summary measures (e.g. total emergency admissions from an entire region to a single provider), rather than providing a disaggregated overview of demand for different demographic groups, different residential areas, multiple providers and different areas of care (e.g. specialties). Thus Milner (1988) presents an ARIMA model of attendances at emergency units, while use such a model to forecast the need for surgical beds; a later review of such work appears in Milner (1997), with Jones et al. (2002) using a GARCH model for emergency bed occupancy at one provider. Operational research models based on queuing theory include Utley et al. (2003), Gilchrist (1985) and a number of papers by Millard and co-workers (e.g. Millard et al. 2000; Mackay and Millard 1999). Downing and Wilson (2002) present a descriptive analysis of temporal and demographic variations in A&E attendances, describing variations over broad age groups. By contrast, Harper et al. (2005) present a simulation model for patient flows that includes area of residence and travel mode but without allowance for patient demography.

The present analysis contributes to these developments using a Bayesian modelling approach which includes both demand generation and allocation to specialties and providers using a gravity model and extends to forecasting medium-term acute health demand. The model provides a demand overview in the sense of being disaggregated by age group, specialty, provider, area of residence and potentially other stratifiers. The model may be extended to include supply changes or to allow for capacity parameters (length of stay, occupancy)—see Sects. 8 and 10. In this connection, Taket and Mayhew (1981) develop an earlier gravity model of patient flows in London (though without including demographic group or specialty), while Tebaldi and West (1998) illustrate a Bayesian gravity model approach to transportation flows.

In the present analysis, projections of acute demand for future years (t) are made in terms of specialty (s), hospital provider (h), patient age (a) and local authority of residence i). Other variables might potentially be included in such an analysis, such as patient's sex, income group or social class. The latter two variables are not available routinely in the UK health records but might be especially relevant in projections of health care which also involved a split between private and public health care.

Because study region data on specialty use by age are not available on an extended time series basis, the analysis includes retrospective analysis of England-wide specialty referral trends (from 1991 to 1992). These data are analysed with a view to establishing past trends and projecting likely future growth in specialty-specific referral rates by age. The study region health demand data pertains to a single year (1997–1998) and contains hospital referral data by specialty, patient age, patient area of residence and hospital provider. Also available for the study region are projections of the population and its age structure in the year 2005–2006. These two sources of data, combined with the England-wide projections of specialty use, form the basis for future projections of health demand by local authority area and specialty and of the way it is distributed across the ten acute providers.

In summary terms the model involves a demand generation sub-model which predicts numbers of referrals by area, age, specialty and year $n_{i,a,s,t}$. This model uses national data on past specialty trends by age N_{ast} to project likely specialty demand, together with regional projections of population by age and area p_{iat}. With sufficient regional data on specialty trends the need to involve national reference data might be avoided. The projected referrals $n_{i,a,s,t}$ are then allocated to providers via an extended gravity model formulation (i.e. with areas as origins and hospitals as destinations but with additional stratification by specialty).

The allocation stage may be based simply on projections of existing activity, with referrals serving as mass measures in the gravity allocation. This may be broadly termed a demand-led projection. Alternatively, the gravity model may be adjusted to allow for new hospital sites, closure of sites or expansion or contraction of bed numbers at existing sites (Congdon 2000). Under this option, total acute beds (for example) may provide the mass measure in the gravity model.

3 Projecting Referral Patterns by Age and Specialty

Referrals for the study region illustrate the pronounced age gradients of inpatient usage rates and therefore the potential impact in the next decade of an increasing proportion of elderly populations. Table 14.1 shows the high rates of usage by persons over 75 for specialties such as urology, ophthalmology and general medicine. For the broad categories of general surgery and general medicine, which together account for around a third of all regional referrals (i.e. around 150,000 in an annual total of 450,000), the referral rate among the over 75s is 3–4 times the average rate for all ages (Fig. 14.2). The main objectives of the generation stage of the analysis are to project these usage profiles by age and specialty at local authority area level to mid-decade.

Let $n_{ias} = \sum_h n_{i,a,s,h}$ denote the pattern of referrals in a particular year, aggregating over hospitals h. Here the year concerned is 1997–1998. The ages are as in Table 14.1, and there are $S = 18$ specialties. Since these are count data, with potentially small numbers of events involved, a Poisson model is assumed in relation to age-specific populations by area, denoted p_{ia}, for the same year. Thus for a given year

$$n_{ias} \sim \text{Poi}(\mu_{ias})$$

$$\mu_{ias} = p_{ia} \times \rho_{ias}$$

where ρ_{ias} are age- and specialty-specific usage rates by local area ($i = 1, \ldots, 13$; $a = 1, \ldots, 7$; $s = 1, \ldots, 18$). These in turn are modelled in terms of two factors:

- Age and specialty effects across the region γ^{R}_{as}
- Area- and specialty-specific effects δ_{is}

Table 14.1 Hospital referrals, numbers and rates, in study region, by age band and specialty 1997–1998

	Age group								
	0–4	5–14	15–44	45–64	65–74	75–84	85+	All ages	
Total episodes	51,619	20,996	154,752	87,685	59,699	55,269	24,313	454,333	
Total population of region	158,073	293,474	946,915	444,862	163,639	103,451	35,030	2,145,444	
Age-specific referral rates per 100,000 population (specialties with all-ages rate over 500/100,000)									Ratio of 75+ rate to average
General surgery	621	870	1,707	2,975	4,657	5,687	5,161	2,249	3.3
Urology	289	206	405	1,404	3,826	5,048	3,548	1,112	5.6
Trauma and orthopaedics (T&O)	631	969	1,001	1,487	2,207	3,331	4,799	1,337	3.7
Ear nose and throat (ENT)	1,313	1,488	565	564	599	508	474	744	0.9
Ophthalmology	317	146	152	624	2,121	3,955	4,373	664	8.2
General medicine	288	206	2,238	6,135	12,793	13,563	14,705	4,179	4.4
Paediatrics	21,379	1,655	68	–	–	–	–	1,832	–
Geriatric medicine	–	–	94	347	2,132	13,282	28,510	1,382	–
Gynaecology, obstetrics	5,138	29	7,389	1,296	645	587	468	3,997	0.2
Mental illness	9	18	774	590	510	888	1,108	567	2.2
Total (including minor specialties)	32,655	7,154	16,343	19,711	36,482	53,425	69,406	21,177	3.6

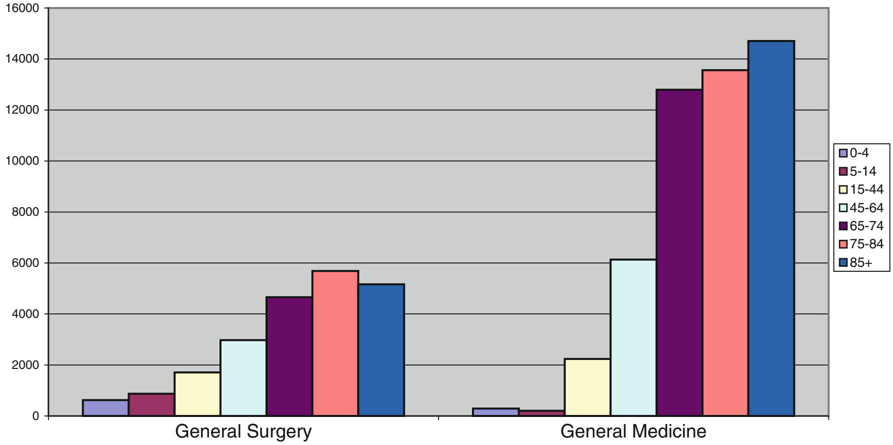


Fig. 14.2 Annual referral rates per 100,000 (1997–1998) by age group and study region

Both sets of effects are taken to be random. They reflect, respectively, the contrasts in age usage rates within and between specialties (as illustrated in Table 14.1) and area level departures from regional norms in terms of specialty use.

The analysis may therefore be seen "provision constrained" as it incorporates current variations between areas in their referral rates to different specialties. Some of the differences δ_{is} may be related to patient need (e.g. populations may differ in their need for maternity services or cancer services because of social factors) and others to variations in supply of services.

Adding main effects for specialty and age was found to make little difference to outputs from subsequent stages of the model, such as the final projections of regional hospital activity by specialty. So the model for region-wide referral rates in the base year $t = B$ (here 1997–1998) is

$$\log(\rho_{ias}) = \alpha + \delta_{is} + \gamma^{\mathrm{R}}_{as} \tag{14.1}$$

We assume that the first component of this model, namely, area differences in specialty usage rates δ_{is}, remains constant in future years but anticipate that demand growth is likely to vary by specialty and by age also. Hence the element γ^{R}_{as} will be subject to change and need to be projected on the basis of past trends. Projections will be facilitated if these trends are consistent and provide a clear basis for forecasting.

Table 14.2 Trends in England-wide specialty referral rates by age (per 100,000 population)

										Annual growth
	Age group									Ages
	0–4	5–14	15–44	45–64	65–74	75–84	85+	All Ages	All Ages	75+
1998/99										
General surgery	569	686	2,051	3,803	6,116	6,902	6,230	2,870	0.8	2.1
Urology	191	143	465	1,375	3,564	4,419	3,115	1,141	9.0	11.7
T&O	686	992	1,226	1,747	2,530	3,676	5,243	1,602	1.6	2.8
ENT	1,341	1,499	602	587	633	642	551	765	0.0	6.5
Ophthalmology	369	157	170	671	2,273	4,605	5,333	813	8.0	11.2
General medicine	156	154	2,473	6,522	13,636	17,390	19,428	5,049	9.9	18.8
Paediatrics	29,722	2,556	114	1	2	4	20	2,226	0.9	
Geriatric medicine	1	1	62	193	1,881	9,371	20,536	1,135	0.3	0.1
Obstetrics and gynaecology	1,552	46	7,775	1,477	713	623	422	3,817	0.0	0.0
Mental illness	53	185	588	394	659	1,145	1,573	513	−1.7	−2.7
Total	33,370	8,163	18,142	22,399	40,540	57,189	70,644	24,097	3.5	5.9
1991–1992										
General surgery	961	1,007	2,012	3,462	5,339	6,225	5,258	2,720		
Urology	179	134	333	842	2,017	2,504	1,632	699		
T&O	728	1,023	1,190	1,483	2,041	3,089	4,362	1,443		
ENT	1,601	1,864	559	484	524	457	363	763		
Ophthalmology	455	200	127	439	1,344	2,662	2,909	522		
General medicine	121	130	1,652	4,265	8,145	8,437	7,481	2,985		
Paediatrics	28,417	2,006	97	44	74	100	106	2,090		
Geriatric medicine	5	1	24	105	1,862	9,936	19,811	1,114		
Obstetrics and gynaecology	1,552	46	7,775	1,477	713	623	422	3,817		
Mental illness	49	187	639	492	702	1,443	1,904	584		
Total	36,715	8,057	>6,335	16,336	27,260	40,439	49,965	19,404		

4 Projecting National Specialty Referral Rates

The national evidence indeed provides clear trends in acute hospital referrals. Thus referral rates (per 100,000 population) across all specialties have been increasing throughout England at around 6 % annually for the over 75s and around 3.5 % for all ages—see Table 14.2. The all-ages referral rate, aggregating over all specialties, has grown from 19,400 in 1991–1992 to 24,100 in 1998–1999. However, the highest increase in referral rates is for the age groups 65–74, 75–84 and 85+. These changes are purely an effect of increased usage and not a reflection of ageing population except in so far as a more long-lived elderly population may generate extra morbidity. In terms of annual numbers, England-wide referrals have risen from 9.4 to 12 million for all ages and from 1.4 million to 2.2 million among the 75s.

Changes in demand are also differentiated by specialty. All types of medical referral have risen appreciably both for the over 85s and for all ages combined (Figs. 14.3 and 14.4). The highest growth rates for surgical specialties are for urology and ophthalmology.

We use such national trends in specialty activity as a basis for projections of future demand by age and specialty in the study region, i.e. we consider time trends in national age and specialty referral rates, denoted $\gamma_{ast}^{\mathrm{N}}$, and project them to a future year T. Such trends reflect many factors, including changing patterns of illness, reductions in length of overnight stays, increases in day case referrals and changes in technology and service provision across sectors. They may also reflect clinical policies (e.g. on age-related criteria) regarding suitability of surgical interventions for older people and improved procedures for managing risks associated with such interventions (Chalfin and Nasraway 1994).

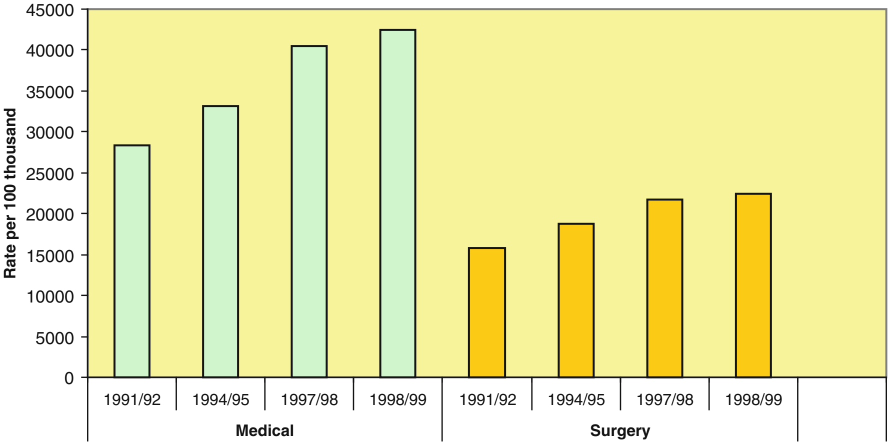


Fig. 14.3 Trend in medical and surgical referral rate for ages over 85 (England)

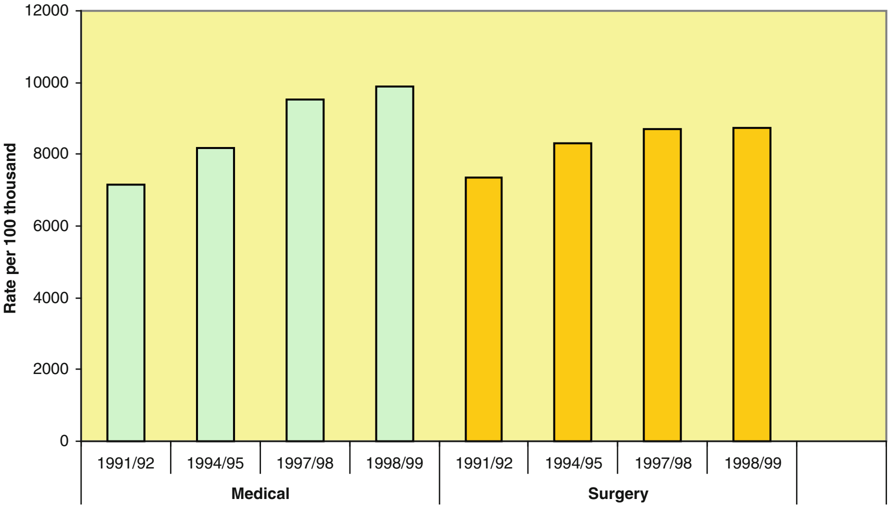


Fig. 14.4 Trend in medical and surgical referral rate (all ages, England)

Changes in the study region in terms of demand by specialty and age between now and 2005–2006 are assumed to follow the forecast national trend in age-specific specialty use, the results of which (from 1998–1999 to 2005–2006) are summarised in Table 14.3. Thus general medicine usage across all ages is forecast to increase by 36 % over 1998–1999 to 2005–2006 or by around 5 % per year. These forecasts allow for changing referral over time by age group and by specialty and so for the increase in usage in many specialties observed up to 1998–1999. However, the increase observed between 1991–1992 and 1998–1999 is "damped down" in the extrapolation from 1998–1999 to 2005–2006. This damping produces a greater compatibility with the National Beds Inquiry age usage forecasts (from 1998/1999 to 2003/2004) which in a sense provide a guide forecast.

Specifically, let N_{ast} denote the observed England-wide specialty referral totals by specialty s, age group a and year t, and let national age-specific populations for year t be denoted P_{at}. Then referrals by specialty s, age a and year t (=1,8) from 1991–1992 to 1998–1999 are binomial as follows:

$$\begin{aligned} N_{ast} &\sim \mathrm{Bin}\left(\gamma^{\mathrm{N}}_{ast}, P_{at}\right) \\ \mathrm{log}it\left(\gamma^{\mathrm{N}}_{ast}\right) &= A + a_s + b_a + c_{sa} + d_{st} + e_{at} \end{aligned} \tag{14.2}$$

The prior for the intercept is $A \sim N(0,10)$, and the age, specialty and age specialty effects, a_s, b_a and c_{sa}, are assumed to be random:

$$\begin{aligned} a_s &\sim N(0, \tau_a), \qquad s = 1, \ldots, 21 \\ b_a &\sim N(0, \tau_b), \qquad a = 1, \ldots, 7 \end{aligned}$$

Table 14.3 Forecast change in population hospitalisation rate, 1998–1999 to 2005–2006 (ratios of rates, later to earlier period)

	Age group							
Specialty	0–4	5–14	15–44	45–64	65–74	75–84	85+	All ages
General surgery	1.13	1.10	0.98	1.02	1.10	1.17	1.23	1.04
Urology	1.21	1.24	1.30	1.33	1.41	1.44	1.52	1.31
T&O	0.94	0.95	1.05	1.07	1.16	1.24	1.31	1.06
ENT	0.99	1.01	1.02	1.06	1.15	1.11	1.17	1.04
Ophthalmology	1.33	1.39	1.22	1.25	1.31	1.34	1.41	1.27
Oral/dentistry	1.06	1.03	1.22	1.16	1.23	1.06	1.16	1.16
Other surgery	1.00	1.06	1.14	1.18	1.28	1.27	1.37	1.15
Plastic surgery	1.12	1.13	1.30	1.34	1.45	1.48	1.55	1.30
General medicine	1.34	1.37	1.35	1.39	1.39	1.29	1.27	1.36
Audiological, etc.	1.17	1.02	1.18	1.23	1.30	1.18	1.22	1.18
Cardiology	1.66	1.66	1.46	1.55	1.54	1.47	1.51	1.53
Medical oncology	2.12	1.95	1.51	1.48	1.53	1.59	1.65	1.61
Neurology	1.32	1.28	0.94	1.02	1.07	1.20	1.12	1.06
Rheumatology	0.71	0.76	1.14	1.19	1.24	1.22	1.18	1.09
Paediatrics	1.05	1.03	1.05	–	–	–	–	1.05
Geriatric medicine	1.27	0.67	0.63	0.73	0.95	1.08	1.07	0.76
Obstetrics and GP maternity	0.54	0.61	1.06	1.29	1.56	0.67	0.55	1.03
Gynaecology	1.32	1.32	1.00	1.09	1.12	1.06	1.18	1.10
Mental illness, old age psychiatry	0.82	0.85	0.86	0.98	0.98	1.21	1.32	0.92
Oncology/radiology	2.13	2.14	2.02	2.09	2.20	2.35	2.77	2.10
Other specialties	0.59	0.43	1.03	1.10	1.23	1.44	1.60	0.99
Total	1.03	1.04	1.11	1.27	1.32	1.27	1.26	1.16

$$c_{sa} \sim N(0, \tau_c).$$

The precisions $1/\tau_a$, $1/\tau_b$ and $1/\tau_c$ are taken to have gamma priors $G(1{,}0.001)$.

The forecasting element of the model for England specialty rates is based on the time-specific parameters d_{st} and e_{at}, namely, trends in usage by specialty and by age. The year 1 parameters d_{s1} and e_{a1} are distributed randomly as $N(0.\tau_d)$ and $N(0{,}\tau_e)$, respectively. Subsequent years of the $d_{s,t}$ are modelled in terms of specialty-specific growth rates δ_s with damping when taken into future years:

$$\begin{array}{ll} d_{s,t} = d_{s,t-1} + \delta_s \times \log(t) & t = 2, \ldots, 8 \\ d_{s,t} = d_{s,t-1} + \delta_s \times \log(t) \times \left(0.75\right)^{t-8} & t = 9, \ldots, 15 \end{array} \tag{14.3a}$$

The parameters δ_s are random effects, with prior $N(0{,}\tau_\mathrm{f})$. Trends in use by age are similarly forecast using age-specific growth rates ε_a:

$$\begin{array}{ll} e_{a,t} = e_{a,t-1} + \varepsilon_a \times \log(t) & t = 2, \ldots, 8 \\ e_{a,t} = e_{a,t-1} + \varepsilon_a \times \log(t) \times \left(0.75\right)^{t-8} & t = 9, \ldots, 15 \end{array} \tag{14.3b}$$

The ε_a are also random effects, with prior $N(0{,}\tau_\mathrm{g})$.

Table 14.4 Specialty and age groups, referral growth rate parameter estimates

	Mean	2.5 %	97.5 %
Specialty demand growth parameters			
General surgery	−0.80	−0.81	−0.78
Urology	2.43	2.40	2.45
T&O	−0.19	−0.21	−0.17
ENT	−0.40	−0.43	−0.36
Ophthalmology	2.00	1.97	2.04
Oral/dentistry	1.07	1.03	1.10
Other surgery	1.12	1.08	1.16
Plastic surgery	2.48	2.41	2.52
General medicine	3.47	3.46	3.48
Audiological, etc.	1.62	1.58	1.65
Cardiology	4.64	4.60	4.69
Medical oncology	4.35	4.31	4.40
Neurology	−0.81	−0.88	−0.74
Rheumatology	0.75	0.67	0.81
Paediatrics	1.70	1.66	1.73
Geriatric medicine	−2.63	−2.66	−2.60
Obstetrics and GP maternity	−0.73	−0.74	−0.71
Gynaecology	0.22	0.19	0.24
Mental illness, old age psychiatry	−2.59	−2.62	−2.55
Oncology/radiology	9.40	9.36	9.43
Other specialties	0.02	−0.03	0.06
Age group demand growth parameters			
0–4	−0.71	−0.73	−0.65
5–14	−0.55	−0.57	−0.52
15–44	0.57	0.56	0.58
45–64	1.29	1.27	1.30
65–74	2.13	2.11	2.14
75–84	2.56	2.55	2.57
85+	3.25	3.22	3.28

Table 14.4 shows the estimates of the growth rate parameters by age and specialty, ε_a and δ_s, together with 95 % credible intervals. Interpretation may need to reflect demographic demand: thus growth in paediatrics may have outpaced the growth in demographic demand per se, so that the age group parameter ε_1 is negative. On the other hand growth in geriatric referrals may have been less than implied by demography (as efforts are made to reduce bed blocking by older patients and improve community care transfers).

Evaluations of these forecasts involved prediction of 1998–1999 observations using the data for the first 7 years only. The performance of the model above (denoted model N1) was compared to two more highly parameterised variants. The first involves forecasting at each age and specialty combination, so that (14.2) is replaced by

$$\text{log}it\left(\gamma_{ast}^{\mathrm{N}}\right) = A + a_s + b_a + c_{sa} + f_{ast} \tag{14.4}$$

with the evolution of f_{ast} based on logarithmic growth rates specific for age and specialty

$$f_{a,s,t} = f_{a,s,t-1} + \eta_{as} \times \log(t) \qquad t = 2, \ldots, 8 \tag{14.5a}$$

and with $f_{a,s,t}$ modelled as a separate set of random effects (this may be termed model N2). The second variant extends this approach by allowing the power of t (implicitly zero for a' trend in log time) to be a free parameter at specialty level. Thus in model N3

$$f_{a,s,t} = f_{a,s,t-1} + \eta_{as} \times t^{\xi_s} \qquad t = 2, \ldots, 8 \tag{14.5b}$$

Comparisons are made between actual referral rates by age and specialty in 1998–1999, namely, $R_{as9} = 1{,}000N_{a,s,9}/P_{a,\mathrm{Q}}$, and predictions for $\hat{R}_{as9}$. A chi-square statistic is accumulated over the 147 specialty–age combinations and shows the models N2 and N3 with age–specialty combination forecasting to have lower prediction error. This is especially so when combined with the model (14.5b) with different growth rate scales. Whereas the chi square for N1 is 207, that for model N2 is 103 and that for N3 is 39.

More heavily parameterised models will generally improve on the fit of the model to observed data and on the accuracy of cross-validatory predictions within the observed data. On the other hand, there is no reason why incorporating the forecasts from N2 or N3 should lead to "better" forecasts for the future year 2005–2006 (i.e. for out of sample rather than cross-validatory predictions) at a regional level. Therefore we might view alternative methods to obtain national forecasts as one scenario to adopt in a sensitivity analysis of the regional activity forecasts.

5 Comparison with the UK National Beds Inquiry

Table 14.5 shows the match between the activity growth assumptions (to 2003–2004) made in the UK National Beds Inquiry, which are not specialty specific, and those resulting from the modelling analysis here (with model N1), when aggregated over specialties. There is a 22 % growth in the all-ages usage rate here as compared to 27 % (7-year equivalent) in the National Beds Inquiry.

These projected demand growth rates are lower than in the recent past in England. However, to simply extrapolate past trends would neglect efforts being made in England (as in other countries) to manage demand for acute care by encouraging community and primary care, especially policies to reduce avoidable admissions and bed blocking. Also much of the rise in admissions has been due to reduced length of stay and an increase in day cases, and national projections are for a slower fall in average length of stay in the next 5 years.

Table 14.5 Forecasts to 2005–2006 of England specialty referral rates by age

	Age group							
Specialty	0–4	5–14	15–44	45–64	65–74	75–84	85+	All ages
General surgery	642	758	2,014	3,886	6,745	8,072	7,656	3,031
Urology	232	178	602	1,826	5,031	6,364	4,723	1,569
TOO	645	943	1,289	1,877	2,926	4,540	6,857	1,761
ENT	1,322	1,516	614	620	728	711	643	791
Ophthalmology	490	219	206	836	2,983	6,186	7,497	1,069
General medicine	209	211	3,350	9,088	19,250	22,470	24,580	6,848
Paediatrics	31,250	2,636	120	9	18	26	39	2,338
Geriatric medicine	1	1	39	141	1,785	10,110	22,060	1,175
Obstetrics and gynaecology	1,536	42	8,157	1,790	1,012	949	726	4,097
Mental illness	43	156	507	386	647	1,390	2,071	495
All specialties	36,369	6660	16,898	20,459	41,124	60,817	76,852	28,728
Change in all-specialties rate, 1998–1999 to 2005–2006	0.957	0.804	0.933	0.943	1.057	1.109	1.126	1.217
National beds Inquiry change	1.113	1.054	1.137	1.386	1.384	1.332	1.267	1.268
Age group trends in usage rates and average length of stay (National Beds Inquiry) 1969–1990 to 1998–1999 (annual change rate)								
Length of stay (excl day cases)	−4.1	−5	−3.1	−3.9	−4.3	−6	−9.1	−4
Usage rate per 100,000	0.6	0.2	1.6	5.3	7.0	5.9	5.9	3.5
Length of stay	3.34	2.21	3.12	5.86	8.75	12.34	16.29	7
1998–1999								
Usage rate per 100,000	22,400	7,800	13,300	20,300	35,400	47,200	56,600	19,300
1998–1999 to 2003–2004 (annual change rate)								
Length of stay	−2.6	−2.0	−1.6	−1.9	−2.5	−2.5	−2.5	−1.9
Usage rate per 100,000	1.6	0.8	2.0	5.5	5.5	4.7	3.8	3.8
2003–2004								
Length of stay	2.87	1.99	2.87	5.3	7.65	10.79	14.28	6
Usage rate per 100,000	24,200	8,100	14,600	25,900	45,100	58,400	67,400	23,000

In this context, Table 14.5 presents the assumptions of the National Beds Inquiry over the period 1998–1999 to 2005–2006. It shows the smaller decline in average length of stay expected to occur in that period as compared to the 1990s, together with the expected increase in age-specific usage of all kinds.

6 Projections of Total Region-Wide Acute Activity

The national projections of age- and specialty-specific usage rates (as summarised in Tables 14.3 and 14.4) are now incorporated in forecasts of region-wide activity by specialty. These region-wide forecasts also take account of projected population change p_{iat} in areas i, up to a future forecast year T. Let $\kappa_{as} = \gamma^{\mathrm{R}}_{asB}/\gamma^{\mathrm{N}}_{asB}$ denote the ratio, comparing the study region to England, of age specialty usage in the base year B. We assume (in the absence of data to confirm differential national-regional trends) that this differential remains constant between years B and T.

Projected study region referral rates in a future year T are modelled as

$$\log(\rho_{iasT}) = \alpha + \delta_{is} + \kappa_{as} + \gamma^{\mathrm{N}}_{asT} \tag{14.6}$$

So the expected total referrals in the future year T are given by

$$\mu_{iasT} = p_{iaT} \times \rho_{iasT}$$

where p_{iaT} are regional population projections. To allow fully for sampling variability we can sample projected referrals, conditional on these projected means:

$$n_{iasT} \sim \mathrm{Poi}(\mu_{iasT})$$

With regard to expected population changes p_{iat} up to the year T (i.e. 2005–2006), two sources are available for London boroughs and a single source for S. Essex areas. One source for the London boroughs (the UK Office of National Statistics projections) does not take account of planned housing development, whereas projections from the Greater London Authority do. In the analysis below, an *average* of the two sets of population projections is taken for NE London local authorities. This option is chosen in preference to making variant activity forecasts according to GLA or ONS projections.

7 Modelling the Distribution of Total Activity to Alternative Sites: Extended Gravity Model for Patient Flows

At the allocation stage, a patient flow model is established for age- and specialty-specific patient flows from area i (origins) to hospital sites h (destinations). This is therefore a version of the gravity model for health demand (e.g. Taket and Mayhew

1981) but with an additional dimension of medical specialty. To this end the "destination" mass effect of the gravity model may be taken as specific to both provider and specialty, and so the model includes specialty-specific mass parameters λ_s which are applied to a mass given by the total number M_{sh} of referrals to different specialties in each hospital from the entire region. The model also includes distance decay effects in terms of distance d_{ih} (between grid references) from the area i to the hospital h. The decay parameters are specific both to origin and specialty, since some specialties occur at all sites but some at only one or two—so that the latter will have less steep decay.

Developments of this stage may also include the overall impact of beds and other types of contiguity or association between areas and hospitals. A distinction may be made between allocation models which are based on the following:

(a) Existing referral patterns relating areas to providers but taking account of changing health demands across demographic groups and specialties (demand-led models).
(b) Allowing a change in spatial supply patterning as well as in health demand by including changes in mass (e.g. beds, staff) at different hospitals (composite demand and supply models): Major changes in provider configuration (e.g. openings of new hospital sites or closures of existing sites) may also be included (Congdon 2000).

The former demand-led model can be translated into bed requirements corresponding to an ageing population or increased demand for different types of care. Thus with LOS denoting length of stay in hospital we can use the standard relationship (at specialty level)

$$\text{Projected beds} = (\text{Average LOS} \times \text{projected episodes})/(\text{Occupancy rate} \times 365)$$

to project numbers of non-day beds. Day beds may be projected by assuming an average number of days worked each week, and the average number of patients (e.g. 1.5) occupying a day bed in each day worked.

The supply-led model (b) may, by contrast, be used in conjunction with modelling methods which allow for opening new hospitals and closing existing ones as well as for simple changes in bed numbers. With the first approach we therefore translate changes in the demographic structure of catchment populations to the demands for different acute care (e.g. implied bed numbers). In the latter we convert projected changes in health demand *and* adjustments in supply to predict a reconfiguration of spatial flows.

Under either option, the impact of changes in age structure on specialty demand by hospital depends on the catchment area of each hospital, since areas differ in their overall population growth rates and the extent of their ageing populations. In 1998/1999 total referrals from the 13 areas were around 455,000, of which 90 % (402,000) were to the ten study hospitals—so emphasising the high self-containment of the "region" defined by the ten hospitals. Just as the impact on hospital workload of demographic change depends on the trusts' catchment areas,

Table 14.6 Trust workload by specialty, 1998–1999 (numbers of referrals from study region)

Specialty	Southend Acute	Basildon and Thurrock	Forest Health Care	Redbridge Health Care	Havering Hospitals	Newham Health Care	Royal London	Mid Essex	Princess Alexandra	Homerton	Total
General surgery	12.0	12.2	8.9	13.3	13.6	14.6	7.6	2.4	24.4	13.7	11.5
Urology	5.2	6.7	4.8	6.1	8.2	3.4	6.1	1.1	4.9	1.9	5.6
Trauma/orthopaedics	7.9	8.6	4.6	6.5	7.7	6.1	4.5	1.8	15.7	5.1	6.4
ENT	3.6	5.9	4.9	0.0	5.2	0.6	3.4	0.5	3.6	0.1	3.3
Ophthalmology	7.6	0.0	5.1	0.2	2.3	0.0	0.7	0.2	1.8	0.0	2.4
Oral/dentistry	1.2	2.8	1.3	0.0	1.9	0.0	4.4	0.4	4.0	0.8	1.7
Other surgery	2.2	0.6	4.1	0.3	1.8	0.0	7.3	0.7	0.0	0.0	2.4
Plastic surgery	1.0	0.0	0.0	0.0	0.3	0.0	2.2	72.3	2.2	0.3	2.1
General medicine	21.1	19.2	19.5	20.6	22.2	21.3	23.7	3.4	15.2	23.9	21.0
Audiological, etc.	0.5	0.6	2.6	1.4	1.6	4.9	5.7	0.2	0.0	5.5	2.5
Cardiology	0.2	3.3	1.4	0.0	0.4	0.0	7.8	0.0	0.0	0.0	1.9
Medical oncology	0.0	0.0	0.0	2.9	0.0	0.6	3.6	0.3	9.3	0.0	1.0
Paediatrics	3.1	10.3	13.5	14.3	3.8	10.0	6.2	6.5	0.0	15.7	8.5
Geriatric medicine	10.3	9.4	8.4	3.7	6.4	10.0	0.0	0.4	0.0	1.4	6.2
Maternity, etc.	19.5	18.9	18.1	20.5	23.3	26.3	15.7	9.7	17.8	31.6	20.7
Mental illness + old age psychiatry	0.3	0.0	2.6	5.8	0.0	0.0	0.0	0.0	0.1	0.0	1.0
Oncology/radiology	4.1	0.0	0.0	0.0	0.9	0.0	1.0	0.0	0.0	0.0	1.0
Other specialties	0.0	1.4	0.2	4.3	0.5	0.3	0.0	0.1	0.9	0.0	0.7

so does the impact of changing specialty referral rates depends in part on the specialty profile of the trusts. Table 14.6 shows the distribution of specialty referrals by patient area and hospital in 1998–1999 for patient flows originating in the study region. Certain specialties only figure significantly at one or two provider sites and that some providers have a more restricted specialty range than others. To reflect local patterns of supply, hospital- and specialty-specific effects α_{sh} are included in the allocation model.

8 Incorporating Supply Change

In the present analysis option (b) above is adopted, in that a change in provider configuration in the study region is also included in the model calibration. Thus, following a review of health care in London (Department of Health 1999), a major planned change affecting this part of London is a centralisation of acute services in the outer North East London. This involves an expanded and more central Oldchurch site and conversion of the Harold Wood site on the London boundary to an outpatient facility only (see Fig. 14.1). Actual 1998/1999 beds and future assumed bed numbers B_h (2005/2006) at the ten providers are as follows:

Hospital provider	1998/1999	Projection year
Newham	362	362
King Georges (Redbridge Health Care)	395	416
Whipps Cross (Forest Health Care)	446	446
Oldchurch Hospital	400	722
Harold Wood Hospital	300	–
Royal London	1,116	1,150
Broomfield (Mid-Essex Acute Trust)	601	601
Basildon and Thurrock	440	440
Southend	495	495
Homerton	321	321
Princess Alexandra	329	329

To see how these factors are included in the allocation stage, we assume for simplicity that all age groups are subject to the same allocation model. Accordingly consider the total observed referrals by area and specialty in the base year B:

$$n_{i,s} = \sum\nolimits_a n_{iasB}$$

To model allocations of this total between sites we consider the data

$$n_{i,s,h} \sim \sum\nolimits_a n_{i,a,s,B,h}$$

as a multinomial response

$$n_{i,s,1:\mathrm{H}} \sim \mathrm{Mult}(\theta_{i,s,1:\mathrm{H}}, n_{is})$$

where the choice probabilities are parameterised as

$$\theta_{ish} = \phi_{ish} / \sum\nolimits_h \phi_{ish}$$

The effects ϕ_{ish} take account of distances from areas i to hospitals h, the sizes of different specialties in different hospitals (in terms of annual referrals M_{sh}) and the total acute bed mass B_h in hospital h. Thus we take

$$\log(\phi_{sh}) = b_0 + \alpha_{sh} + \beta_{is} \times \log d_{ih} + \lambda_s \times \log(M_{sh} + 1) + b_1 \times S_{ih} + b_2 \times \log(B_h) \tag{14.7}$$

Note that under option (a) above (i.e. under a demand-led projection) we would omit mass factors at hospitals which are potentially subject to deliberate strategic revision, such as the term in B_h in (14.7)—as opposed to the M_{sh}. As well as the impact of area–hospital "similarity" through distances between them, there may be a secondary effect of a hospital being in the same health authority as the patient (represented by dummy indices $S_h = 1$ if this applies and $S_h = 0$ otherwise).

Given that there are 18 specialties and 10 hospitals to consider, we have 180 specialty and hospital effects α_{sh} to estimate (their impact in practice is illustrated in Table 14.7). We assume that these are drawn from a population of effects and therefore are random with

$$\alpha_{sh} \sim N(0, \tau_\alpha) \tag{14.8a}$$

The remaining effects are taken as fixed and specified in terms of prior assumptions about the likely nature of mass effects and distance decay. Thus the mass parameters have priors

$$\lambda_s \sim N(1, 1) \qquad s = 1, \ldots, 18 \tag{14.8b}$$

$$b_2 \sim N(0.5, 1) \tag{14.8c}$$

and the distance decay parameters have priors

$$\beta_{is} \sim N(-2, 1) \qquad i = 1, \ldots, 13; \quad s = 1, \ldots, 8 \tag{14.8d}$$

We would expect the effect of hospital being in the same health region as the patient's home to raise referral rates but take a vague prior on the associated parameter

$$b_1 \sim N(0, 100) \tag{14.8e}$$

Table 14.7 Parameter estimate summary, allocation mode

		Mean	2.5 %	97.5 %
Distance decay (by patient origin area)				
Barking and Dagenham	β_1	−3.33	−3.46	−3.22
Hackney	β_2	−3.61	−3.8	−3.43
Havering	β_3	−1.35	−1.45	−1.28
Newham	β_4	−2.76	−2.90	−2.61
Redbridge	β_5	−2.34	−2.42	−2.26
TH + City	β_6	−1.47	−1.60	−1.33
Waltham Forest	β_7	−1.97	−2.07	−1.88
Basildon	β_8	−2.57	−2.73	−2.43
Brentwood	β_9	−4.66	−4.84	−4.49
Castle Point	β_{10}	−2.16	−2.32	−2.00
Rochford	β_{11}	−2.3	−2.49	−2.13
Southend-on-Sea	β_{12}	−1.43	−1.52	−1.34
Thurrock	β_{13}	−5.79	−5.98	−5.61
Specialty mass effects				
General surgery	λ_1	0.45	0.29	0.65
Urology	λ_2	0.63	0.5	0.76
Trauma/orthopaedics	λ_3	0.41	0.17	0.62
ENT	λ_4	0.80	0.55	1.02
Ophthalmology	λ_5	1.64	1.19	2.12
Oral/dentistry	λ_6	1.13	0.95	1.42
Other surgery	λ_7	1.63	1.45	1.82
Plastic surgery	λ_8	2.82	2.58	3.00
General medicine	λ_9	2.02	1.84	2.20
Audiological, etc.	λ_{10}	2.22	1.83	2.58
Cardiology	λ_{11}	2.22	2.05	2.36
Medical oncology	λ_{12}	1.60	1.14	1.99
Paediatrics	λ_{13}	0.51	0.17	0.83
Geriatric medicine	λ_{14}	1.59	0.81	2.22
Maternity, etc.	λ_{15}	0.45	0.26	0.65
Mental illness + old age psychiatry	λ_{16}	1.56	1.20	2.03
Oncology/radiology	λ_{17}	1.82	1.48	2.22
Other specialties	λ_{18}	1.33	0.94	1.82
Other factors				
Hospital and LA in same HA	b_1	1.19	1.16	1.21
Beds	b_2	0.53	0.41	0.73

The predicted mean flows are then

$$v_{ish} = \theta_{ish} R_{is}$$

and may be compared (e.g. in terms of deviance fit measures) with the actual flows $n_{i,s,h}$.

The model of (14.7) is based on analyzing current flows. Changes in total bed mass ($B_{h,\mathrm{new}}$) or in specialty balance at hospitals (e.g. via bed loads by specialty $B_{sh,\mathrm{new}}$) may, however, be incorporated in the model to derive forecasts of allocation parameters for year T, denoted $\phi_{ish,\mathrm{new}}$:

$$\log(\phi_{\mathrm{ish.new}}) = b_0 + \alpha_{sh} + \beta_{is} \times \log d_{ih} + \lambda_s \times \log(M_{sh} + 1) + b_1 \times S_{ih} \\ + b_2 \times \log(B_{h.new}) \tag{14.9}$$

Table 14.7 shows the estimates of λ_s, b_1, b_2 and average distance decay parameters in model (14.7) by patient area of residence β_i (averaging over specialties). Local authorities without a major hospital within their boundaries (e.g. Brentwood, Thurrock) but with a high dependence on a few sites have more highly negative values of β_i. Specialty-specific supply effects λ_s are greater for those specialties (e.g. plastic surgery) which are located at only a minority of the ten hospitals. The secondary effect of a hospital being in the same health authority as the patient is also clear, which may in part reflect referral preferences among primary care practitioners. The positive impact of bed numbers is consistent with the conventional gravity model.

With new bed numbers and revised health demand by specialty we therefore are in a position to predict health flows in the "new" situation of 2005–2006. Specifically, we use the predicted allocation rates

$$\theta_{ish.\mathrm{new}} = \phi_{ish.\mathrm{new}} / \sum\nolimits_h \phi_{ish.\mathrm{new}}$$

to predict new mean flows

$$v_{ish.\mathrm{new}} = \theta_{ish.\mathrm{new}} \times n_{isT}$$

where $n_{isT} = \sum{}_a n_{iasT}$ and the n_{iasT} are projected as in (14.6) using the England-wide specialty forecasts. To allow for sampling variability we can then sample new flows

$$n_{i,s,h,\mathrm{new}} \sim \mathrm{Poi}(v_{ish.\mathrm{new}})$$

9 Forecasts and Their Sensitivity

The resulting forecasts of specialty demand by age in the study region in 2005–2006 are given in Table 14.8. They show a projected growth of around 75,000 episodes across the region by 2005–2006 as compared to 402,000 to the ten providers in 1998–1999. Note that the latter total relates to the ten providers excluding flows to non-study hospitals.

Table 14.8 Patient flow forecasts in 2005–2006 by specialty and hospital

Specialty	Hospital Southend Acute	Basildon and Thurrock	Forest Health Care	Redbridge Health Care	Havering Hospitals	Newtiam Health Care	Royal London	Mid Essex	Princess Alexandra	Homer-ton	Total
General surgery	6,446	6,944	4,602	4,936	9,575	5,691	5,171	319	241	3,384	49,512
Urology	4,765	5,122	3,341	2,825	7,459	1,723	5,396	179	59	632	31,501
Trauma and orthopaedics (T&O)	5,762	5,092	2,562	2,644	5,432	2,462	3,152	219	142	1,294	28,761
Ear nose and throat (ENT)	2,409	3,172	2,528	–	3,383	208	2,226	54	27	30	14,037
Ophthalmology	7,040	–	3,342	79	1,830	–	554	32	19	0	12,896
General medicine	16,970	14,270	13,610	9,398	20,310	10,870	20,830	686	208	7,770	116,821
Paediatrics	2,020	5,019	6,591	5,275	3,045	3,708	4,181	674	0	3,753	34,266
Geriatric medicine	6,963	5,540	4,009	1,140	4,593	3,468	0	60	0	319	26,092
Gynaecology, obstetrics	12,770	9,616	9,033	7,072	16,070	10,580	10,370	1,198	164	7,161	64,054
Menial illness	212	–	1,259	2,327	–	–	–	–	–	–	3,798
Total	79,457	60,748	57,242	40,469	78,185	41,464	79,641	11,823	1,022	26,257	476,309

Table 14.9 Forecast trust workload under variant priors

		Credible interval	
Hospital trust	Mean	2.5 %	97.5 %
Southend Acute	79,410	78,590	80,260
Basildon and Thurrock	60,740	60,020	61,490
Forest Health Care	57,410	56,770	58,050
Redbridge Health Care	40,160	39,590	40,740
Havering Hospitals	78,430	77,610	79,250
Newham Health Care	41,520	40,970	42,080
Royal London	79,430	78,700	80,310
Mid Essex	11,850	11,540	12,170
Princess Alexandra	1,035	963.6	1,111
Homerton	26,340	25,900	26,700
Total	476,325	470,654	482,161

The expanded Oldchurch site has a growth of around 15,000 episodes as compared to the 1997–1998 totals in Table 14.6. The fast-growing general medicine specialty increases by around 6,000 episodes at this site.

The repercussions on neighbouring hospitals of an expansion at one site are important to consider. Thus the workload of King George Hospital, relatively close by to the expanded Oldchurch hospital, is forecast to be static, despite the region-wide growth in demand.

Impacts of population ageing by contrast are greater in areas of suburban Essex, most notably in post-war New Towns (of which Basildon was one). These experienced a major influx of young families in the 1950s and 1960s, and consequent cohort ageing is apparent in increased demand at Basildon and Thurrock Trust (from 48,000 episodes in 1997–1998 to around 60,000 in 2005–2006).

As a sensitivity analysis of these forecasts we may consider first a modification of the above prior model assumptions, as in (14.8a)–(14.8e). To this end, we adopt a heavy-tailed alternative to (14.8a), namely, a t density with 5 degrees of freedom, and replace priors (14.8b)–(14.8d) by normal priors with zero mean and variance 14. This has little impact on the overall forecasts of total hospital workload, as given in the last row of Table 14.8. The set of average workloads and the 2.5 and 97.5 % forecast limits are in Table 14.9.

A second sensitivity analysis involves using the model N3 for national usage forecasts, without damping as in (14.3). These should be compared to the forecast N1 summarized in Tables 14.3, 14.4 and 14.5 and show a much higher growth rate in demand, especially at older ages. The overall growth rates (2005–2006 vs. 1998–1999) are 57, 51 and 53 % at ages 65–74, 75–84 and 85+, respectively. These compare to growth implied by the National Beds Inquiry of 38, 33 and 27 % (see Table 14.5). The resulting prediction of the total regional flow in 2005–2006 is around 520,000 (see Table 14.10), as compared to around 475,000 using model N1 with damping.

Table 14.10 Forecast trust workload under variant national model

	Credible interval		
Hospital trust	Mean	2.5 %	97.5 %
Southend Acute	89,270	88,400	90,230
Basildon and Thurrock	65,660	64,910	66,330
Forest Health Care	61,280	60,650	61,930
Redbridge Health Care	43,200	42,640	43,890
Havering Hospitals	84,620	83,610	85,410
Newham Health Care	45,180	44,440	45,820
Royal London	88,010	87,210	88,820
Mid Essex	11,480	11,120	11,790
Princess Alexandra	1,076	985	1,162
Homerton	28,380	27,860	28,930
Total	518,156	511,825	524,312

10 Implications for Capacity Planning

Hospital planning focuses especially on providing an adequate number of beds while making efficient use of those available to avoid operational overload (Gallivan et al. 2002). Bed numbers need to reflect total demand (referrals to hospitals) and length of stay distributions and to be sufficient to avoid excess occupancy rates and high refusal rates (referrals not admitted because sufficient beds are not available) and also meet daily and seasonal fluctuations in referrals (Harper and Shahani 2002). The gravity model approach considered in this chapter, with adjustment for supply as in Sect. 8, is only a preliminary to an extended model framework that allows for hospital dynamics. For example, instead of assuming an externally set number of beds as in Sect. 8, one may model bed requirements in terms of length of stay and occupancy rates. Thus a deterministic calculation for beds B involves an average length of stay L times admitted referrals R, and an allowance for bed days not used though notionally available, via the occupancy rate O. Thus,

$$B = R \times L/(365 \times O).$$

More ideally, one should disaggregate bed requirements according to patient groups with similar broad diagnosis or sub-specialty and allowing for admission method (emergency vs. elective in-patient and day case). In this way excessively broad specialty groups (e.g. medical, surgical) are sufficiently differentiated in terms (for example) of different length-of-stay distributions (Harper and Shahani 2002). Denoting groups by g and differentiating also by hospital lead to a deterministic equation that allows also for group-specific occupancy and length of stay, namely,

$$B_{gh} = R_{gh} \times L_{gh}/\left(365 \times O_{gh}\right)$$

Such a model might incorporate forecasts of average lengths of stay by specialty and hospital (i.e. a statistical forecasting approach as in Farmer and Emami 1990) or allow for varying scenarios over future lengths of stay and occupancy under a simulation-based approach. A model might also allow for refused referrals, which will occur when there is under-capacity and excess length of stay (including bed blocking) (Harper and Shahani 2002). Several papers have argued that the impact of length of stay on bed requirements ideally requires a probabilistic approach, whereby simulation of bed requirements using skew length-of-stay densities (e.g. mixed exponential, Weibull, lognormal) improves in predicting bed requirements compared to using average length of stay (Costa et al. 2003; Harrison 2001). This would require forecasts of some or all parameters of such densities.

11 Avenues for Further Research

Much discussion has focussed on national impacts of population ageing on health care costs and demand for medical staff (Metz 1999; Kane 1994). However, impacts of ageing on health care are likely to be spatially differentiated. There have been a number of advances in spatial allocation modelling as applied to health care which facilitate a regional perspective. The gravity model approach has been applied to modelling patient flows between areas and hospitals, but disaggregation to patient demographic groups or specialties has not generally been a feature of such models.

This chapter has developed an estimation and simulation model which takes account of changing age structure in 13 local authority areas and the differential provision of care over hospitals and specialties in a region of two million people in England. The forecasting element of the model is consistent with the assumptions made in the national Hospital Beds Inquiry but introduces an extra specialty dimension and reflects the clear patterning of demand by specialty according to patient age.

Assessment in terms of alternative national forecasts of specialty use by age and in terms of the prior assumptions of the allocation stage was carried out. This shows the forecast of hospital workload to be more sensitive to the method adopted for national forecasting of specialty usage by age than to variant prior assumptions about the allocation model. It may be noted that the model N1 for national specialty usage forecasting comes closer (in the scenario period) to the “guideline” provided by the National Beds Inquiry than the more parameterised model N3. The latter, by contrast, has better fit within the period 1991/1992–1998/1999 and in a one-step-ahead validation using 7 years data to predict 1998/1999. Such contradictions emphasise the possible problems involved in extrapolating a best fit model for current data to the future.

There is obvious scope for further sensitivity analysis using different population projections or different bed numbers at the study hospitals. One might also differentiate hospital episodes by their mode of admission (e.g. emergency vs. elective) or their destination at discharge (e.g. return to home, nursing home, community care). Such disaggregation will add to the potential strategic and planning potential of such models. On the other hand extensive disaggregation may make modelling unwieldy, especially at broader regional or even national scales.

A Bayesian modelling perspective may assist in simulation (i.e. prediction) of new flows in the face of multiple sources of uncertainty and in allowing for information from prior studies (e.g. on the likely degree of distance decay or direction of mass effects) to be used in guiding parameter estimation.

Application of the modelling framework here to other locations might, as in this chapter, involve a full perspective on acute health demand. Alternatively a modified version of it might focus on one type of care (e.g. geriatric or maternity care) and might add certain classifications (e.g. method of admission) and perhaps drop others (e.g. consider only one provider). Despite the potential for such modification, the framework rests essentially on the following: demographic projections for a set of residential areas, projections of future need by area of care (e.g. by specialty or case-mix group) and a flow model that "allocates" patients to providers. As in Sect. 10, bed capacity parameters may also be introduced. The ideal data requirement would be a regional panel of patient flow observations n_{iasth} over providers h and times $t = 1, \ldots, B$. Often such data are not available because of changes in area definitions or because providers are subject to restructuring. In this case national guideline projections of demand in different care areas are one option, as in the work described here. If a full regional panel is available then one may estimate a dynamic flow model, which extends equations (14.7) and (14.9) to include time-varying data on beds B_{ht} and area–hospital specialty mix M_{sht} as well as allow some parameters to be time specific.

References

Chalfin, D., & Nasraway, S. (1994). Preoperative evaluation and postoperative care of the elderly patient undergoing major surgery. *Clinics in Geriatric Medicine, 10*(1), 51–70.

Clarke, M., & Wilson, A. (1984). Health services planning: An outline and an example. In M. Clarke (Ed.), *Planning and analysis in health care systems* (London papers in regional science 13, pp. 22–56). London: Pion.

Congdon, P. (2000). A Bayesian approach to prediction using the gravity model, with an application to patient flow modeling. *Geographical Analysis, 32*(3), 205–224.

Costa, A., Ridley, S., Shahani, A., Harper, P., De Senna, V., & Nielsen, M. (2003). Mathematical modelling and simulation for planning critical care capacity. *Anaesthesia, 58*, 320–327.

Department of Health. (1999). *Modernising the NHS in London, HMSO*. London: HMSO.

Department of Health. (2000). *Shaping the future NHS: Long term planning for hospitals and related services, HMSO*. London: HMSO.

Downing, A., & Wilson, R. (2002). Temporal and demographic variations in attendance at accident and emergency departments. *Emergency Medicine Journal, 19*, 531–535.

Farmer, R., & Emami, J. (1990). Models for forecasting hospital bed requirements in the acute sector. *Journal of Epidemiology and Community Health, 44*, 307–312.

Fries, J., Koop, C., Sokolov, J., Beadle, C., & Wright, D. (1998). Beyond health promotion: Reducing need and demand for medical care. *Health Affairs, 17*(2), 70–84.

Gallivan, S., Utley, M., Treasure, T., & Valencia, O. (2002). Booked inpatient admissions and hospital capacity: Mathematical modelling study. *British Medical Journal, 324*(7332), 280–282.

Gilchrist, R. (1985). Some aspects of modeling operational problems in the National Health Service. *The Statistician, 34*, 209–214.

Harper, P., & Shahani, A. (2002). Modelling for the planning and management of bed capacities in hospitals. *Journal of the Operational Research Society, 53*, 11–18.

Harper, P., Shahania, A., Gallagher, J., & Bowie, C. (2005). Planning health services with explicit geographical considerations: A stochastic location-allocation approach. *Omega, 33*, 141–152.

Harrison, G. (2001). Implications of mixed exponential occupancy patient flow models for health. *Health Care Management Science, 4*, 37–45.

Hull, S., Rees Jones, I., & Moser, K. (1997). Factors influencing the attendance rate at accident and emergency departments in East London: The contributions of practice organization, population characteristics and distance. *Journal of Health Services Research & Policy, 2*, 6–13.

Jones, S., Joy, M., & Pearson, J. (2002). Forecasting demand of emergency care. *Health Care Management Science, 5*, 297–305.

Kane, V. (1994). An older America: Strategic challenges for the acute-care hospital. *Health Manage Quarterly, 16*(4), 9–12.

Lowe, J., & Sen, A. (1996). Gravity model applications in health planning: Analysis of an urban hospital market. *Journal of Regional Science, 3*, 437–461.

Mackay, M., & Millard, P. (1999). Application and comparison of two modelling techniques for hospital bed management. *Australian Health Review, 22*, 118–143.

Metz, D. (1999). Can the impact of ageing on health care costs be avoided? *Journal of Health Services Research & Policy, 4*(4), 249–252.

Millard, P., Mackay, M., Vasilakis, C., & Christodoulou, G. (2000). Measuring and modelling surgical bed usage. *Annals of the Royal College of Surgeons of England, 82*, 75–82.

Milner, P. (1988). Forecasting the demand on accident and emergency departments in health districts in the Trent region. *Statistics in Medicine, 10*, 1061–1072.

Milner, P. (1997). Ten year follow-up of ARIMA forecasts of attendance at accident and emergency departments in the Trent region. *Statistics in Medicine, 16*, 2117–2125.

Puig-Junoy, J., Saez, M., & Martinez-Garcia, E. (1998). Why do patients prefer hospital emergency visits? A nested multinomial logit analysis for patient-initiated contacts. *Health Care Management Science, 1*(1), 39–52.

Sibbritt, D. (1992). Trends and projections for day only admissions in NSW acute hospitals. *Australian Clinical Review, 12*(3), 115–124.

Taket, A., & Mayhew, L. (1981). Interactions between the supply of and demand for hospital services in London. *Omega, 9*, 519–526.

Tebaldi, C., & West, M. (1998). Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association, 93*, 557–573.

Utley, M., Gallivan, S., Davis, K., Daniel, P., Reeves, P., & Worrall, J. (2003). Estimating bed requirements for an intermediate care facility. *European Journal of Operational Research, 150*, 92–100.

Chapter 15
Queueing Analysis in Health Care

Linda Green

Abstract Many organizations, such as banks, airlines, telecommunications companies, and police departments, routinely use queueing models to help determine capacity levels needed to respond to experienced demands in a timely fashion. Though queueing analysis has been used in hospitals and other health care settings, its use in this sector is not widespread. Yet given the pervasiveness of delays in health care and the fact that many health care facilities are trying to meet increasing demands with tightly constrained resources, queueing models can be very useful in developing more effective policies for bed allocation and staffing, and in identifying other opportunities for improving service. Queueing analysis is also a key tool in estimating capacity requirements for possible future scenarios, including demand surges due to new diseases or acts of terrorism. This chapter describes basic queueing models as well as some simple modifications and extensions that are particularly useful in the health care setting, and gives examples of their use. The critical issue of data requirements is also discussed, as well as model choice, model-building, and the interpretation and use of results.

Keywords Queueing • Capacity management • Staffing • Hospitals

1 Introduction

1.1 Why Is Queueing Analysis Helpful in Health care?

Health care is riddled with delays. Almost all of us have waited for days or weeks to get an appointment with a physician or schedule a procedure, and upon arrival we

L. Green (✉)
Graduate School of Business, Columbia University, New York, NY 10027, USA
e-mail: lvg1@columbia.edu

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_15, © Springer Science+Business Media New York 2013

wait some more time until being seen. In hospitals, it is not unusual to find patients waiting for beds in hallways, and delays for surgery or diagnostic tests are common.

Delays are the result of a disparity between demand for a service and the capacity available to meet that demand. Usually this mismatch is temporary and due to natural variability in the timing of demands and in the duration of time needed to provide service. A simple example would be a health care clinic where patients walk in without appointments in an unpredictable fashion and require anything from a flu shot to the setting of a broken limb. This variability and the interaction between the arrival and service processes make the dynamics of service systems very complex. Consequently, it is impossible to predict levels of congestion or to determine how much capacity is needed to achieve some desired level of performance without the help of a queueing model.

Queueing theory was developed by A.K. Erlang in 1904 to help determine the capacity requirements of the Danish telephone system (see Brockmeyer et al. 1948). It has since been applied to a large range of service industries including banks, airlines, and telephone call centers (e.g., Brewton 1989; Stern and Hersh 1980; Holloran and Byrne 1986; Brusco et al. 1995; Brigandi et al. 1994) as well as emergency systems such as police patrol, fire, and ambulances (e.g., Larson 1972; Kolesar et al. 1975; Chelst and Barlach 1981; Green and Kolesar 1984; Taylor and Huxley 1989). It has also been applied in various health care settings as we will discuss later in this chapter. Queueing models can be very useful in identifying appropriate levels of staff, equipment, and beds as well as in making decisions about resource allocation and the design of new services.

Unlike simulation methodologies, discussed in Chap. 9, queueing models require very little data and result in relatively simple formulae for predicting various performance measures such as mean delay or probability of waiting more than a given amount of time before being served. This means that they are easier and cheaper to use and can be more readily used to find “optimal” solutions rather than just estimating the system performance for a given scenario.

Timely access has been identified as one of the key elements of health care quality (Institute of Medicine 2001) and consequently, decreasing delays has become a focus in many health care institutions. Given the financial constraints that exist in many of these facilities, queueing analysis can be an extremely valuable tool in utilizing resources in the most cost-effective way to reduce delays. The primary goal of this chapter is to provide a basic understanding of queueing theory and some of the specific queueing models that can be helpful in designing and managing health care systems. For more detail on specific models that are commonly used, a textbook on queueing theory such as Hall (1991) is recommended.

Before discussing past and potential uses of queueing models in health care, it is important to first understand some queueing theory fundamentals.

1.2 Queueing System Fundamentals

A basic queueing system is a service system where "customers" arrive to a bank of "servers" and require some service from one of them. It is important to understand that a "customer" is whatever entity is waiting for service and does not have to be a person. For example, in a "back-office" situation such as the reading of radiologic images, the "customers" might be the images waiting to be read. Similarly, a "server" is the person or thing that provides the service. So when analyzing delays for patients in the emergency department (ED) awaiting admission to the hospital, the relevant servers would be inpatient beds.

If all servers are busy upon a customer's arrival, they must join a queue. Though queues are often physical lines of people or things, they can also be invisible as with telephone calls waiting on hold. The rule that determines the order in which queued customers are served is called the queue *discipline*. The most common discipline is the familiar first-come, first-served (FCFS) rule, but other disciplines are often used to increase efficiency or reduce the delay for more time-sensitive customers. For example, in an ED, the triage system is an example of a *priority* queue discipline. Priority disciplines may be preemptive or non-preemptive, depending upon whether a service in progress can be interrupted when a customer with a higher priority arrives. In most queueing models, the assumption is made that there is no limit on the number of customers that can be waiting for service, i.e., there is an *infinite waiting room*. This is a good assumption when customers do not physically join a queue, as in a telephone call center, or when the physical space where customers wait is large compared to the number of customers who are usually waiting for service. Even if there is no capacity limit on waiting room, in some cases new arrivals who see a long queue may "balk" and not join the queue. This might happen in a walk-in clinic. A related characteristic that is incorporated in some queueing systems is "reneging" which occurs when customers grow inpatient and leave the queue before being served. An example of this behavior is found in some EDs where the patients who renege are often referred to as "left without being seen".

Finally, queues may be organized in various ways. In most cases, we will consider a *single line* that feeds into all servers. But sometimes each server has his/her own queue as may be the case for a primary care office in which patients have their own physician. This is usually referred to as queues in *parallel*. In other situations, we may want to consider a *network* design in which customers receive service from different types of servers in a sequential manner. For example, a surgical inpatient requires an operating room (OR), then a bed in the recovery unit, followed by a bed in a surgical intensive care unit (ICU), and/or other part of the hospital. However, it might still make sense to analyze a single queue in these situations to determine the capacity requirements of a single type of resource, particularly if there is reason to believe that the resource is a bottleneck.

A queueing model is a mathematical description of a queueing system which makes some specific assumptions about the probabilistic nature of the arrival and service processes, the number and type of servers, and the queue discipline and organization. There are countless variations possible, but some queueing models are more widely used and we will focus on these in this chapter. For these models, as well as many others, there are formulae available that enable the fast calculation of various performance measures that can be used to help design a new service system or improve an existing one.

2 Basic Queueing Principles and Models

Most of queueing theory deals with system performance in *steady-state*. That is, most queueing models assume that the system has been operating with the same arrival, service time and other characteristics for a sufficiently long time that the probability distribution for the queue length and customer delay is independent of time. Clearly, there are many service systems, including health care systems, for which there are time-of-day, day-of-week or seasonality affects. In this section, we will assume that we are looking at systems in steady state and in subsequent sections, we will discuss how to deal with systems that have some time-varying characteristics.

2.1 Delays, Utilization, and System Size

In queueing theory, utilization, defined as the average number of busy servers divided by the total number of servers times 100, is an important measure. From a managerial perspective, utilization is often seen as a measure of productivity and therefore it is considered desirable for it to be high. For example, in hospital bed planning, utilization is called occupancy level and historically, an average hospital occupancy level of 85 % has been used as the minimum level for the states to make a determination under Certificate of Need (CON) regulations that more beds might be needed (see Brecher and Speizio 1995). Since the actual average occupancy rate for nonprofit hospitals has recently been about 66 %, there has been a widely held perception in the health care community that there are too many hospital beds. Largely because of this perception, the number of hospital beds has decreased almost 25 % in the last 20 years.

But determining bed capacity based on occupancy levels can result in very long waiting times for beds (Green 2003). In all queueing systems, the higher the average utilization level, the longer the wait times. However, it is important to note that this relationship is nonlinear. This is illustrated in Fig. 15.1 which shows the fundamental relationship between delays and utilization for a queueing system. There are three critical observations we can make from this figure. First, as average utilization (e.g., occupancy rate) increases, average delays increase at an increasing rate. Second, there is an "elbow" in the curve after which the average delay increases more dramatically in response to even small increases in utilization.

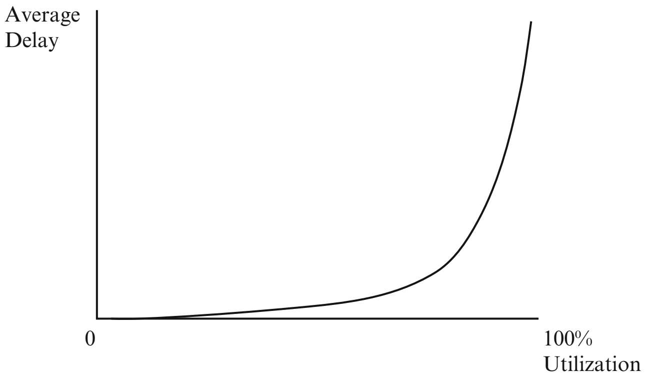


Fig. 15.1 Trade-off between average delay and utilization in a queueing system

Finally, the average delay approaches infinity as utilization approaches one. (It is important to note that this assumes there is no constraint on how long the queue can get and that customers continue to join and remain in the queue.)

The exact location of the elbow in the curve depends upon two critical characteristics of the system: variability and size. Variability generally exists in both the time between arrivals and the duration of service times and is usually measured by the ratio of the standard deviation to the mean, called the coefficient of variation (CV). The higher the degree of variability in the system, the more to the left the elbow will be so that delays will be worse for the same utilization level. System size is defined as the ratio of the average demand over the average service time, which is a determinant of the number of servers needed. The larger the system, the closer the elbow will be to 100 %, so that delays will be smaller for the same utilization level.

These basic queueing principles have several important implications for planning or evaluating capacity in a service system. First, the average total capacity, defined as the number of servers times the rate at which each server can serve customers, must be strictly greater than the average demand. In other words, unless average utilization is strictly *less than* 100 %, the system will be "unstable" and the queue will continue to grow. Though this fact may appear counter-intuitive on the surface, it has been well known by operations professionals for decades. So if an emergency room has ten patients arriving per hour on average and each health care provider (physician or physician assistant) can treat two patients per hour, a minimum of six providers is needed. (Of course, in many contexts, if arrivals see a long queue they may not join it or they may renege after waiting a long time. If so, the system may stabilize even if the average demand exceeds the average capacity.) Second, the smaller the system, the longer the delays for a given utilization level. In other words, queueing systems have economies of scale so that, for example, larger hospitals can operate at higher utilization levels than smaller ones yet maintain similar levels of congestion and delays. Finally, the greater the variability in the service time (e.g., length-of-stay), the longer the delays at a given utilization level. So a clinic or physician office that specializes in for example vision testing or mammography will experience shorter patient waits than a university based clinic of the same size and with the same provider utilization that treats a broad variety of illnesses and injuries. These properties will be more specifically illustrated when we discuss applications of queueing models.

2.2 Some Simple But Useful Queueing Models

2.2.1 The Poisson Process

In specifying a queueing model, we must make assumptions about the probabilistic nature of the arrival and service processes. The most common assumption to make about arrivals is that they follow a *Poisson* process. The name comes from the fact that the number of arrivals in any given time period has a Poisson distribution. So if $N(t)$ is the number of arrivals during a time period of duration t and $N(t)$ has a Poisson distribution,

$$\text{Probability}\{N(t) = n\} = e^{-\lambda t}(\lambda t)^n/n!$$

where λ is called the *rate* and is the expected number of arrivals per unit time. For example, if $\lambda = 10$ customers per hour, then the expected number of arrivals in any 60 min interval is 10 and the expected number to arrive in a 15 min interval is 2.5. Notice that these are averages so that λ need not have an integer value. Another way to characterize the Poisson process is that the time between consecutive arrivals, called the interarrival time, has an *exponential* distribution. So if IA is the interarrival time of a Poisson process with rate λ,

$$\text{Probability}\{\text{IA} \leq t\} = 1 - e^{-\lambda t}$$

and $1/\lambda$ is the average time between arrivals.

An important property of the exponential distribution is that it is "memoryless". This means that the time of the next arrival is independent of when the last arrival occurred. This property also leads to the fact that if the arrival process is Poisson, the number of arrivals in any given time interval is independent of the number in any other nonoverlapping time interval. Conversely, it can be shown analytically that if customers arrive independently from one another, the arrival process is a Poisson process. For this reason, the Poisson process is considered the most "random" arrival process.

In determining whether the Poisson process is a reasonable model for arrivals in a specific service system, it is useful to consider its three defining properties:

1. Customers arrive one at a time.
2. The probability that a customer arrives at any time is independent of when other customers arrived.
3. The probability that a customer arrives at a given time is independent of the time.

In most contexts, customers generally do arrive one at a time. Though there may be events, such as a major accident, that trigger multiple simultaneous arrivals, this is likely to be an exceptional circumstance which will not significantly affect the usefulness of this modeling assumption. Intuitively, the second property is also often a reasonable assumption. For example, in an emergency room, where the population of potential patients is very large, it is unlikely that someone arriving with a broken arm has anything to do with someone else's injury or illness, or that the fact that the number of patients who arrived between 9 a.m. and 10 a.m. was four provides information about the number of patients that are likely to arrive between 10 a.m. and 11 a.m. Again, there may be occasional exceptions, such as a flu outbreak, for which there is an exogenous factor responsible for generating multiple arrivals over a time period. However, this assumption is still likely to be a reasonable one in most situations. The third property may be more suspect. More typically, the average arrival rate varies over the day so that, e.g., it is more likely for an arrival to occur in the morning than in the middle of the night. Certain days of the week may be busier than others as well. However, we may be able to use the standard Poisson process as a good model for a shorter interval of time during which the arrival rate is fairly constant. We will discuss this in more detail in a subsequent section.

So the assumption of a Poisson process will generally be a good one when the three properties above are a reasonable description of the service system in question. However, it is possible to perform more rigorous tests to determine if it is a good fit. The simplest tests are based on the relationship of the standard deviation to the mean of the two distributions involved in the Poisson process. Since the variance (square of the standard deviation) of the Poisson distribution is equal to its mean, we can examine the number of arrivals in each fixed interval of time (e.g., 30 min) and determine whether the ratio of the mean to the variance is close to one. Alternatively, since the exponential distribution is characterized by its standard deviation being equal to its mean, we can look at the interarrival times and compute the ratio of the standard deviation to the mean to see if it is close to one. Hall (1991) describes goodness of fit tests in greater detail.

Many real arrival and demand processes have been empirically shown to be very well approximated by a Poisson process. Among these are demands for emergency services such as police, fire and ambulance, arrivals to banks and other retail establishments, and arrivals of telephone calls to customer service call centers. Because of its prevalence and its assumption of independent arrivals, the Poisson process is the most commonly used arrival process in modeling service systems. It is also a convenient assumption to make in terms of data collection since it is characterized by a single parameter—its rate λ. In health care, the Poisson process has been verified to be a good representation of unscheduled arrivals to various parts of the hospital including ICUs, obstetrics units and EDs (Young 1965; Kim et al. 1999; Green et al. 2005).

2.2.2 The M/M/s Model

The most commonly used queueing model is the *M/M/s* or *Erlang delay* model. This model assumes a single queue with unlimited waiting room that feeds into *s* identical servers. Customers arrive according to a Poisson process with a constant rate, and the service duration (e.g., LOS or provider time associated with a patient) has an exponential distribution. (These two assumptions are often called Markovian, hence the use of the two "M's" in the notation used for the model.)

One advantage of using the *M/M/s* model is that it requires only three parameters and so it can be used to obtain performance estimates with very little data. Given an average arrival rate, λ, an average service duration, $1/\mu$, and the number of servers, s, easy-to-compute formulae are available to obtain performance measures such as the probability that an arrival will experience a positive delay, p_{D}, or the average delay, W_q:

$$p_{\mathrm{D}} = 1 - \sum_{n=0}^{s-1} p_n \tag{15.1}$$

$$W_q = p_{\mathrm{D}}/[(1 - \rho s\mu)] \tag{15.2}$$

for

$$\rho = \lambda/s\mu \tag{15.3}$$

and

$$p_n = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} p_0 & (1 \leq n \leq s) \\ \dfrac{\lambda^n}{s^{n-s} s! \mu^n} p_0 & (n \geq s) \end{cases} \tag{15.4}$$

where

$$p_0 = \left[\sum_{n=0}^{s-1} \frac{(\rho s)^n}{n!} + \frac{\rho^s s^{s+1}}{s!(s - \rho s)} \right]^{-1} \quad \rho < 1 \tag{15.5}$$

Note that ρ is the average utilization for this queueing system and the equation is only valid when the utilization is strictly less than one. Also note that average delay increases as utilization approaches one. These quantitative observations support the discussion of utilization and delays in the previous section.

Many other measures of performance can be calculated as well and many of the formulae for both the *M/M/s* and other common queueing models are available in software packages or are easily programmable on spreadsheets. One common performance constraint is often referred to as the *service level*—a requirement that *x* % of customers start service within *y* time units. For example, many customer call centers have a target service level that 85 % of calls be answered within 20 s. The delay is always measured from the time of the demand for service (e.g., patient registered in the ED) to the time at which service begins (e.g., a provider is available to treat that patient). It is important to note that the model's delay predictions pertain only to waiting times due to the unavailability of the server.

2.2.3 Some Useful Extensions of the M/M/s Model

There are several variations on the basic *M/M/s* queueing model. One important one for many health care organizations is the *M/M/s* with priorities. While the fundamental model assumes that customers are indistinguishable and are served FCFS, the priority model assumes that customers have differing time-sensitivities and are allocated to two or more service classes $i = 1, 2, \ldots, N$, and that customers are served in priority order with 1 being the highest priority and *N* the lowest. Within any given class, customers are served FCFS. But when there is a queue and a server becomes available, a customer belonging to class *i* will be served only if there are no waiting customers of class $1, \ldots, i - 1$. A priority queueing model would be appropriate if a facility is interested in identifying the capacity needed to assure a targeted service level for the highest priority customers. For examples, in an ED, while many arriving patients would not incur any particular harm if they had to wait more than an hour to be seen by a physician, some fraction, who are emergent or urgent, need a physician's care sooner to prevent serious clinical consequences. In this case, a priority queueing model could be used to answer a question like: How many physicians are needed to assure that 90 % of emergent and urgent patients will be seen by a physician within 45 min?

There are two types of priority queueing disciplines: preemptive and non-preemptive. In the preemptive model, if a higher priority customer arrives when all servers are busy and a lower priority customer is being served, the lower priority customer's service will be interrupted (preempted) so that the higher priority customer can begin service immediately. The preempted customer must then wait for another server to become free to resume service. In the non-preemptive model, new arrivals cannot preempt customers already in service. While priority queueing models are usually either purely preemptive or non-preemptive, it is possible to model a service system that has both preemptive and non-preemptive customer classes. This might be appropriate for a hospital ED where the normal triage system which classifies patients as emergent, urgent or nonurgent is usually assumed to be non-preemptive, but will use a preemptive discipline for certain urgent patients whose conditions are extremely time-sensitive, such as stroke victims. In addition to the usual input parameters for the *M/M/s* model, priority models also require the fraction of customers in each of the priority classes.

Another common variant of the *M*/*M*/*s* model assumes a finite capacity $K \geq s$ and is notated as *M*/*M*/*s*/*K*. In this model, if a customer arrives when there are *K* customers already in the system (being served and waiting), the customer cannot join the queue and must leave. A common application of this would be a telephone trunk line feeding into a call center. Such a system has a finite number of spaces for calls being served or on hold and when a new call comes in and all the spaces are already taken, the new arrival hears a busy signal and hangs up. A similar phenomenon might occur in a walk-in health clinic which has a waiting room with a fixed number of seats. Though some patients may choose to wait even if there is no seat available upon arrival, many patients may leave and try to return at a less busy time. Customers who are "blocked" from joining the queue are called "lost" and may show up again or never return. In these types of systems, queueing analysis might be used to help determine how large the waiting or holding area should be so that the number of customers who are blocked is kept to an acceptably low level.

A specific special case of these finite capacity models is the one where $K = s$ so that there is no waiting room for those who arrive when all servers are busy. These are called pure "loss" models and they are often used to analyze service systems in which it is considered either impractical or very undesirable to have any customers wait to begin service. For example, Shmueli, Sprung and Kaplan (2003) used a loss model to analyze the impact of various admissions policies to ICU facilities.

2.3 The *M*/*G*/*1* and *G*/*G*/*s* Models

An important characteristic of the exponential distribution used in the *M*/*M*/*s* is that the standard distribution equals the mean and so the CV of the service time equals one. If the actual CV of service is a bit less than or greater than one, the *M*/*M*/*s* will still give good estimates of delay. However, if the CV is substantially different than one, the *M*/*M*/*s* may significantly underestimate or overestimate actual delays. (Recall that if variability is lower, the model will overestimate delays while the converse is true if variability is greater.) In this case, if the arrival process is Poisson, and there is only one server, the average delay can still be calculated for any service distribution through use of the following formula for what is known as the *M*/*G*/*1* system:

$$W_q = [\lambda\rho/(1-\rho)]\left[\left(1+\mathrm{CV}^2(S)\right)/2\right] \tag{15.6}$$

where $\mathrm{CV}^2(S)$ is the square of the coefficient of variation of the service time. Clearly, this formula requires knowledge of the standard deviation of the service time in addition to the mean in order to compute $\mathrm{CV}^2(S)$. This formula also illustrates the impact of variability on delays. Notice that, as mentioned previously, as the coefficient of variation of the service time increases, so does the average delay.

Though there are no exact formula for non-Markovian multi-server queues, there are some good, simple approximations. One such approximation (Allen 1978) is given by:

$$W_q = W_{q,M/M/s}\left[\mathrm{CV}^2(A) + \mathrm{CV}^2(S)\right]/2 \tag{15.7}$$

where $\mathrm{CV}^2(A)$ is the square of the coefficient of variation of the arrival time and $W_{q,M/M/s}$ is the expected delay for an *M*/*M*/*s* system, (15.2). So this formula requires the standard deviation of the interarrival time as well and again demonstrates that more variability results in longer delays.

3 Analyses of Fixed Capacity: How Many Hospital Beds?

Many resources in health care facilities have a fixed capacity over a long period of time. These are usually “things” rather than people: beds, operating rooms, imaging machines, etc. Queueing models are not always appropriate for analyzing such resources. In particular, if the patients for a resource are scheduled into fixed time slots, there is little or no likelihood of congestion unless patients routinely come late or the time slots are not large enough to accommodate most patients. An example of this would be a magnetic resonance imaging (MRI) facility which is only used by scheduled outpatients. It should be noted that the use of an appointment system can be an effective way to minimize or eliminate variability in the arrival stream of a service system and therefore minimize delays. See Chapter for more on appointment systems.

However, the difficulty of managing many health care facilities is that the demand for resources is unscheduled and hence random, yet timely care is important. This is the case for many parts of a hospital that deal primarily with nonelective admissions. In these cases, queueing models can be very helpful in identifying long-term capacity needs.

3.1 Applying the M/M/s Model

To illustrate the use of a queueing model for evaluating capacity, consider an obstetrics unit. Since it is generally operated independently of other services, its capacity needs, e.g., number of postpartum beds, can be determined without regard to other parts of the hospital. It is also one for which the use of a standard *M*/*M*/*s* queueing model is quite good. Most obstetrics patients are unscheduled and the assumption of Poisson arrivals has been shown to be a good one in studies of unscheduled hospital admissions (Young 1965). In addition, the CV of length of

Table 15.1 Probability of (Delay) and utilization for obstetrics unit

No. beds	Pr (Delay)	Utilization
45	0.666	0.953
46	0.541	0.933
47	0.435	0.913
48	0.346	0.894
49	0.272	0.875
50	0.212	0.858
51	0.163	0.841
52	0.124	0.825
53	0.093	0.809
54	0.069	0.794
55	0.051	0.78
56	0.037	0.766
57	0.026	0.753
58	0.018	0.74
59	0.013	0.727
60	0.009	0.715
61	0.006	0.703
62	0.004	0.692
63	0.003	0.681
64	0.002	0.67
65	0.001	0.66

stay is typically very close to 1.0 (Green and Nguyen 2001) satisfying the service time assumption of the $M/M/s$ model.

A queueing model may be used either descriptively or prescriptively. As an example of the descriptive case, we can take the current operating characteristics of a given obstetrics unit: arrival rate, average LOS, and number of beds; and use these in (15.1) to determine the probability that an arriving patient will not find a bed available. Let us assume that Big City Hospital's obstetrics unit has an average arrival rate of $\lambda = 14.8$ patients per day, an average LOS of $1/\mu = 2.9$ days, and $s = 56$ beds. Then the $M/M/s$ formula for probability of delay (15.1) produces an estimate of approximately 4 %. To use the $M/M/s$ prescriptively to find the minimum number of beds needed to attain a target probability of delay, we can enter (15.1) in a spreadsheet and produce a table of results for a broad range of bed capacities to find the one that best meets the desired target. Table 15.1 is a partial table of results for our example obstetrics unit.

Though there is no standard delay target, Schneider (1981) suggested that given their emergent status, the probability of delay for an obstetrics bed should not exceed 1 %. Applying this criterion, Table 15.1 indicates that this unit has at least 60 beds. Table 15.1 also shows the utilization level for each choice of servers and that at 60 beds, this level is 71.5 %. This is what hospitals call the average occupancy level and it is well below the 85 % level that many hospitals and health care policy officials consider the minimum target level. It is also below the maximum level of 75 % recommended by the American College of Obstetrics

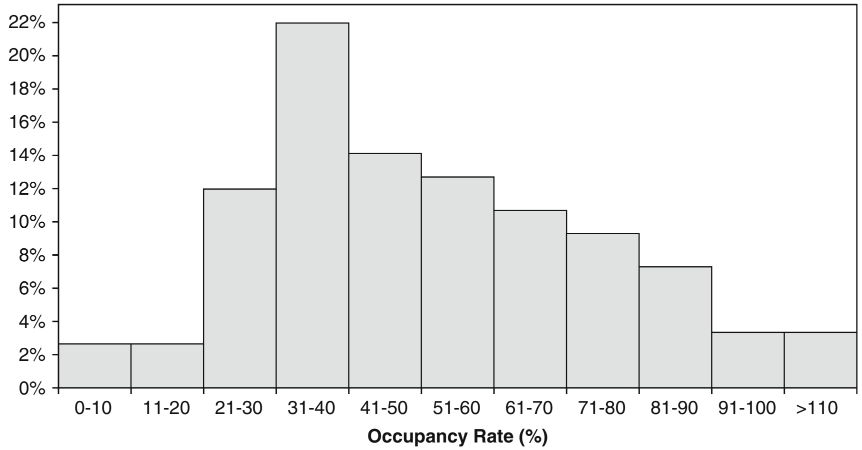


Fig. 15.2 Average occupancy rates of New York State maternity units, 1997

and Gynecology (ACOG) to assure timely access to a bed (Freeman and Poland 1997). So does this example show that as long as an obstetrics unit operates below this ACOG occupancy level of 75 %, the fraction of patients who will be delayed in getting a bed will be very low?

3.2 The Problem with Using Target Occupancy Levels

Hospital capacity decisions traditionally have been made, both at the government and institutional levels, based on target occupancy levels—the average percentage of occupied beds. Historically, the most commonly used occupancy target has been 85 %. Estimates of the number of "excess" beds in the USA, as well as in individual states and communities, usually have been based on this "optimal" occupancy figure (Brecher and Speizio 1995, p. 55). In addition, low occupancy levels are often viewed as indicative of operational inefficiency and potential financial problems. So hospital administrators generally view higher occupancy levels as desirable. However, as we saw previously in this chapter, higher occupancy levels result in longer delays and so basing capacity on target occupancy levels can lead to undesirable levels of access for patients.

In the basic *M*/*M*/*s* model is used to demonstrate the implications of using target occupancy levels to determine capacity in both obstetrics and ICU units in New York State. Figure 1 from that paper (shown below as Fig. 15.2) shows the distribution of average occupancy rates for 148 obstetrics units in New York State for 1997. These data, representing nearly all obstetrics units in New York, were obtained from Institutional Cost Reports (ICRs), and unlike most other published data, reflect staffed beds rather than certified beds. The graph shows that many maternity units had low average occupancy levels with the overall

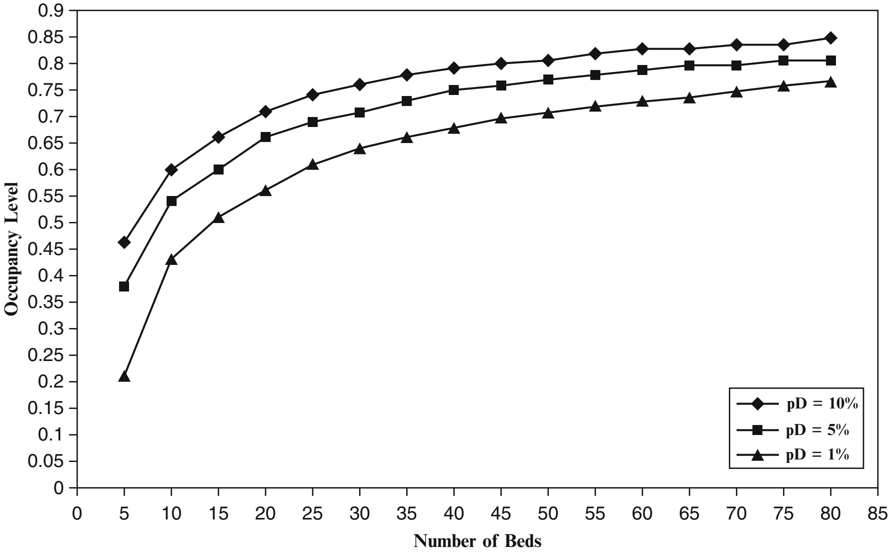


Fig. 15.3 Probability of Delay (p_D) by occupancy and size

average occupancy level for the study hospitals was only 60 %, which, based on the ACOG standard, would imply significant excess capacity. Applying this 75 % standard to the 1997 data, 117 of the 148 New York state hospitals had "excess" beds, while 27 had insufficient beds.

However, if one considers a bed delay target as a more appropriate measure of capacity needs, the conclusions can be quite different. Now the number of beds in each unit becomes a major factor since, for a given occupancy level, delays increase as unit size decreases. While obstetrics units usually are not the smallest units in a hospital, there are many small hospitals, particularly in rural areas, and the units in these facilities may contain only five to ten beds. Of the New York state hospitals considered here, more than 50 % had maternity units with 25 or fewer beds.

In the *M*/*M*/*s* model, probability of delay is a function of only two parameters: s and ρ, which in our context is the number of beds and occupancy level. Each of the three curves shown in Fig. 15.3 represents a specific probability of delay as a function of these two variables as generated by (15.1). Thus, using the unit size and occupancy level reported on the ICR report for a given maternity unit, we can determine from this figure if the probability of delay meets or exceeds any one of these targets. For example, if a maternity unit has 15 beds and an occupancy level of 45 %, it would fall below all three curves and hence have a probability of delay less than 0.01 or 1 %, meeting all three targets.

Doing this for every hospital in the database, 30 hospitals had insufficient capacity based on even the most slack delay target of 10 %. (It is interesting to note that two of the hospitals that would be considered over utilized under the 75 % occupancy standard had sufficient capacity under this delay standard.) Tightening the probability of delay target to 5 %, yields 48 obstetrics units that do not meet this standard. And adopting a maximum probability of delay of 1 % as was suggested in the only publication identified as containing a delay standard for obstetrics beds (Schneider 1981), results in 59, or 40 %, of all New York state maternity units with insufficient capacity.

How many hospitals in New York State had maternity units large enough to achieve the ACOG-suggested 75 % occupancy level and also meet a specified probability of delay standard? Using Fig. 15.3, we see that for a 10 % target, an obstetrics unit would need to have at least 28 beds, a size that exists in only 40 % of the state hospitals. For a 5 % standard, the minimum number of beds needed is 41, a size achieved in only 14 % of the hospitals; for a 1 % standard, at least 67 beds are needed, leaving only 3 of the 148 or 2 % of the hospitals of sufficient size.

3.3 Choosing a Delay Standard

As the previous analysis illustrates, the number of required beds can change substantially depending upon what level of delay is considered tolerable. There is no single right choice and in choosing a delay standard, several factors are relevant.

First, what is the expected delay of those patients who experience a delay? This performance measure can be easily calculated once both the probability of delay (15.1) and the average or mean delay (15.2) are known. Specifically,

$$\text{Expected delay of delayed customers} = W_q/p_{\mathrm{D}} \tag{15.8}$$

So returning to our obstetrics example above, Table 15.1 shows that the average delay is 0.008 days (note that since the input was expressed in days, so is the output) which multiplying by 24 gives us 0.19 h. So dividing this by the probability of delay of 0.04 results in an expected delay for delayed patients of about 4.75 h. This may indicate that the probability of delay standard should be lower. This, of course, should be considered in light of what this level of congestion means for the particular hospital.

What are the possible consequences of congestion? In the obstetrics case, while patients in some hospitals remain in the same bed through labor, delivery, recovery, and postpartum, in most maternity units, there are separate areas for some or all of these stages of birth. Therefore, a delay for an obstetrics bed often means that a postpartum patient will remain in a recovery bed longer than necessary. This, of course, may cause a backup in the labor and delivery areas so that newly arriving patients may have to wait on gurneys in hallways or in the emergency room. Some hospitals have overflow beds in a nearby unit that is opened (staffed) when all regular beds are full. (This is likely the case for the five hospitals that reported average occupancy levels exceeding 100 %.) While these effects of congestion likely pose no medical threat for most patients who experience normal births, there could be adverse clinical consequences in cases in which there are complications.

In particular, whether patients are placed in hallways or overflow units, the nursing staff is likely to be severely strained, thereby limiting the quantity and quality of personal attention. Even if a hospital is able to obtain additional staffing, it is usually by using agency nurses who are more expensive and not as familiar with the physical or operating environment, thereby jeopardizing quality of patient care. In addition, telemetry devices, such as fetal monitors that are usually in labor and delivery rooms, may be unavailable in other locations, thus compromising the ability to monitor often need the resources of an intensive care vital body functions of both mother and baby. Finally, it is worth noting that such results of congestion may negatively affect patients' perceptions of service quality.

Of course, all major capacity decisions need to be made in light of financial constraints, competing demands, and predictions concerning future demands for the service.

3.4 Planning for Predictable Changes in Demand

When making capacity decisions about resources that will be used over several years, it is clearly necessary to consider how conditions may change over that period of time. So in determining the choice of arrival rate or average LOS for a queueing analysis of a hospital unit, it would be important to engage in analyses and discussion to gauge how these parameters may change and then run the model to determine the sensitivity of capacity levels to these changes.

However, what may not be so obvious is the need to consider the changes in the arrival rate that are likely to occur on a regular basis due to predictable day-of-week or time-of-year patterns. For example, obstetrics units often experience a significant degree of seasonality in admissions. An analysis performed on data from a 56-bed maternity unit at Beth Israel Deaconess Hospital in Boston (Green and Nguyen 2001) revealed that the average occupancy levels varied from a low of about 68 % in January to about 88 % in July. As indicated by Fig. 15.4, the *M*/*M*/*s* model estimate of the probability of delay of getting a bed for an obstetrics patient giving birth in January is likely to be negligible with this capacity. However, in July, the same model estimates this delay to be about 25 %. And if, as is likely, there are several days when actual arrivals exceed this latter monthly average by say 10 %, this delay probability would shoot up to over 65 %. The result of such substantial delays can vary from backups into the labor rooms and patients on stretchers in the hallways to the early discharge of patients. Clearly, hospitals need to plan for this type of predictable demand increase by keeping extra bed capacity that can be used during peak times, or by using "swing" beds that can be shared by clinical units that have countercyclical demand patterns.

Most hospital units experience different arrival rates for different days of the week. For example, in one surgical intensive care unit, the average admissions per day over a 6 month period varied from a low of 1.44 for Sundays to a high of 4.40 for Fridays. Using the average arrival rate over the week of 3.34 in a queueing

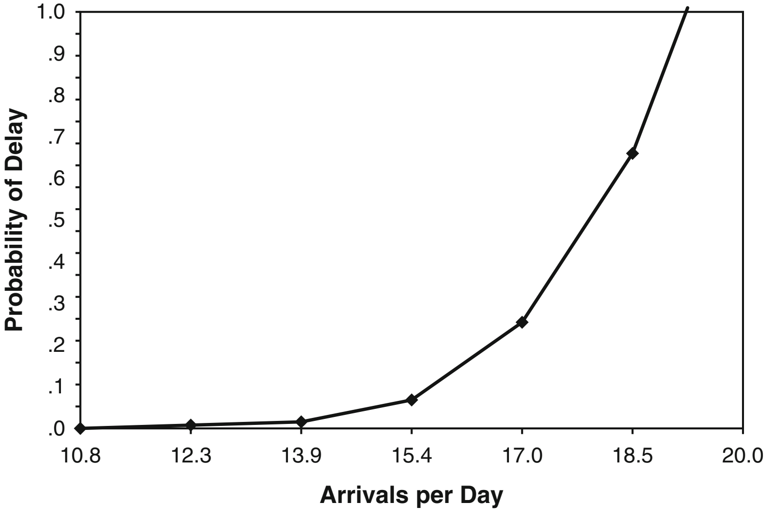


Fig. 15.4 Probability of Delay as a function of arrivals per day for a 56-bed obstetrics unit

model would indicate that given the 12 bed capacity of this unit, the probability of delay for a bed was about 39 %, indicating serious congestion. However, this is very misleading because delays will be significantly greater in the middle of the week and quite small earlier in the week due to the large differences in the admissions rates (Green and Nguyen 2001). This illustrates a situation in which a steady-state queueing model is inappropriate for estimating the magnitude and timing of delays and for which a simulation model will be far more accurate.

It is important to note that while in the obstetrics unit case, most arrivals are unscheduled and cannot be controlled, in the surgical unit case, the converse is true since most surgeries are elective. So while there is little that can be done to minimize the seasonal variability in arrivals for the former, the intra-week variability of the surgical unit could be reduced by adjusting the scheduling of surgeries so as to smooth out the demand over the week. This would result in higher average levels of bed occupancy and shorter delays for beds.

3.5 Using Queueing Models to Quantify the Benefits of Flexibility

Health care facilities often have to make a choice as to the extent to which resources should be dedicated to specific patient types. For example, should there be a imaging facility just for the use of inpatients, or for emergency patients? Should there be a “fast-track” unit in the emergency room to deal with simpler, nonurgent cases. How many distinct clinical service units should be used for hospital inpatients? In many of these situations, a queueing analysis can be useful in evaluating the potential trade-offs between more flexible and more specialized facilities.

For example, seriously ill patients arriving to a hospital ED often experience serious delays in being admitted due to highly variable patient demands and insufficient inpatient bed capacity. Yet hospitals are often reluctant or unable to add capacity because of cost pressures, regulatory constraints, or a shortage of appropriate personnel. This makes it extremely important to use existing capacity most efficiently. Increasing bed flexibility can be a key strategy in alleviating congestion. For example, hospitals vary in the degree to which they segregate patients by diagnostic type. While all hospitals have separate units for pediatrics, obstetrics and psychiatric patients, some also have distinct units for clinical services such as cardiology, neurology, oncology, urology, neurosurgery, etc. Other hospitals may make no such distinctions and simply designate all of these as medical/surgical beds. What are the implications of these differing bed assignment policies on delays for beds?

As mentioned in Sect. 2.1, service systems have economies of scale and so in general, the less specialized the beds, the larger the pool of beds that can be used for any type of patient, and therefore fewer beds should be needed to achieve a given standard of delay. In other words, if one hospital has 100 general medical/surgical beds, and another has the same 100 beds, but allocated into ten distinct clinical services, each of which can only be used for patients falling into the appropriate category, the second hospital will likely have considerably longer delays for beds (which usually show up as longer stays in the ED) and lower average occupancy levels than the first. This is pretty clear once you consider that by creating separate categories of beds, there is the possibility of patients waiting for beds even when beds are available if they are the “wrong” kind. This also happens when beds are distinguished by capability, for example telemetry beds.

Clearly, there are many instances in which there are compelling clinical and/or managerial reasons for maintaining particular patient types in specialized units. From a medical perspective, there may be benefits derived from having patients clustered by diagnostic categories in dedicated units managed and staffed by specialized nurses. These include shorter LOS, fewer adverse events and fewer readmits. Yet many hospital managers believe that nurses can be successfully cross-trained and that increasing bed flexibility is ultimately in the best interests of patients by increasing speedy access to beds and minimizing the number of bed transfers. By incorporating waiting times, percentage of “off-placements” and the effects on LOS, queueing models can be used to better evaluate the benefits of greater versus less specialization of beds or any other resource. This would be done by simply modeling the general-use unit as a single multi-server queueing system fed and comparing the results to those from modeling each distinct service as an independent queue. In the latter case, the overall patient delay can be obtained from an arrival rate weighted average of the individual queue delays (see e.g., Green and Nguyen 2001).

4 Analyses of Flexible Capacity: Determining Staffing Levels to Meet Time-Varying Demands

As mentioned previously, health care facilities generally experience very different levels of demand over the day, over the week, and even over the year. Many facilities adjust their staffing—e.g., physicians, nurses, technicians, housekeeping staff—in order to respond to the demands in a timely fashion at minimal cost. This is often done without the help of a quantitative model and can lead to an inefficient and ineffective allocation of resources. Here we use the example of determining physician staffing levels in an ED to illustrate how queueing models can be used to improve performance in these types of situations.

4.1 Data Collection and Model Choices

In order to use a queueing model to determine how to adjust staffing to meet time-varying demands, it is first necessary to collect fairly detailed data on the volume of demand that must be handled by that staff by time-of-day and day-of-week. In collecting demand data, the goal is twofold. First, and most obviously, the data will be used to parameterize the queueing model. However, before that can be done, it must first be determined how many staffing models are needed. That is, will staffing be identical for all days of the week or vary from day to day? For example, in a study conducted in the ED of a mid-size urban hospital in New York City (Green et al. 2005), the overall volume varied from a low of 63 patients per day on Saturdays to a high of 72 per day on Monday. This degree of variation indicated that the then-current policy of identical staffing levels for all days of the week was likely suboptimal. However, it was deemed impractical to have a different provider schedule every day and so it was decided to use queueing analyses to develop two schedules: weekday and weekend. This required aggregating ED arrival data into these two groups. For each, demand data was then collected for each hour of the day using the hospital's admissions database to understand the degree of variation over the day (see Fig. 15.5). This level of detail also allows for the use of queueing analysis to determine the impact of different shift starting times on delays and/or staffing levels.

A queueing model also requires an average provider service time per patient, which must include the times of all activities related to a patient. In the ED, these activities include direct patient care, review of X-rays and lab tests, phone calls, charting, and speaking with other providers or consults. In many, if not most, hospitals, these data are not routinely collected. At the time of the study, provider service times were not recorded and had to be estimated indirectly from direct observation and historical productivity data.

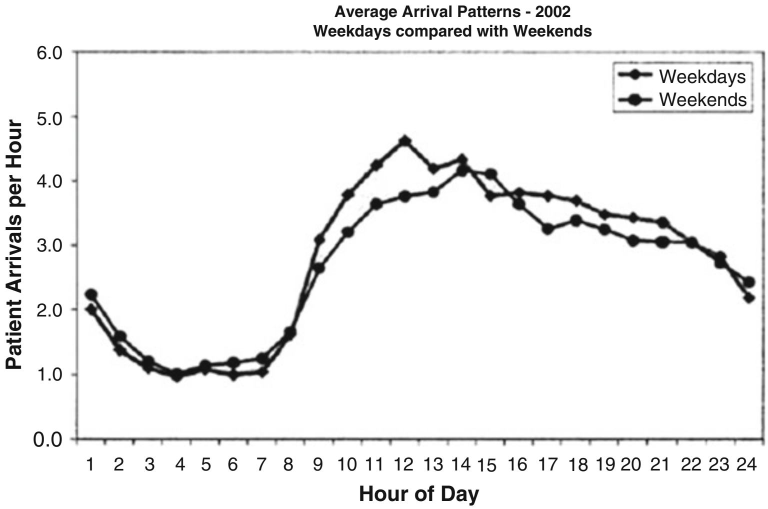


Fig. 15.5 Average arrival patterns for the Allen Pavilion

4.2 Constructing the Queueing Models

Since the *M*/*M*/*s* model assumes that the arrival rate does not change over the day, actual service systems that have time-varying demands typically use this model as part of a *SIPP* (stationary independent period-by-period) approach to determine how to vary staffing to meet changing demand. The *SIPP* approach begins by dividing the workday into staffing periods, e.g., 1, 2, 4, or 8 h. Then a series of *M*/*M*/*s* models are constructed, one for each staffing period. Each of these period-specific models is independently solved for the minimum number of servers needed to meet the service target in that period. The service target might be a desired maximum mean delay or probability of delay standard. However, recent research has shown that the *SIPP* approach is often unreliable, particularly when average service times are 30 min or more, and that a simple modification, called *Lag SIPP*, is often more effective in identifying staffing levels that achieve the desired performance standard (Green et al. 2001). This is because in many service systems with time-varying arrival rates, the time of peak congestion significantly lags the time of the peak in the arrival rate (Green et al. 1991). While the standard *SIPP* approach ignores this phenomenon, the *Lag SIPP* method incorporates an estimation of this lag and thus does a better job of identifying staffing levels to limit delays. For the *M*/*M*/*s* model, the lag can be well approximated by an average service time.

4.3 Choosing a Delay Standard and Applying the Queueing Results

In our ED physician staffing study, the *Lag SIPP* approach was applied by first advancing the arrival rate curve by our estimate of the average physician time per patient, 30 min. We then constructed a series of $M/M/s$ models for each 2-h staffing interval, using the average arrival rate for each based on the time-advanced curve and the average 30 min service time. The delay standard we choose was that no more than 20 % of patients wait more than 1 h before being seen by a provider. The use of 1 h is consistent with the time standards associated with emergent and urgent patient groups used in the National Hospital Ambulatory Medical Care Survey (McCaig and Burt 2004). The 20 % criterion reflects the approximate percentage of nonurgent arrivals at the study institution.

The modeling results gave the number of ED physicians needed in each of the 2-h staffing intervals to meet the delay standard. In total, 58 physician-hours were needed on weekdays to achieve the desired service standard, which represented an increase of 3 h over the existing staffing level of 55 h. Model runs for the weekend indicated that the target performance standard could be achieved with a total of 53 provider-hours. In both these cases, the queueing analyses suggested that some physician hours should be switched from the middle of the night to much earlier in the day. A more subtle change suggested by the model was that the increase in staffing level to handle the morning surge in demand needed to occur earlier than in the original schedule. Though resource limitations and physician availability prevented the staffing suggested by the queueing analyses from being implemented exactly, the insights gained from these analyses were used to develop new provider schedules. More specifically, as a result of the analyses one physician was moved from the overnight shift to an afternoon shift, 4 h were moved from the weekends and added to the Monday and Tuesday afternoon shifts (since these were the two busiest days of the week) and a shift that previously started at noon was moved to 10 a.m. These changes led to shorter average delays and a reduced fraction of patient that left before being seen by a physician.

5 Using Queueing Models to Improve Health care Delivery: Opportunities and Challenges

As this chapter has illustrated, service systems are very complex due to both predictable and unpredictable sources of variability in both the demands for service and the time it takes to serve those demands. In health care facilities, decisions on how and when to allocate staff, equipment, beds, and other resources in order to minimize delays experienced by patients are often even more difficult than in other service industries due to cost constraints on the one hand and the potentially serious adverse consequences of delays on the other hand. Therefore, it is imperative that these decisions should be as informed as possible and rely upon the best methodologies available to gain insights into the impact of various alternatives.

Queueing theory is a very powerful and very practical tool because queueing models require relatively little data and are simple and fast to use. Because of this simplicity and speed, they can be used to quickly evaluate and compare various alternatives for providing service. Beyond the most basic issue of determining how much capacity is needed to achieve a specified service standard, queueing models can also be useful in gaining insights on the appropriate degree of specialization or flexibility to use in organizing resources, or on the impact of various priority schemes for determining service order among patients.

On the other hand, though queueing models do not require much data, the type of operational data needed as input to a queueing model is often unavailable in health care settings. Specifically, though demand or arrival data are often recorded, service times are usually not documented. So a queueing analysis might require a data collection effort to estimate, for example, the time that a care provider spends with a patient. However, as information technology systems become more prevalent in health care, this type of data will be increasingly available.

In developing the data inputs for a model, it is also very important to make sure that all of the data needed for the model is collected and/or estimated. On the demand side, this means including all demands for care, including the ones that may not have been met in the past because of inadequate capacity. For example, in a hospital ED, some patients who are forced to wait a long time before seeing a physician leave the ED before being seen. If these are not captured in the data collection system that is being used to measure demands, the model will underestimate the capacity needed to meet the desired performance standard. On the service side, it is important to include all of the time spent by the servers that is directly associated with caring for the patient. For a physician, this may include reviewing medical history and test results in addition to direct examination of the patient.

In addition to data, a queueing analysis of a particular health care system requires the identification of one or more delay measures that are most important to service excellence for that facility. These measures should reflect both patient perspectives as well as clinical realities. For example, though hospital ED arrivals with nonurgent problems may not require care within an hour or so from a clinical perspective, clearly very long waits to see a physician will result in high levels of dissatisfaction, and perhaps even departure, which could ultimately lead to lost revenue. Trying to decide on what might be a reasonable delay standard in a specific health care facility is not trivial due to a lack of knowledge of both patient expectations as well as the impact of delays on clinical outcomes for most health problems.

In summary, health care managers are increasingly aware of the need to use their resources as efficiently as possible in order to continue to assure that their institutions survive and prosper. This is particularly true in light of the growing threat of sudden and severe demand surges due to outbreaks of epidemics such as SARS and avian flu, or terrorist incidents. As this chapter has attempted to demonstrate, effective capacity management is critical to this objective as well as to improving patients' ability to receive the most appropriate care in a timely fashion. Yet effective capacity management must deal with complexities such as trade-offs between bed flexibility and quality of care, demands from competing sources and types of patients, time-varying demands, and the often differing perspectives of administrators, physicians, nurses and patients. All of these are chronic and pervasive challenges affecting the ability of hospital managers to control the cost and improve the quality of health care delivery. To meet these challenges, managers must be informed by operational and performance data and use these data in models to gain insights that cannot be obtained from experience and intuition alone. Queueing analysis is one of the most practical and effective tools for understanding and aiding decision-making in managing critical resources and should become as widely used in the health care community as it is in other major service sectors.

References

Allen, A. O. (1978). *Probability, statistics and queueing theory, with computer science applications*. New York, NY: Academic.

Brecher, C., & Speizio, S. (1995). *Privatization and public hospitals*. New York, NY: Twentieth Century Fund Press.

Brewton, J. P. (1989) Teller staffing models. *Financial Manager's Statement*, July–August: 22–24.

Brigandi, A. J., Dargon, D. R., Sheehan, M. J., & Spencer, T., III. (1994). AT&Ts call processing simulator (CAPS) operational design for inbound call centers. *Interfaces, 24*, 6–28.

Brockmeyer, E., Halstrom, H. L., & Jensen, A. (1948). *The life and works of A.K. Erlang (Transactions of the Danish Academy of Technical Science)* (Vol. 2). Copenhagen, Denmark: Danish Academy of Science.

Brusco, M. J., Jacobs, L. W., Bongiorno, R. J., Lyons, D. V., & Tang, B. (1995). Improving personnel scheduling at airline stations. *Operations Research, 43*, 741–751.

Chelst, K., & Barlach, Z. (1981). Multiple unit dispatches in emergency services. *Management Science, 27*, 1390–1409.

Freeman, R. K., & Poland, R. L. (1997). *Guidelines for perinatal care* (4th ed.). Washington, DC: American College of Obstetricians and Gynecologists.

Green, L. V. (2003). How many hospital beds? *Inquiry, 39*, 400–412.

Green, L. V., Giulio, J., Green, R., & Soares, J. (2005). Using queueing theory to increase the effectiveness of physician staffing in the emergency department. *Academic Emergency Medicine, 13*, 61–68.

Green, L. V., & Kolesar, P. J. (1984). The feasibility of one-officer patrol in New York City. *Management Science, 20*, 964–981.

Green, L. V., Kolesar, P. J., & Svoronos, A. (1991). Some effects of nonstationarity on multiserver Markovian queueing systems. *Operations Research, 39*, 502–511.

Green, L. V., Kolesar, P. J., & Svoronos, A. (2001). Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research, 49*, 549–564.

Green, L. V., & Nguyen, V. (2001). Strategies for cutting hospital beds: The impact on patient service. *Health Services Research, 36*, 421–442.

Hall, R. W. (1991). *Queueing methods for service and manufacturing*. Upper Saddle River, NJ: Prentice Hall.

Holloran, T. J., & Byrne, J. E. (1986). United Airlines station manpower planning system. *Interfaces, 16*, 39–50.

Institute of Medicine, Committee on Quality of Health Care in America. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Kim, S., Horowitz, I., Young, K. K., & Buckley, T. A. (1999). Analysis of capacity management of the intensive care unit in a hospital. *European Journal of Operational Research, 115*, 36–46.

Kolesar, P. J., Rider, K., Crabill, T., & Walker, W. (1975). A queueing linear programming approach to scheduling police cars. *Operations Research, 23*, 1045–1062.

Larson, R. C. (1972). *Urban police patrol analysis*. Cambridge, MA: MIT Press.

McCaig, L. F., & Burt, C. W. (2004). National hospital ambulatory medical care survey: 2002 Emergency department summary. *Advance Data from Vital and Health Statistics, 340*, 1–35.

Schneider, D. (1981). A methodology for analysis of comparability of services and financial impact of closure of obstetrics service. *Medical Care, 19*(4), 393–409.

Shmueli, A., Sprung, C. L., & Kaplan, E. H. (2003). Optimizing admissions to an intensive care unit. *Health Care Management Science, 6*(3), 131–136.

Stern, H. I., & Hersh, M. (1980). Scheduling aircraft cleaning crews. *Transportation Science, 14*, 277–291.

Taylor, P. E., & Huxley, S. J. (1989). A break from tradition for the San Francisco police: Patrol officer scheduling using an optimization-based decision support system. *Interfaces, 19*, 4–24.

Young, J. P. (1965). Stabilization of inpatient bed occupancy through control of admissions. *Journal of the American Hospital Association, 39*, 41–48.

Chapter 16
Rapid Distribution of Medical Supplies

Maged Dessouky, Fernando Ordóñez, Hongzhong Jia, and Zhihong Shen

Abstract Some important issues in the design of an efficient pharmaceutical supply chain involve deciding where to place the warehouses/inventories and how to route distribution vehicles. Solving appropriate facility location and vehicle routing problems can ensure the design of a logistic network capable of rapid distribution of medical supplies. In particular, both these problems must be solved in coordination to quickly disburse medical supplies in response to a large-scale emergency. In this chapter, we present models to solve facility location and vehicle routing problems in the context of a response to a large-scale emergency. We illustrate the approach on a hypothetical anthrax emergency in Los Angeles County.

Keywords Emergency supply • Facility location • Vehicle routing

1 Introduction

Rapid distribution of medical supplies plays a critical role in assuring the effectiveness and efficiency of the healthcare system. The medical supply distribution involves the movement of a large volume of different items that usually must be delivered rapidly. For example, in the USA, the distribution system must serve more than 130,000 pharmacy outlets every day on demand and a typical pharmacy relies on the distributors to have more than 10,000 SKUs accessible for delivery, often within 12 h (HDMA 2005).

In broad terms, most pharmaceuticals distributed in the USA go through a supply chain that comprises the following steps (Belson 2005):

M. Dessouky (✉) • F. Ordóñez • H. Jia • Z. Shen
Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089-0193, USA
e-mail: maged@usc.edu

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_16, © Springer Science+Business Media New York 2013

- Manufacturers produce various pharmaceuticals necessitated by demand.
- Distributors manage large warehouses and control the movement of supplies from manufacturers to the retailers.
- Retailers, including hospitals, clinics, independent pharmacies, chain pharmacies, and grocery stores, sell or dispense the pharmaceuticals to customers.

The pharmaceutical supply chain is relatively complex compared to the supply chains for other products, particularly when considering the strict deadline and sufficiency requirements. Different information technologies such as product identification, bar coding, usage related information, and electronic identification have been applied to facilitate the rapid distribution of the pharmaceuticals in the supply chain (Belson 2005). Furthermore, logistic and inventory control of the pharmaceuticals have also been widely investigated in the research community in the past decades; for example, see Rebidas et al. (1999), Rubin and Keller (1983), and McAllister (1985).

It is the design of the distribution system in particular, that most significantly affects the rapid disbursement of pharmaceuticals, directly impacting the quality of healthcare. The design of an effective distribution system comprises the careful consideration of two strategic planning issues:

- Where to place the facilities including warehouses and inventories in support of rapid distribution of the medical supplies
- What is the best strategy to distribute the medical supplies and what routes need to be used?

Operations research models play an important role in addressing these logistical problems for distribution systems. At the heart of both questions there is a transportation network to distribute the medical supplies. The question of where to place warehouses/inventories is essentially a facility location problem within this supply network, while the disbursement of supplies can be posed as a vehicle routing problem (VRP) on this network. The benefits of modeling and solving the facility location problem and vehicle routing problem are twofold. First, from a planning perspective, the models and solutions can aid planners to optimally determine the facility locations and vehicle routes and thus maximize the efficiency and effectiveness of the pharmaceutical supply chain system as a whole. Second, these plans can become well tested operating policies, which can further improve performance. Clearly, the plans need to be flexible enough to accommodate contingencies of daily operations. For the plans to be robust, they must take into consideration the stochastic nature of the problem such as uncertain demand, traffic conditions, etc.

Large-scale emergencies create situations that demand a rapid distribution of medical supplies and thus require an efficient and coordinated solution to both the facility location and vehicle routing problems. In particular, the response to a large-scale emergency must take into consideration that:

- A huge demand for medical supplies appears within a short time period and thus large quantities of medical supplies must be brought to the affected area.
- The local first-responders and resources will be overwhelmed.

- Although tremendous in their magnitude, large-scale emergencies occur with low frequency.

An additional parallel distribution system is envisioned in response to large-scale emergencies such as earthquakes, terrorist events, etc. as massive supplies that are brought to the affected area have to be rapidly disbursed among the affected population. Indeed, many countries maintain national stockpiles of medical supplies that can be delivered in push packages to the Emergency Staging Area (ESA) in case of a large-scale emergency. For example, to address emergencies of infectious disease outbreak, the federal government of the USA maintains a Strategic National Stockpile (SNS) which contains about 300 million doses of smallpox vaccines and enough antibiotic to treat 20 million people for anthrax (CDC website 2005). Furthermore, a vendor managed inventory system (VMI) has also been developed to augment the SNS from pharmaceutical vendors to ESAs within 21–36 h. During a large-scale emergency, the medical supplies at the national stockpile and VMI require direct delivery and disbursement to ESAs and dispensing centers from which the population could receive the medical supplies. Rapid delivery and disbursement of the large volume of supplies need careful planning and professional execution to save lives, particularly in high-density urban regions like Southern California.

In this chapter, we analyze the facility location and vehicle routing problems, which are crucial for a rapid distribution of medical supplies in response to large-scale emergencies. We use the anthrax disease as an emergency example to investigate the problems of determining where to locate the staging areas to receive the national stockpile and how to route the vehicles to distribute the medical supplies. The rest of the chapter is organized as follows: Section 2 presents a literature review of the facility location and vehicle routing problems that are related to emergency services. In Section 3, we describe an anthrax emergency example in a metropolitan area and then analyze the requirements for locating the facilities and routing the vehicles for rapid medical supply distribution. In Section 4, we propose a facility location model and a vehicle routing model that address the characteristics of an anthrax emergency. In Section 5, we demonstrate how the proposed models can be used to solve the facility location problem and the VRP. The solutions, including the selected staging areas and vehicle routes to store and distribute the medical supplies, are discussed. Finally, we conclude the chapter and give future research directions in Sect. 6.

2 Literature Review

Facility location problems and VRPs have been extensively investigated by different researchers and practitioners. In this section, we review the prior work that is related to different emergencies settings.

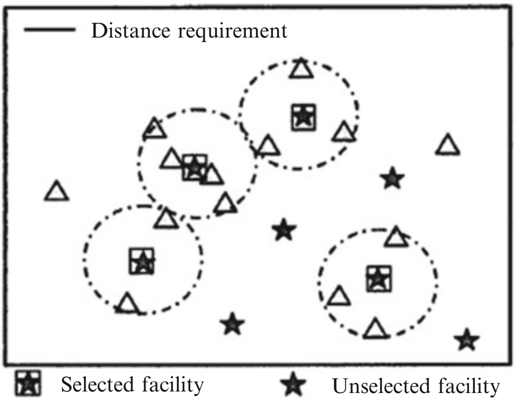


Fig. 16.1 Covering problem example

2.1 *Review of Facility Location Problems*

Various location models have been proposed to formulate different facility location problems for emergency services. Based on the objectives, these location models can be classified into covering models, *P*-median models, and *P*-center models.

2.1.1 Covering Models

Covering models are the most widespread location models for formulating the emergency facility location problem. The objective of covering models is to provide "coverage" to the demand points. A demand point is considered as covered only if a facility is available to service the demand point within a distance limit. Figure 16.1 presents an illustration of an infeasible covering problem, where the coverage area of a facility is indicated by circles around the four selected locations.

Toregas et al. (1971) first proposed the location set covering problem (LSCP), aiming to locate the least number of facilities to cover all demand points. Since all the demand points need to be covered in the LSCP, the resources required for facilities could be excessive. Recognizing this problem, Church and ReVelle (1974) and White and Case (1974) developed the MCLP model that does not require full coverage to all demand points. Instead, the model seeks the maximal coverage with a given number of facilities. The MCLP and different variants of it have been extensively used to solve various emergency service location problems (see e.g., Benedict 1983, and Hogan and ReVelle 1986).

Research on emergency service covering models has also been extended to incorporate the stochastic and probabilistic characteristics of emergency situations so as to capture the complexity and uncertainty of these problems. Examples of these stochastic models can be found in recent papers by Goldberg and Paz (1991), ReVelle et al. (1996), and Beraldi and Ruszczynski (2002). There are several approaches to model stochastic emergency service covering problems. The first approach is to use chance constrained models (Chapman and White 1974). Daskin (1983) used an estimated parameter (q) to represent the probability that at least one server is free to serve the requests from any demand point. He formulated the Maximum Expected Covering Location Problem (MEXCLP) to place *P* facilities on a network with the goal to maximize the expected value of population coverage. ReVelle and Hogan (1986) later enhanced the MEXCLP and proposed the Probabilistic Location Set Covering Problem (PLSCP). In the PLSCP, a server busy fraction and a service reliability factor are defined for the demand points. Then the locations of the facilities are determined such that the probability of service being available within a specified distance is maximized. The MEXCLP and PLSCP later were further modified to tackle other EMS location problems by ReVelle and Hogan (MALP) (1989a), Bianchi and Church (MOFLEET) (1988), Batta et al. (AMEXCLP) (1989), Goldberg et al. (1990), and Repede and Bernardo (TIMEXCLP) (1994). A summary and review to the chance constrained emergency service location models can be found in ReVelle (1989).

Another approach to modeling stochastic emergency medical service (EMS) covering problems is to use scenario planning to represent possible values for parameters that may vary over the planning horizon in different emergency situations. A compromise decision is made to optimize the expected/worst-case performance or expected/worse-case regret across all scenarios. For example, Schilling (1982) extended the MCLP by incorporating scenarios to maximize the covered demands over all possible scenarios. Individual scenarios are respectively used to identify a range of good location decisions. A compromise decision is made to the final location configuration that is common to all scenarios in the horizon.

2.1.2 P-Median Models

Another important way to measure the effectiveness of facility location is by evaluating the average (total) distance between the demand points and the facilities. When the average (total) distance decreases, the accessibility and effectiveness of the facilities increase. This relationship applies to both private and public facilities such as supermarkets, post offices, as well as emergency service centers, for which proximity is desirable. The *P*-median problem takes this measure into account and is defined as: minimize the average (total) distance between the demands and the selected facilities. We illustrate a *P*-median model in Fig. 16.2. The total cost of the solution presented is the sum of the distance between demand points and selected locations represented by the black lines.

Since its formulation, the *P*-median model has been enhanced and applied to a wide range of emergency facility location problems. Carbone (1974) formulated a deterministic *P*-median model with the objective of minimizing the distance traveled by a number of users to fixed public facilities such as medical or day-care centers. Recognizing the number of users at each demand node is uncertain, the author further extended the deterministic *P*-median model to a chance constrained

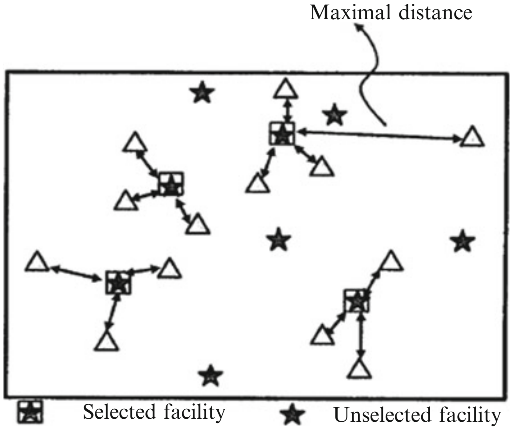


Fig. 16.2 P-median/P-center problem example

model. The model seeks to maximize a threshold and meanwhile ensure the probability that the total travel distance is below the threshold is smaller than a specified level a. Paluzzi (2004) discussed and tested a P-median based on a heuristic location model for placing emergency service facilities for the city of Carbondale, IL. The goal of this model is to determine the optimal location for placing a new fire station by minimizing the total aggregate distance from the demand sites to the fire station.

One major application of the P-median models is to dispatch EMS units such as ambulances during emergencies. For example, Carson and Batta (1990) proposed a P-median model to find the dynamic ambulance positioning strategy for campus emergency service. Mandell (1998) developed a P-median model and used priority dispatching to optimally locate emergency units for a tiered EMS system that consists of advanced life-support (ALS) units and basic life-support (BLS) units.

Uncertainties have also been considered in many P-median models. Mirchandani (1980) examined a P-median problem to locate fire-fighting emergency units with consideration of stochastic travel characteristics and demand patterns. Serra and Marianov (1998) implemented a P-median model and introduced the concept of regret and minmax objectives. The authors explicitly addressed in their model the issue of locating facilities when there are uncertainties in demand, travel time or distance.

2.1.3 P-Center Models

In contrast to the P-median models, which concentrate on optimizing the overall (or average) performance of the system, the P-center model attempts to minimize the worst performance of the system and thus is also known as minimax model. The P-center model considers a demand point is served by its nearest facility and therefore full coverage to all demand points is always achieved. In the last several decades, the P-center model and its extensions have been investigated and applied in the context of locating facilities such as EMS centers, hospitals, fire station, and other public facilities, The objective function for the *P*-center model of the location solution in Fig. 16.2 quantifies only the longest distance between a demand point and a selected location.

In order to locate a given number of emergency facilities along a road network, Garfinkel et al. (1977) examined the fundamental properties of the *P*-center problem. He modeled the *P*-center problem using integer programming and the problem was successfully solved by using a binary search technique and a combination of exact tests and heuristics. ReVelle and Hogan (1989b) formulated a *P*-center problem to locate facilities so as to minimize the maximum distance within which EMS service is available with α reliability. System congestion is considered and a derived server busy probability is used to constrain the service reliability level that must be satisfied for all demands. Stochastic *P*-center models have also been formulated for EMS location problems. For example, Hochbaum and Pathria (1998) considered the emergency facility location problem that must minimize the maximum distance on the network across all time periods. The cost and distance between the locations vary in each discrete time period. The authors used k underlying networks to represent different periods and provided a polynomial time 3-approximation algorithm to obtain the solution for each problem. Talmar (2002) utilized a *P*-center model to locate and dispatch three emergency rescue helicopters to serve the growing EMS demands from accidents of tourist activities such as skiing, hiking and climbing at the north and south end of the Alpine mountain ranges. One of the model's aims is to minimize the maximum (worst) response times and the author used effective heuristics to solve the problem.

2.2 *Review of VRPs*

Routing vehicles in response to a large-scale emergency typically include various uncertainties such as stochastic demands and/or travel times. In this section, we first review the literature on the stochastic vehicle routing problem (SVRP). We then review other vehicle routing/dispatching problems in the literature that have been formulated for emergency situations.

2.2.1 Stochastic Vehicle Routing Problems (SVRPs)

SVRPs differ from the deterministic VRPs in several aspects, such as problem formulations and solution techniques. SVRPs are usually divided, according to the uncertainties in consideration, into SVRPs with stochastic customers and/or demands, and SVRPs with stochastic travel time and/or service time.

The VRP with stochastic customers (VRPSC) addresses the probabilistic presence of customers (see e.g., Jezequel (1985), Jaillet (1987), and Jaillet and Odoni

(1988). Bertsimas (1988) gave a systematic analysis to this problem. The properties, the bounds, and the heuristics to solve the problem were also presented.

The VRP with stochastic demand (VRPSD) captures the uncertainty of customer demands (i.e., the demands at the individual delivery (pickup) locations behave as random variables). An early investigation on the VRPSD comes from Stewart and Golden (1983), who applied the chance constraint modeling and resource methods to model the problem. Dror and Trudeau (1986) later illustrated the effects of route failure on the expected cost of a route, as well as the impact that a redirection of the predesigned route has on the expected cost. In the late 1980s and early 1990s, along with the conventional stochastic programming framework, Markovian Decision Processes for single-stage and multi-stage stochastic models were introduced by Dror (1989, 1993) to investigate the VRPSD. Another major contribution to the study of VRPSD comes from Bertsimas (1988, 1992). Their work illustrates different recourse policies that could be applied to re-optimize the routes. More recently, a re-optimization based routing policy for the VRPSD has been demonstrated by Secomandi (2001). In their work, a rollout algorithm is proposed to improve a given a priori solution.

The vehicle routing problem with stochastic customers and demands (VRPSCD) combines the VRPSC and the VRPSD. Early work on this topic includes Jezequel (1985), Jaillet (1987), and Jaillet and Odoni (1988). Motivated by applications in strategic planning and distribution systems, Bertsimas (1992) constructed an a priori customer visit sequence with minimal expected total distance and analyzed the problem using a variety of theoretical approaches. Gendreau et al. (1995) proposed a L-shaped method for the VRPSCD and solved it to optimality for instances of up to 46 customers. Another strategy to account for the demand uncertainties is to develop a waiting strategy for vehicles to strategically wait at predetermined locations in order to maximize the probability of meeting any future anticipated demand (Branke et al. 2005).

VRP with stochastic travel time (VRPSTT) addresses the unknown knowledge of the road conditions. Systematic research on the VRP with service time and travel time (VRPSSTT) has been done by Laporte et al. (1992). They proposed three models for the VRPSTT: chance constrained model, 3-index recourse model, and 2-index recourse model. The VRPSSTT model was also applied by Lambert et al. (1992), and Hadjiconstrantinou and Roberts (2002) to optimize the customer service in the banking and other commercial systems. Jula et al. (2005) has developed an approximate solution approach for random travel times with hard time windows. Their approximation approach is based on developing estimations for the first two moments of the arrival time distribution.

2.2.2 Vehicle Locating/Routing/Dispatching for Emergency Services

Emergency service systems (e.g., police, fire, etc.) need to dispatch their response units to service requests. In an emergency, the primary objective is to save lives, and thus sending response units to the incident site at the earliest time has the highest priority. However the requests for emergency services are usually unpredictable and furthermore they come with a relatively low frequency. Therefore the planner is generally faced with two major problems. First, an allocation problem in which the response units that are sent for service need to be determined; and second, a re-deployment problem in which the available response units need to be deployed at the potential sites in preparation to incoming requests needs to be determined.

One important thrust and cornerstone in vehicle locating/routing/dispatching for emergency services is the development and application of the queuing approach. The most well known queuing models for emergency service problems are the hypercube and the approximated hypercube by Larson (1974, 1975), which consider the congestions of the system by calculating the steady-state busy fractions of servers on a network. The hypercube model can be used to evaluate a wide variety of output performance such as vehicle utilization, average travel time, inter-district service performance, etc. Particularly important in the hypercube models is the incorporation of state-dependent interactions among vehicles that preclude applications of traditional vehicle locating/routing/dispatching models. Larson (1979) and Brandeau and Larson (1986) further extended and applied the hypercube models with locate-allocate heuristics for optimizing many realistic EMS systems. For example, these extended models have been successfully used to optimize the ambulance deployment problems in Boston and the EMS systems in New York. Based on the hypercube queuing model, Jarvis (1977) developed a descriptive model for operation characteristics of an EMS system with a given configuration of resources and a vehicle locating/dispatching model for determining the placement of ambulances to minimize average response time or other geographically based variables. Marianov and ReVelle (1996) created a realistic vehicle locating/dispatching model for emergency systems based on results from queuing theory. In their model, the travel times or distances along arcs of the network are considered as random variables. The goal is to place a limited numbers of emergency vehicles, such as ambulances, in a way as to maximize the calls for service. Queueing models formulating other theoretical and practical problems have also been reported by Berman and Larson (1985), Batta (1989), and Burwell et al. (1993).

3 A Large-Scale Emergency: An Anthrax Attack

In this section, we use an anthrax attack emergency to demonstrate the characteristics of a large-scale emergency. We then derive the requirements for the facility location problem and vehicle routing problem for the medical supply distribution in a large-scale emergency. Note that different emergency scenarios may require different response plans. The area in which we consider the anthrax attack emergency is Los Angeles (LA) County, which consists of 2054 census tracts and a total population of 9.5 million. In addition, we identify a number of potential eligible

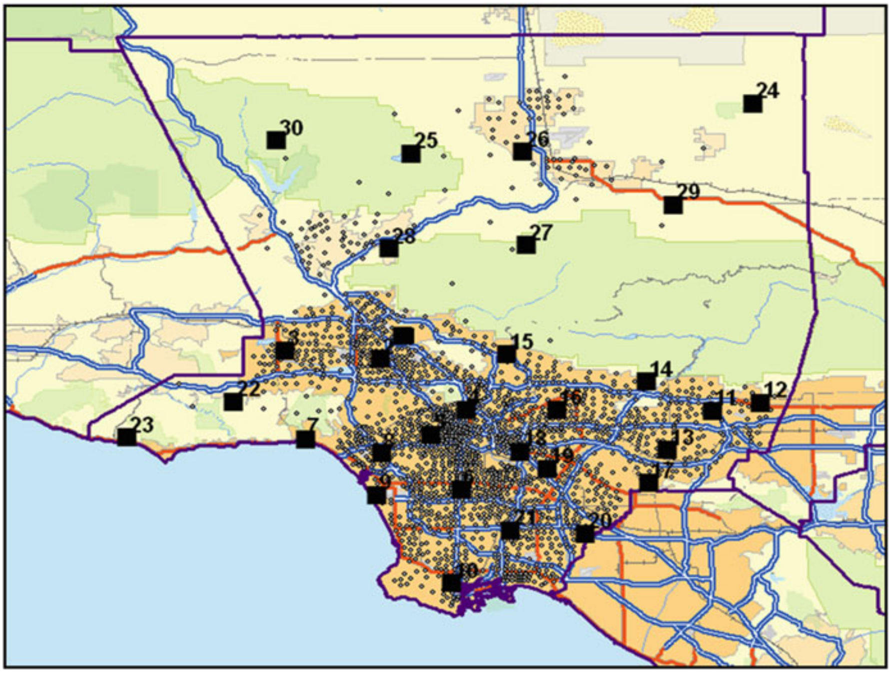


Fig. 16.3 Los Angeles county

medical supply facility sites (see Fig. 16.3) and the goal is to select some of these eligible facility sites as the staging areas to dispense the vaccinations.

3.1 *Characteristics of an Anthrax Emergency*

Anthrax is an acute infectious disease caused by a spore-forming bacterium. The anthrax spores can be used as a bioterrorist weapon, as was the case in 2001, when *Bacillus anthracis* spores had been intentionally distributed through the postal system causing 22 cases of anthrax emergency, including 5 deaths (CDC website 2005). If the anthrax spores had been disseminated in an airborne manner through airplanes or from high buildings, thousands of people and hundreds of blocks would have been severely affected. Anthrax causes disease after inoculation of open or minor wounds, ingestion, or inhalation of the spores. At the earliest sign of disease, patients should be treated with antibiotics and other necessary medications to maximize patient survival. Otherwise, shock and death could ensue within 24–48 h. Although no cases of person-to-person transmission of inhalation anthrax have ever been reported, cutaneous transmissions have occurred. Early treatment of anthrax disease is usually curative and significant for recovery. For example, patients with cutaneous anthrax have reported case fatality rates of 20 % without antibiotic treatment and less than 1 % with it (CDC website 2005).

The impact of an anthrax attack to the population can be tremendous. First, thousands of people could be directly infected by the disease at the incident site. Second, the affected area could quickly spread from the original incident site to a much larger region by the movement of the infected but unaware people because the anthrax attack is usually covert and the appearance of the disease symptom may lag the attack from hours to days. Third, after an anthrax disease emergency becomes known in public, people may panic and become scared. They may request medical treatment or vaccination even if they are not actually infected or not in a high-risk situation.

Huge demands for medical supplies could occur in a short time period after the anthrax attack. Blanket medical service coverage and mass vaccination may be necessary to all the population in a region. As such, a large amount of vaccines may be required. However an anthrax emergency has a low occurrence frequency and it is very expensive for any local region to maintain massive medical supplies for such a rare event. Therefore, large volumes of medical supplies for such an emergency are usually not stored at local sites. Instead, they are inventoried by the government at national stockpiles which consist of large quantities of medications, vaccines, and antibiotics to protect the public. The national stockpile is organized as push packages for flexible response and immediate deployment to a designated site within 12 h (e.g., the SNS of the USA). Once delivered to the local areas, the stockpiles can be repackaged and distributed to various demand points though the local dispensing centers (staging areas). The overall process of a rapid medical supply distribution for a large-scale emergency can be depicted as follows in Fig. 16.4. The details for each procedure in this process are described in the following sections.

It should be noted that anthrax is not contagious from person to person and the medical service coverage should depend on the actual disease spreading pattern. For example, if the attack can be detected at an early stage and the infected people can be identified and quarantined in a timely manner, then only the areas near the incident site need to be serviced with medical supplies. In this example we consider the worst case scenario and assume that the delayed detection of the attack has caused intractable population movements, and thus a blanket medical service coverage to all the areas is required. The logistical problem for such a worst case scenario is much more challenging than other scenarios in which only a portion of a region needs to be provided with medical supplies. Also note that the blanket medical service coverage is similarly applicable to contagious emergencies such as smallpox. During a contagious disease outbreak, it is possible that some areas are more critical than others due to certain disease spreading pattern. However, a mass vaccination to all the areas may be desired since it could effectively stem the disease transmission among the population (CDC 2005).

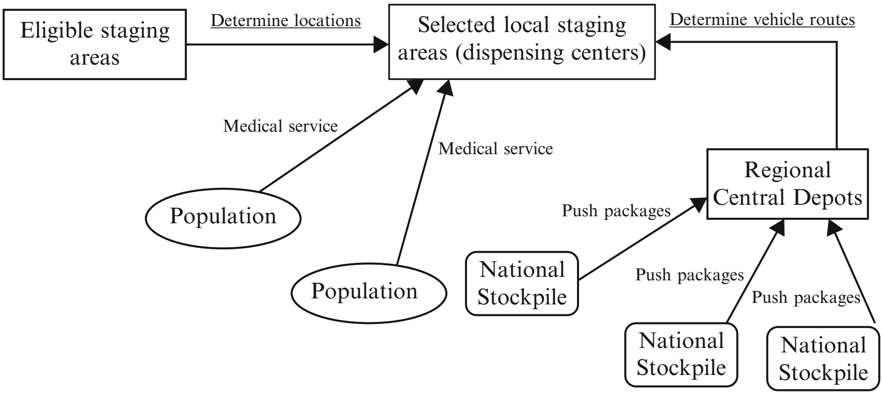


Fig. 16.4 Medical supply distribution process

3.2 Requirements to Facility Location Deployment

As mentioned in the last section, the medical supplies are usually not stored at the local level, and during an anthrax emergency the national stockpile will be called to service the demands at the local areas. Therefore, the primary goal of the facility location problem is to determine a number of local staging areas so that the supplies from the national stockpile can be received, repackaged, and distributed.

The deployment of medical facility sites (staging areas) in response to a large-scale emergency must account for massive service requirements. In most traditional facility location problems, each individual demand point is covered only by one facility given the fact that demand does not appear in large amounts. However, in the event of an anthrax emergency, if a mass vaccination to the population is necessary, the demands for medical services will be significant. As a result, a redundant and dispersed placement of the facilities (staging areas) is required so that more medical supplies could be mobilized to service different demand points to reduce mortality and morbidity.

Another important aspect of the facility locations for the anthrax emergency is the fact that given the occurrence of the emergency at an area, the resources of a number of facilities will be applied to quell the impact of the emergency, not only those located closest to the emergency site. This implies that there are different types of coverage, or quality of coverage, which can be classified in terms of the distance (time) between facilities and demand points. Thus, a facility that is close to a demand point provides a better quality of coverage to that demand point than a facility located far from that demand point. When planning the emergency medical services, it is important to consider adequate staging areas of various qualities for each demand point.

Furthermore, potential demand areas for medical services need to be categorized in a different way than other regular emergencies. Each demand area has distinct attributes, such as population density, economic importance, geographical feature, weather pattern, etc. Therefore, different requirements of facility quantity and quality should be assigned for each demand point so that all demand points can be serviced in a balanced and optimal manner. For example, for the demand points at a downwind and populous downtown area, a larger quantity of facilities should be located at a relatively better quality level, as opposed to the demand points at an upwind and/or less populous area.

Moreover, the facility location objective for an anthrax disease emergency should be carefully defined. An anthrax emergency is bound to impact lives regardless of the solution. Thus care should be taken in prioritizing one solution over another. Since the blanket medical service coverage and mass vaccination may be carried out, all the demand points in the affected areas need to be serviced simultaneously. To optimize the overall performance of the medical distribution system, it is desirable that the total (average) distance from all demand points to the staging areas be minimized. Thus, a P-median model with multiple facility quantity-of- coverage and quality-of-coverage requirements is applicable. It is important to note the model that is selected should be in accordance with the characteristics of the emergency, and different models may be suitable for different emergency scenarios. For example, for the emergency of a dirty bomb attack in which only a portion of a region needs be serviced by the medical supplies, the covering model may be more applicable since the model ensures a maximal population coverage by the medical supply facilities.

Finally, the selection of eligible staging area sites for the anthrax emergency must consider a different set of criteria that are used for regular emergencies. For instance, the facilities should have easy access to more than one major road/ highway including egress and ingress. The sites should be secure and invulnerable to damages caused by the emergencies. In this paper we consider eligible staging area sites as given.

3.3 Requirements to Vehicle Routing

In an anthrax emergency, the primary goal of vehicle routing is to deliver the medical supplies to the affected areas as soon as possible. To reach this goal, a fast and efficient vehicle routing/dispatching plan needs to be executed. To maximize life-saving in an anthrax emergency, medications, antibiotics, and vaccines should be administered to the affected population within a specified time limit (within 24–36 h). This implies that vehicles need to have a hard time-window for medical supply delivery. To minimize the loss of life at any demand area, the medical supplies must be sent to the demand area within this hard time-window. Note that although a hard time-window is used to model the VRP for the anthrax emergency, it may not always be applicable to other emergencies. For example, for a contagious disease outbreak, such as smallpox, the demand for medical supplies could be a continuous function of time. In such a case, a soft time- window approach may be more suitable to model the VRP.

The input parameters to the vehicle routing problem in the anthrax emergency have a probabilistic/stochastic nature. For example, the traffic conditions may change and therefore the vehicle travel times can be highly uncertain. In addition, the demands for medical services may be stochastic because of the way the disease disseminates, the wind direction, the geographic conditions, etc. As such, the vehicle routing problem needs to capture the demand uncertainties and provide a robust solution that performs well in a variable environment.

Moreover, because of the massive service requirements, the demand at a location is not necessarily satisfied by a single truckload. As such, the vehicle routing problem for the anthrax emergency should allow for split delivery (i.e., a point can be visited more than once if the demand exceeds the load capacity of available vehicles). Also, the VRP for the anthrax emergency should be a multi-depot problem since many local depots are dispersed across a region. However, unlike the traditional multi-depot VRP, which requires each vehicle to return to its origination depot, the vehicles are now allowed to return to any depot for reloading and then continue serving other demand points. This requirement enables the vehicles to distribute the medical supplies in a more flexible manner.

Finally, the primary objective of the vehicle routing problem during the emergency should be the minimization of loss of lives, which is caused by minimizing the unmet demands for the medical service.

As mentioned before, the facility location and vehicle routing models can be used as a planning tool to determine the optimal staging areas and vehicle routes considering the probabilistic/stochastic nature of the emergency. These plans can serve as practical drills for the first responders to prepare and train them for a possible emergency, and they may be to be altered in the event that an emergency has occurred once the characteristics of the emergency become known.

4 Mathematical Model Formulations

Based on the analysis stated in the last section, we now formulate a facility location model and a vehicle routing model that take into account the characteristics of the anthrax disease emergency. Generalizations of the models discussed in this section can found in Jia et al. (2005) for the facility location problem and Shen et al. (2005) for the vehicle routing problem.

4.1 Formulation of the Facility Location Model

To formulate the facility location model, we use I to denote the set of demand points and J to denote the set of eligible facility sites (staging areas). Indexed on these sets we define two types of integer variables:

Decision variables:

$x_j = \begin{cases} 1 \\ 0 \end{cases}$ if a facility is placed at site j; otherwise

$z_{ij} = \begin{cases} 1 \\ 0 \end{cases}$ if a facility j services demand point i; otherwise

Furthermore, we define the following parameters:

Input Parameters:

$\text{Pop}_i =$ the population of demand point i

$d_{ij} =$ the distance between demand point i and facility location j

$D_t =$ the distance limit within which a facility could service demand point i

$N_i = \{j | d_{ij} \le D_i\}$, the set of eligible facility sites that are located within the distance limit and thus are able to service demand point i

$Q_i =$ the required number of facilities that must be assigned to demand point i so that i is considered as covered

$P =$ the maximal number of facilities that can be placed in J

We can now formulate the model to locate P facilities to service the population during an anthrax emergency, requiring that Q_i facilities service demand point i with the same quality.

$$\text{Minimize} \sum_{i \in I} \sum_{j \in J} \text{Pop}_i d_{ij} z_{ij} \tag{16.1}$$

Subject to:

$$\sum_{j \in J} x_j \le P \tag{16.2}$$

$$\sum_{j \in N_i} z_{ij} \ge Q_i \ \forall i \in I, \tag{16.3}$$

$$z_{ij} \le x_j \ \forall i \in I, j \in J, \tag{16.4}$$

$$x_j, z_{ij} = \{0, 1\} \ \forall i \in I, j \in J, \tag{16.5}$$

The objective (16.1), as mentioned in Sect. 3.2, is to minimize the total demand-weighted distance between the demand points and the facilities. Constraint (16.2) states that there are P facilities to be located in a set J of possible locations. Constraint (16.3) ensures that demand point i is assigned with a required quantity (Q_i) of facilities servicing it. This constraint also requires that all the facilities assigned to demand point i need to be located within the given distance limit. Constraint (16.4) allows assignment only to the sites at which facilities have been located. Finally constraint (16.5) enforces the integrality of variables Z_{ij} and x_i

Consider now the problem with multiple quality-of-coverage requirements at each demand point. Let us assume that at demand point i we must have Q_i^1, Q_i^2, …, Q_i^q, facilities for each quality from 1 to q, where quality Q_i^1 represents the facilities that are closest to demand point i, Q_i^2 are the facilities located farther than those of quality 1, and so on. Thus the facility location model needs to be modified as follows:

1. *Objective function*: Since multiple quality-of-coverage is considered, the objective function needs to be optimized across different quality levels. Because the facilities with a higher quality level (i.e., closer to the demand points) are usually considered to be more crucial in servicing the demand points, as opposed to the facilities with lower quality levels (i.e., farther from the demand points), we introduce a weight parameter, h^r, to prioritize the importance of the facilities at each different quality level r. Also we modify Z_{ij} to z_{ij}^r in order to differentiate the facilities that are servicing the demand points at different quality levels. Thus, we obtain the modified objective function:

$$\text{Minimize} \sum_r \sum_{i \in I} \sum_{j \in J} h^r \text{Pop}_i d_{ij} z_{ij}^r \tag{16.6}$$

2. *Constraints*: First, the group of constraints (16.3) needs to be changed to:

$$\sum_{j \in N_i^r} z_{ij}^r \geq Q_i^r \ \ \forall i \in I, r = 1, \ldots q \tag{16.7}$$

   The modified constraints state that, for each demand point, there must be more than a required quantity of facilities at each quality level so that this demand point can be considered as properly serviced. In addition, to avoid repeated assignment of a facility to any demand point for different quality requirements, we introduce another group of constraints:

$$\sum_r z_{ij}^r \leq 1 \ \ \forall i \in I, j \in J \tag{16.8}$$

As such, the modified objective (16.6), together with the constraints (112), (16.4), (16.5), (16.7), and (16.8), can be used to formulate the facility location problem for the anthrax emergency with multiple facility quantity- of-coverage and quality-of-coverage requirements. Note that in the problem formulation, all the Z_{ij}, Q_i *and* N_i need to be correspondingly changed to z_{ij}^r, Q_i^r and N_i^r.

Exact algorithms have been developed in the literature to solve different facility location problems; for example, see Holmberg (1999). However exact algorithms can only solve small problem instances in a reasonable computational time. Therefore, to solve the location problems for large-scale emergencies, efficient heuristics, such as greedy algorithms, genetic algorithms, or Tabu search, should be used. References to the heuristics for traditional location problems can be found in Jain et al. (2002) and Jaramillo et al. (2002).

4.2 Formulation of the Vehicle Routing Model

To formulate the vehicle routing model, we use K, I, J, and A to denote the sets of vehicles, demand points, facility sites, and medical supply items. In addition, we use node 0 as a dummy node to represent a virtual/imaginary central depot that is linked to each real depot (facility site). The cost or travel times on these links is set to be a large number. The dummy node is useful in representing the availability of vehicles. To conveniently denote different node combinations in the medical supply network, we further define set $C = I \cup J \cup \{\text{node}\ \_\ 0\}$, and set $\text{RO} = I \cup J$. Furthermore, indexed on these sets, we define the following deterministic parameters. Note that different from the facility location problem, in which the index i is defined as demand point i and the index j is defined as facility site j, here the indices i and j are defined as any node from set C, which could be either a demand point or a facility site (depot).

Deterministic Parameters:

n_i = the initial number of vehicles at facility site (depot) i

W_a = the unit weight of medical supply item a

$C_{a,k}$ = the load capacity of vehicle k for medical supply item a

$e_{a,i}$ = the earliest service start time for medical supply item a at demand point i

$l_{a,j}$ = the latest service start time for medical supply item a at demand point i

$S_{a,i}$ = the amount of medical supply item a supplied at facility site (depot) i

r_i = the service (loading/unloading) time at node i, including both the demand points and the facility sites

We use M as a large constant to transform nonlinear terms to linear ones for the time window constraints. In addition, the parameter α_D is used to represent the upper bound of unsatisfactory rate for demands at each demand point and α_T is used to denote the upper bound of total traveling time for each vehicle. These two parameters represent the probabilistic violation on the demand and travel time constraints.

As mentioned in the previous section, uncertainties exist in the anthrax emergency. We consider the following two parameters as stochastic variables.

Stochastic Parameters:

$\tau_{i,j,k}$ = the time required for vehicle k to travel from point i to j

$\zeta_{a,i}$ = the demand for medical supply item a at demand point i

Finally, four groups of decision variables are defined as follows:

Decision variables:

$X_{i,j,k} = \begin{cases} 1 \\ 0 \end{cases}$ if vehicle k traverses arc (ij):otherwise

$Y_{a,i,j,k}$ = the amount of medical supply item a traversing arc (i,j) using vehicle k

$U_{a,t}$ = the amount of unsatisfied demand for medical supply item a at demand point i

$T_{i,k}$ = the service start time for vehicle k at demand point i

Based on these parameters and variables, we are now in a position to formulate the stochastic vehicle routing problem, with the objective to minimize the unmet demands over all the demand points.

$$\text{Minimize} \sum_{a\in A}\sum_{i\in I} U_{a,i} \tag{16.9}$$

Subject to:

$$\sum_{k\in K} X_{0,i,k} \le n_i \;\; \forall i\in J \tag{16.10}$$

$$\sum_{j\in 1} X_{i,j,k} = 1 \;\; \forall i\in J, k\in K, \tag{16.11}$$

$$\sum_{j\in I} X_{j,i,k} = 1 \;\; \forall i\in J, k\in K, \tag{16.12}$$

$$\sum_{j\in \mathrm{RO}} X_{i,j,k} = \sum_{j\in \mathrm{RO}} X_{j,i,k} \qquad \forall i\in I, k\in K, \tag{16.13}$$

$$P\left\{\tau\middle|\left(T_{i,k}+r_i+\tau_{i,j,k}-T_{j,k}\right) \le \left(1-X_{i,j,k}\right)M\right\} \ge 1-\alpha_T \quad \forall i,j\in C \tag{16.14}$$

$$s_{a,i} - \sum_{k\in K}\left[\sum_{j\in C} Y_{a,i,j,k} - \sum_{j\in C} Y_{a,j,i,k}\right] \ge 0 \;\; \forall a\in A, \forall i\in J \tag{16.15}$$

$$Y_{a,0,i,k} + Y_{a,i,0,k} = 0 \;\; \forall a\in A, \forall i\in \mathrm{RO}, k\in K \tag{16.16}$$

$$X_{i,j,k}c_k \ge \sum_{a} w_a Y_{a,i,j,k} \;\; \forall \{i,j\} \subseteq \mathrm{RO}, k\in K, a\in A \tag{16.17}$$

$$e_{a,i} \sum_{j\in \Delta^+(i)} X_{i,j,k} \le T_{i,k} \le l_{a,i} \sum_{j\in \Delta^+(i)} X_{i,j,k} \;\; \forall a\in A, \forall i,j\in C \tag{16.18}$$

$$P\left\{\zeta\middle|\sum_{k\in K}\left[\sum_{j\in C} Y_{a,j,i,k} - \sum_{j\in C} Y_{a,i,j,k}\right] - U_{a,i} - \zeta_{a,i} \ge 0\right\} \ge 1-\alpha_D \\ \forall a\in A, \forall i\in I \tag{16.19}$$

$$X_{i,j,k} = \{0,1\}; Y_{a,i,j,k} \ge 0; U_{a,i} \ge 0; T_{i,k} \ge 0; \tag{16.20}$$

Constraints (16.10)–(16.14) characterize the vehicle flow on the medical distribution network. Constraint (16.10) states that the number of vehicles in service should not exceed the number of vehicles available at each depot at the beginning of the planning horizon. The number of vehicles in service is the total number of vehicles flowing from the dummy central depot 0 to each facility site. Constraints (16.11) and (16.12) specify that each vehicle can flow from and to only one facility site (depot). Constraint (16.13) states that all vehicles that flow into any demand point must also flow out of it. Constraint (16.14) is a chance constraint for the service start times at the demand points. The inner part, $(T_{i,k} + r_i + \tau_{i,j,k} - T_{j,k}) \leq (1 - X_{i,j,k})M$ guarantees the schedule feasibility with respect to time considerations. Constraint (16.15) gives the balanced material flow requirement for the facility sites. Constraint (16.16) prohibits the medical supply items flow from and to the dummy node. Constraint (16.17) allows the medical supply item to flow as long as there are sufficient vehicle capacities. It establishes the connection between the medical supply flow and vehicle flow. Constraint (16.18) gives the hard time window constraint on each demand point. Chance constraint (16.19) enforces the balanced material flow requirement for the demand points from a probabilistic perspective. It states that a small probability of unmet demands at each demand point is allowed within a threshold level. Finally constraint (16.20) enforces the integrality and non-negativity constraints on the variables.

5 Problem Solution and Analyses

In the preceding section, the facility location problem and the VRP for the anthrax disease emergency have been formulated. In this section, we first specify illustrative values for the input parameters and then we show how these proposed models could be applied to solve the facility location problem and the VRP for the anthrax emergency.

5.1 Facility Location Problem

5.1.1 Parameter Specification

There are 2054 census tracts and 9.5 million people in Los Angeles County. To define the demand distribution for medical services during an anthrax disease emergency, we use the day-time population density pattern that is available for Los Angeles County (ESRI website 2005). Furthermore, we use the centroid of each census tract as a demand point to represent the aggregated population in this tract. Thus we obtain 2054 discrete demand points that have different population densities. We assume that, in the anthrax emergency, the people at different demand points need to visit the selected facilities (staging areas) for vaccination. Note that although we assume that all the population at the demand points need to be serviced by the medical supplies, during an emergency, a more accurate demand pattern for medical supplies can be obtained by using schools, shopping malls and offices as indicators to assess the actual disease exposure.

To determine the staging areas that can be used to receive, repackage, and distribute the medical supplies from the national stockpile to the demand points, we identify 30 eligible facility sites. We assume that the resource limitation allows only 10 eligible facility sites to be selected to services the demand points ($P = 10$).

To ensure effective and efficient medical supply distribution, each demand point needs to be serviced by a required quantity of facilities that are located at each quality level. In practice, different quality levels should be defined for different demand points, based on the attributes of each point such as population density, political/economic importance, etc. In this example, for simplicity, we define a uniform quality requirement for all demand points; that is, each demand point needs two quality levels and the distance requirements for the first and second quality levels are 35 miles and 60 miles, respectively.

Furthermore, we specify the facility quantity requirement at each quality level for each demand point as follows:

1. $Q_i = 1$ if the population of demand point i is less than 4,000.
2. $Q_i = 2$, if the population of demand point i is between 4,000 and 8,000.
3. $Q_i = 3$, if the population of demand point i is greater than 8,000.

Finally, we specify the distances (times) between each pair of demand point and facility site. In practice, the roadway system should be used to define the distances since the medical supplies will be transported by vehicles during the emergency. However, for simplicity, in this illustrative example, we use the straight line distances between the demand points and facility sites.

5.1.2 Solution and Analyses

Based on the input parameters defined above, we solve the facility location problem for the anthrax emergency. The solution is depicted in Fig. 16.5. The problem was solved to optimality using a commercial integer program solver, CPLEX 8.1. The stars in the diagram represent the selected facilities.

In this solution, each demand point is covered by a required quantity of facilities at each of the two quality levels. Therefore, the demand points can be sufficiently serviced by the facilities in an efficient manner. The average distance from the demand points to their servicing facilities at quality level 1 is 25.8 miles; and the average distance at quality level 2 is 50.2 miles. Since the weighted total distance between the demand points and the facilities has been minimized (as defined by the objective function), the effectiveness of facility service performance is optimized.

It should be noted that a tight definition of the input parameters may lead to the facility location problem being infeasible; that is, no subset of P facilities is able to service all demand points within the defined quality levels (distance requirements). In this case, any one of the following four adjustments in the parameters can be made to make the problem feasible:

1. Increase the parameter P, i.e., the number of facilities that can be selected.
2. Relax the distance requirements, within which the facilities need to be located to service the demand points.
3. Drop the insignificant demand points (e.g., the ones with a low population density) from the problem constraints so that the limited resources (facilities) can be leveraged to the other demand points.

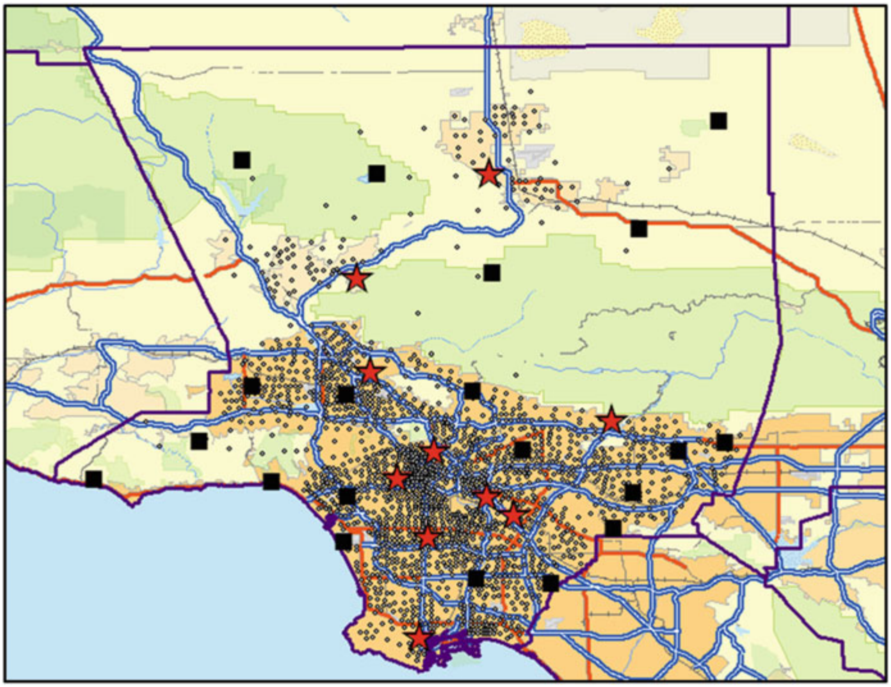


Fig. 16.5 Solution to the facility location problem

5.2 Vehicle Routing Problem (VRP)

5.2.1 Parameter Specification

After the facility sites (staging areas) have been determined, the solution is used as input parameters for the vehicle routing problem. In the anthrax emergency example, the 10 selected facility sites from the location problem are the demand points for the vehicle routing problem. To illustrate the VRP, for simplicity, we will use a single depot (i.e., Los Angeles International Airport as the central distribution warehouse) and a uniform capacity for each of a total of three vehicles to route and service the 10 staging facilities.

We calculate the demand on each selected facility by summing up the population in the tracts that are covered by the facility. The population size will be used as the criterion to specify the demand size; for example, 1 box of 100,000-dose anthrax vaccine is needed for every 100,000 people. As we stated in the previous section, the demand of each facility is stochastic. The exponential distribution, $p(x) = e^{(A-x)/B}\ IB$ (where the mean is $A + B$ and variance is B^2), is assumed with the mean value set according to the population density. The standard deviation is set to be 20 % of its mean value at each facility.

Furthermore, we assume an exponential distribution for the travel times between each pair of facility and central depot. Their mean values are specified as proportional to the Euclidean distances between them. We also set the standard deviation to be 20 % of the mean value of the travel time on each leg of the connection. Such an exponential distribution gives a lower bound and an upper bound for the travel time, which reasonably reflects the fact that travel time is constrained by the physical distance and the maximal speed of the vehicle, and could be prolonged by different traffic conditions.

Shock and death caused by untreated anthrax exposure could ensue within 24–48 h, and the dispatching from the central warehouse to the 10 selected local staging facilities is just one chain of the whole process of dispensing medical supplies. Hence we use a hard time window constraint of up to half of the required time for treatment (i.e., 12 h) to finish this placement.

Finally, we assume the total supply at the depot can meet 120 % of the summation of the mean value of the demand quantity at all points. However, since the demand is stochastic, it is possible that the demand cannot be fully satisfied in some cases.

5.2.2 Solution and Analyses

The routing problem is solved based on the parameters specified in the previous section and its result is compared with that of a deterministic formulation to show the advantage of our chance-constraint model.

The CPLEX solver was used to optimally solve both the deterministic and chance-constraint models to optimality with the given parameters. The deterministic model uses the mean value of the demand quantity and travel time to eliminate uncertainties.

To compare the routing solutions, we generate exponential random variables with the mean and variance specified above for the demand and travel time. For each generated scenario we solve a linear optimization problem to obtain the quantities of supply that minimize the total unmet demand with fixed routing solutions obtained above and constrained by the deadline and the total available quantity at the depot. The comparison shows that out of the 50 test cases, the deterministic routes generate 18 unmet demand cases with an average unmet demand of 9.94 while the chance-constraint routes only generate 2 unmet demand cases with an average unmet demand of 5.50. The chance-constraint routes outperform the deterministic ones because of the conservative nature of the chance-constraint model, which leads to balanced routes with similar number of nodes. The deterministic routes are more prone to have uneven number of nodes on different paths. We observe that this property makes the chance-constraint solution more robust and competitive than the deterministic one especially for the test cases that deviate far away from the mean value.

6 Conclusions and Future Directions

Facility location and vehicle routing are important issues in designing the medical supply distribution system, particularly for large-scale emergencies. This chapter has two primary goals. The first is to review different location models and vehicle routing models in the literature that are related to regular emergency services such as police, fire, etc. The second goal is to present tailored location and vehicle routing models to design rapid distribution systems of medical supplies in response to a large-scale emergency. An illustrative example of an anthrax emergency was discussed to show how the proposed models can be used to determine the facilities locations and vehicle routes for rapid medical supply distribution during the emergency.

In this chapter, we consider an emergency due to an anthrax attack as a representative large-scale emergency. We discuss the characteristics of large-scale emergencies and their requirements for the facility location and vehicle routing problems in the context of this particular emergency. However, other types of emergencies (e.g., chemical incident, dirty bomb attack, contagious disease outbreak) may involve different characteristics and thus will lead to different requirements on the problem formulations and solutions. For example, an emergency caused by a dirty bomb attack may impact not only the population, but also the medical supply facilities themselves. Therefore, reduced service capability of the facilities needs to be taken into account. A chemical incident may need instantaneous medical service to the infected people, and therefore medical supplies may need to be pre-positioned at a local level for immediate deployment. An open research question is how to develop an overall response plan that takes into consideration all the different possible scenarios. Is it more efficient and cost effective to develop a single strategy that is robust to the different possibilities or is it better to develop a separate plan for each possible emergency?

Another research direction is to develop efficient algorithms to solve the facility location and vehicle routing problems. In this chapter, the formulated problems were of relatively small size (i.e., 30 eligible facility sites, 10 selected staging areas, 1 central depot, and 3 vehicles) so the optimal solutions could be readily found using commercially available optimization software. However, for modeling more realistic and larger scenarios, the problem size of the models will increase significantly so that it becomes computationally prohibitive to obtain an optimal solution. Future research direction should also focus on developing efficient heuristics which can identify near optimal solutions to the large problems within a reasonable computational time.

Acknowledgement This research was supported by the US Department of Homeland Security through the Center for Risk and Economic Analysis of Terrorism Events (CREATE), grant number EMW-2004-GR-0112. However, any opinions, findings, and conclusions or recommendations in this document are those of the authors and do not necessarily reflect views of the US Department of Homeland Security. Also the authors wish to thank Richard Larson, Harry Bowman, and Terry O'Sullivan for their valuable input and comments for the improvement of this paper.

References

Batta, R. (1989). A queueing-location model with expected service time-dependent queueing disciplines. *European Journal of Operational Research, 39*, 192–205.

Batta, R., Dolan, J., & Krishnamurthy, N. (1989). The maximal expected covering location Problem: Revisited. *Transportation Science, 23*, 277–287.

Belson, D. (2005). *Storage, distribution and dispensing of medical supplies*. CREATE Interim Report.

Benedict, J. (1983). *Three hierarchical objective models which incorporate the concept of excess coverage for locating EMS vehicles or hospitals*. MSc thesis, Northwestern University.

Beraldi, P., & Ruszczynski, A. (2002). A branch and bound method for stochastic integer problems under probabilistic constraints. *Optimization Methods and Software, 17*, 359–382.

Berman, O., & Larson, R. C. (1985). Optimal 2-facility network districting in the presence of queuing. *Transportation Science, 19*, 261–277.

Bertsimas, D. (1988). *Probabilistic combinational optimization problems*. PhD thesis, Operation Research Center, Massachusetts Institute of Technology, Cambridge, MA.

Bertsimas, D. (1992). A vehicle routing problem with stochastic demand. *Operation Research, 40*, 574–585.

Bianchi, C., & Church, R. (1988). A hybrid FLEET model for emergency medical service system design. *Social Sciences in Medicine, 26*(1), 163–171.

Brandeau, M., & Larson, R. C. (1986). Extending and applying the hypercube queueing model to deploy ambulances in Boston. In A. Swersey & E. Ignall (Eds.), *Delivery of urban services*. New York: North Holland.

Branke, J., Middendorf, M., Noeth, G., & Dessouky, M. M. (2005). Waiting strategies for dynamic vehicle control. *Transportation Science, 39*, 298–312.

Burwell, T., Jarvis, J., & McKnew, M. (1993). Modeling co-located servers and dispatch ties in the hypercube model. *Computers and Operations Research, 20*, 113–119.

Carbone, R. (1974). Public facility location under stochastic demand. *INFOR, 12*, 261–270.

Carson, Y., & Batta, R. (1990). Locating an ambulance on the Amherst campus of the State University of New York at Buffalo. *Interfaces, 20*, 43–49.

CDC website. (2005) http://www.bt.cdc.gov/agent/anthrax/.

Chapman, S.C., & White, J.A. (1974). *Probabilistic formulations of emergency service facilities location problems*. Paper Presented at the ORSA/TIMS Conference, San Juan, Puerto Rico.

Church, R., & ReVelle, C. (1974). The maximal covering location problem. *Papers of the Regional Science Association, 32*, 101–118.

Daskin, M. (1983). The maximal expected covering location model; Formulation, properties and heuristic solution. *Transportation Science, 17–1*, 48–70.

Dror, M. (1989). Vehicle routing with stochastic demands: Properties and solution framework. *Transportation Science, 23*, 166–176.

Dror, M. (1993). Modeling vehicle routing with uncertain demands as a stochastic program: Properties of the corresponding solution. *European Journal of Operational Research, 64*, 432–441.

Dror, M., & Trudeau, P. (1986). Stochastic vehicle routing with modified savings algorithm. *European Journal of Operational Research, 23*, 228–235.

ESRI website. (2005). http://reports.esribis.com/.

Garfinkel, R. S., Neebe, A. W., & Rao, M. R. (1977). The m-center problem: Minimax facility location. *Management Science, 23*, 1133–1142.

Gendreau, M., Laporte, G., & Seguin, R. (1995). An exact algorithm for the vehicle routing problem with stochastic demands and customers. *Transportation Science, 29*, 143–155.

Goldberg, J., Dietrich, R., Chen, J. M., & Mitwasi, M. G. (1990). Validating and applying a model for locating emergency medical services in Tucson, AZ. *European Journal of Operational Research, 49*, 308–324.

Goldberg, J., & Paz, L. (1991). Locating emergency vehicle bases when service time depends on call location. *Transportation Science, 25*(4), 264–280.

Hadjiconstrantinou, E., & Roberts, D. (2002). Routing under uncertainty: An application in the scheduling of field service engineers. In P. Toth & D. Vigo (Eds.), *The vehicle routing problem* (pp. 331–352). Philadelphia: SIAM Monographs.

HDMA, Healthcare Distribution Management Association. (2005) http://www.healthcaredistribution.org/

Hochbaum, D. S., & Pathria, A. (1998). Locating centers in a dynamically changing network and related problems. *Location Science, 6*, 243–256.

Hogan, K., & ReVelle, C. (1986). Concepts and applications of backup coverage. *Management Science, 32*, 1434–1444.

Holmberg, K. (1999). Exact solution methods for uncapacitated location problems with convex transportation costs. *European Journal of Operational Research, 114*, 127–140.

Jaillet, P. (1987). Stochastic routing problem. In P. S. Mason & F. Andreatta (Eds.), *Stochastics in combinatorial optimization world scientific*. Oxford: Oxford University Press.

Jaillet, P., & Odoni, A. (1988). The probabilistic vehicle routing problem. In B. L. Golden & A. A. Assad (Eds.), *Vehicle routing: Methods and studies*. Amsterdam: North-Holland.

Jain, K., Mahdian, M. and Saberi, A. (2002). *A new greedy approach for facility location problems*. Proceedings of the 34th ACM Symposium on Theory of Computing, 731–740.

Jaramillo, J. H., Bhadury, J., & Batta, R. (2002). On the use of genetic algorithms to solve location problems. *Computers and Operations Research, 29*, 761–779.

Jarvis, J. P. (1977). Models for the location and dispatch of emergency medical vehicles. In T. R. Willemain & R. C. Larson (Eds.), *Emergency medical systems analysis*. Lexington, MA: Lexington Books.

Jezequel, A. (1985). *Probabilistic vehicle routing problems*. Master's thesis, Department of Civil Engineering, Massachusetts Institute of Technology.

Jia, H., Ordonez, F., & Dessouky, M.M. (2005). *A modeling framework for facility location of medical services for large-scale emergencies*. USC ISE Working paper #2005-01. Under revision in Special Issue IIE Transactions on Homeland Security.

Jula, H., Dessouky, M.M., & Ioannou, P. (2005). An approximate solution for the TSPTW with stochastic travel and service times, to appear IEEE Transactions on ITS.

Lambert, V., Laporte, G., & Louveaux, F. (1992). Designing collection routes through bank branches. *Computers and Operations Research, 20*, 783–791.

Laporte, G., Louveaux, F., & Mercure, H. (1992). The vehicle routing problem with stochastic travel times. *Transportation Science, 26*, 161–170.

Larson, R. C. (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers and Operations Research, 1*, 67–95.

Larson, R. C. (1975). Approximating the performance of urban emergency service systems. *Operations Research, 23*(5), 845–868.

Larson, R. C. (1979). Structural system models for locational decisions: An example using the hypercube queueing model. In K. B. Haley (Ed.), *Operational research 78, proceedings of the eighth IFORS international conference on operations research*. Amsterdam: North-Holland Publishing Co.

Mandell, M.B. (1998). *A/'-median approach to locating basic life support and advanced life support units*. Presented at the CORS/INFORMS National Meeting, Montreal, April, 1998.

Marianov, V., & ReVelle, C. (1996). The queueing maximal availability location problem: A model for the siting of emergency vehicles. *European Journal of Operational Research, 93*, 110–120.

McAllister, J. C. (1985). Challenges in purchasing and inventory control. *American Journal of Hospital Pharmacy, 42*, 1370–1373.

Mirchandani, P. B. (1980). Locational decisions on stochastic networks. *Geographical Analysis, 12*, 172–183.

Paluzzi, M. (2004). *Testing a heuristic P-median location allocation model for siting emergency service facilities*. Paper Presented at the Annual Meeting of Association of American Geographers, Philadelphia, PA.

Rebidas, D., Smith, S. T., & Denomme, P. (1999). Redesigning medication distribution systems in the OR. *AORN Journal, 69*, 184–190.

Repede, J. F., & Bernardo, J. J. (1994). Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research, 40*, 58–69.

ReVelle, C. (1989). Review, extension and prediction in emergency service siting models. *European Journal of Operational Research, 40*, 58–69.

ReVelle, C., & Hogan, K. (1986). A reliability constrained siting model with local estimates of busy fractions. *Environment and Planning, B15*, 143–152.

ReVelle, C., & Hogan, K. (1989a). The maximum availability location problem. *Transportation Science, 23*, 192–200.

ReVelle, C., & Hogan, K. (1989b). The maximum reliability location problem and a-reliable *p*-center problem: Derivatives of the probabilistic location set covering problem. *Annals of Operations Research, 18*, 155–174.

ReVelle, C., Schweitzer, J., & Snyder, S. (1996). The maximal conditional covering problem. *IN FOR, 34*, 77–91.

Rubin, H., & Keller, D. D. (1983). Improving a pharmaceutical purchasing and inventory control system. *American Journal of Hospital Pharmacy, 40*, 67–70.

Schilling, D. A. (1982). Strategic facility planning: The analysis of options. *Decision Sciences, 13*, 1–14.

Secomandi, N. (2001). A rollout policy for the vehicle routing problem with stochastic demands. *Operation Research, 49*, 796–802.

Serra, D., & Marianov, V. (1998). The *P*-median problem in a changing network: The case of Barcelona. *Location Science, 6*(1), 383–394.

Shen, Z., Dessouky, M., & Ordonez, F. (2005). *Stochastic vehicle routing problem for large-scale emergencies*. ISE Working paper #2005–02.

Stewart, W., & Golden, B. (1983). Stochastic vehicle routing: A comprehensive approach. *European Journal of Operational Research, 14*, 371–385.

Talmar, M. (2002). *Location of rescue helicopters in South Tyrol*. Paper Presented at 37th Annual ORSNZ Conference, Auckland, New Zealand.

Toregas, C., Swain, R., ReVelle, C., & Bergman, L. (1971). The location of emergency service facility. *Operations Research, 19*, 1363–1373.

White, J., & Case, K. (1974). On covering problems and the central facility location problem. *Geographical Analysis, 6*, 281–293.

Part V
Implementing Change

Chapter 17
Using a Diagnostic to Focus Hospital Flow Improvement Strategies*

Roger Resar

Abstract Current methods to evaluate hospital flow primarily measure micros-system issues, such as numbers of diversions from the ED, or how many patients are boarded in some way somewhere in the hospital, or specific delays in hospital units and annual admissions. The narrow focus on delays, although appropriate, generally leads to attempts at fixing a large system problem from the viewpoint of a micro-system. A newly developed hospital diagnostic that utilizes a broader view of flow can be used to evaluate how a hospital might best respond to delays, but also how to maximize the number of turns each bed generates in a year. The diagnostic utilizes easily obtainable and commonly collected hospital metrics. The diagnostic allows a hospital to categorize itself depending on the number of bed turns and the efficiency of using those beds. Based on a self-evaluation a hospital can determine which improvement strategies would be most useful, or if ongoing improvement strategies are properly focused to achieve an ultimate goal of reduced delays and increased bed turns (the goal of 90 or more adjusted turns and a bed use efficiency of around 90 %). A business case for improvements can be based on increasing turns either by accommodation of more demand or by decreasing unneeded capacity. Improvements in throughput per bed allow leadership to justifiably focus resources on the appropriate change concepts. Serial measurements of the bed turns and bed utilization metric allow the organization to measure the effects of flow improvement strategies over time.

Keywords Bed turns • Hospital efficiency • Flow streams • Delays • Flow diagnostic

*The work presented here is from the efforts of the IHI Innovation Team.

R. Resar (✉)
Institute for Healthcare Improvement, 20 University Road, 7th Floor,
Cambridge, MA 02138, USA
e-mail: Roger.resar@mayo.edu

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_17, © Springer Science+Business Media New York 2013

1 Introduction

In 2002 the Institute for Healthcare Improvement (IHI) Innovation Team started work on hospital flow and the hospital flow diagnostic. Although the characteristics of the diagnostic and how the diagnostic is used continue to develop, it has become an important focus for the direction of flow improvement work for many hospitals. Initially the diagnostic was designed to direct hospitals towards specific change concepts that would be most appropriate for their particular flow problems. Currently the diagnostic is first being used as a high-level hospital evaluation and then second as an entry point to improvement strategies. The diagnostic can validate a given approach to flow improvement in a hospital and provide a measurement overtime of eventual success of the improvement work. Prior to the use of the diagnostic it was far too easy to tackle hospital delays as isolated unit specific problems. Trying to solve a system problem by working on a single unit has proven to be very wasteful and foolhardy since the best plans of a single unit redesign (since they are connected to the rest of the hospital) can be ruined by changing practices of upstream or downstream units. For example, the best of efforts in the ICU can be totally discounted by changing policies on admissions or transfers to a step down telemetry unit. So rather than making the ICU responsible for delays, the hospital flow system needs to be accountable for delays. It is now clear hospital flow is a system problem and needs to be looked at from a more global point of view.

The chapter will describe the methodology of the hospital diagnostic, demonstrate the degree of variability in the number of bed turns and bed utilization, and lastly suggest strategies for improvement and introduce the use of flow streams to work on delays. Patient flow management is a very complex problem not limited to just patient flow within the hospital. It also includes study of pre-hospital patient flow (e.g., why patients use the ED rather than their primary care physicians) and post-hospital patient flow (e.g., availability of rehabilitation or skilled nursing facility beds). Comprehensive coverage of all these issues would not be possible in this chapter. Those who are interested in those issues are referred to the references at the end of this chapter.

2 History

Characteristically most hospital flow solutions relate to fixing the “squeaky wheel”, commonly entwined with demands from high producers, or by looking at isolated unit problems with solutions from a narrow point of view. Both usually accomplish nothing in terms of overall hospital flow.

Both solutions, however, usually assume demand is greater than capacity by measuring delays. One only has to look at the building boom in hospitals to understand the agreement to the general assumption that demand is far greater than capacity. These approaches might lead an occasional hospital to the right solution, but luck or happenstance not deliberate design would be responsible. Another possible solution to hospital flow has been the approach of the flow communities working with the IHI. The hospital diagnostic was used by each hospital as a self-evaluation. A description of the hospital flow problem was generated. Next each hospital was encouraged to select appropriate strategies for improving hospital flow. In this scenario new or interesting ideas, such as discharge appointments or extending the flow chain, are tested knowing the appropriateness of the intervention based on the hospital diagnostic and how in many cases how it will be applied specifically to a high volume flow stream.

The hospital diagnostic was developed by the IHI Innovation Team with a specific set of outcomes around bed turns. For those hospitals with a high number of turns (greater than 90), the goal is to maintain or increase turns, but with minimal delays. For those hospitals with low bed turns, increasing the number of bed turns by capturing deflected demand or reducing unneeded capacity is the goal.

The hospital diagnostic views flow from a viewpoint of the whole system including beyond the walls of the hospital in some cases. One viewpoint that will not be examined is how adequately the hospital services are fulfilling the needs of the community as a societal resource. Hospital throughput will be measured as bed turns and revenue per bed, as well as bed utilization (different than occupancy rate). The strategy will demand minimal delays as an adjunct to any improvement in turns and utilization.

The methodology, tools and other documents related to the topic of hospital flow can be found on the IHI Web site (www.ihi.org). The following definitions are used in the diagnostic calculation:

- *Unadjusted Bed Turns* = (Admissions + OIIB)/by functional beds.
- *OIIB* Stands for any outpatient in an inpatient bed. These may include the standard observation patients in an inpatient bed, in addition to any other use of an inpatient bed without being an actual admission. Common uses might be post heart catheterization observation, post outpatient endoscopy procedure observation, etc. Since the bed is being used even for a short period of time, it does force admissions to compete for the bed and must be counted. (A good method to determine this is to take several weeks and just survey each unit once each 24 h for any outpatient use of an inpatient bed and average it out for the year.)
- *Adjusted Bed Turns* = (Admissions × case-mix-index)/functional beds + OIIB/functional beds.
- *Utilization* (efficiency) = unadjusted bed turns divided by potential bed turns. This number is different than occupancy, since occupancy is measured at midnight and usually only once per day.
- *Potential Bed Turns* = 365 divided by aggregate average length-of-stay (LOS).
- *Average Length of Stay (LOS):* Average overall length of stay for inpatient admissions. This does not include the standard observation classification and should exclude newborns, which most hospitals do automatically.
- *Admissions:* The number of inpatient admissions (excluding newborns).

- *Observations:* Those patients defined by billing purposes as not being an admission, but being cared for in the hospital for 24 h or less. If they become an admission they should not be counted twice when the OIIB calculation is made.
- *Case Mix Index (CMI):* All payer case mix as defined by CMS (Center for Medicare and Medicaid Services).
- *Bed Turns:* Number of times (to the nearest whole number) an inpatient bed is used by a patient whether for admission or for any type of OIIB.
- *Functional Beds:* The number of beds normally open and staffed in a hospital over a year on average, excluding newborn and delivery beds, but including all other beds. The beds must be staffed around the clock which would exclude admission units etc. For rehab beds they are counted if under the same Medicare number.

The performance metrics from Sect. 3 are now demonstrated with example calculations.

Unadjusted Bed Turns = (admissions + OIIB) divided by functional beds.
Example:

Admissions:	10,000/year
Observations:	1,000/year
Functional beds:	200

10,000 + 1000 = 11,000/200 = 55 turns on average per bed per year

Adjusted Bed Turns = ((Admissions × Case Mix Index) + OIIB)/functional beds.
Example:

Admissions	10,000/year
Observations	1,000/year
Functional beds	200
Case Mix Index	1.4

10,000 × 1.4 = 14,000 + 1,000 = 15,000/200 = 75 turns on average/bed/year

Potential Bed Turns = 365/LOS. Example:
LOS 4.0 days
365/4.0 = 91 potential bed turns per year

Utilization = Unadjusted Bed Turns/Potential Bed Turns.
In the above example:
Unadjusted Bed Turns = 55
Potential Bed Turns = 91
Utilization = 55/91 = 60.4 %

3 Observations

Figure 17.1 shows the data submitted by hospitals in the IHI Flow Community in 2004. The plot uses adjusted bed turns versus bed utilization percentage. None of the hospitals were unique specialty hospitals, although there was a mixture of academic and nonacademic hospitals. The number of actual turns adjusted by the case mix index attempted to level the field by consideration of severity. Even with the adjustment for severity of illness considerable variation exists between these hospitals. The IHI team looked at the plot and selected 90 adjusted bed turns and 90 % utilization as desirable goals. The turns and utilization goals were based on reviewing best performers and using the knowledge of queuing theory. Based on these guidelines, the data collected from the hospitals place each hospital into one of four quadrants. Representative samples of hospitals from each quadrant were studied and some visited to learn about the characteristics of the hospitals in each of the quadrants.

One obvious observation needing an explanation is utilization well over 100 %. Obviously consistent over 100 % utilization is difficult to imagine. The partial answer is the fact that the calculation for utilization uses bed turns. The bed turns associated with OIIB do not have a LOS measure since many of the stays are only for a few hours. Hospitals with very high OIIB fractions commonly will have unusually high utilization rates. If a hospital has a very high OIIB rate with significant delays, this particular flow stream needs to be examined for opportunity.

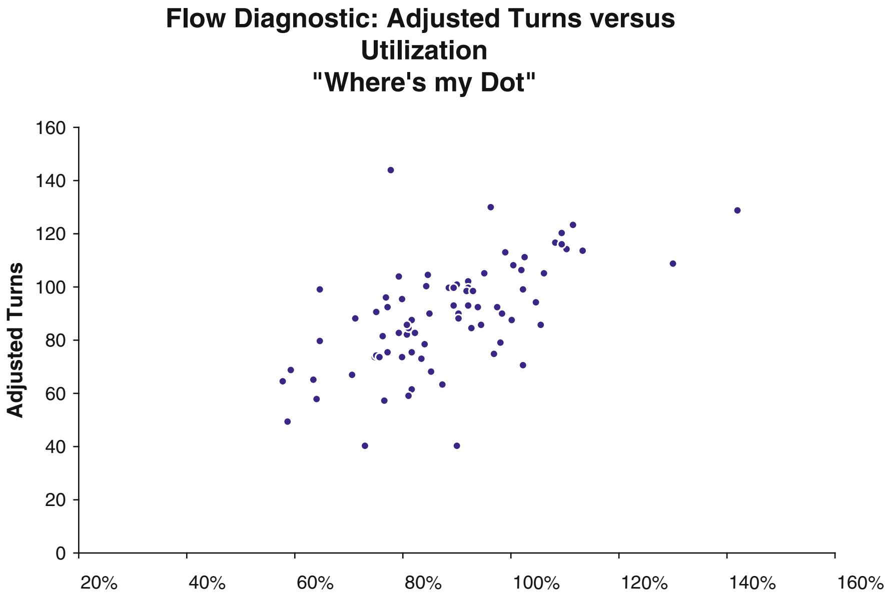


Fig. 17.1 Flow diagnostic: adjusted turns versus utilization, “Where’s My Dot”

Figure 17.1 also gives the visual representation of tremendous variability between hospitals in terms of flow and utilization. Yet each of these hospitals have a common problem of significant flow delays otherwise they would not have joined the collaborative. It is hard to imagine how a hospital with only 40 turns and a utilization of 60 % should have delays, while it is very understandable why a hospital with 140 turns and 100 % utilization should have delays. Those questions and the subsequent quest for the answers will form the basis for the rest of the chapter but should also tempt the reader to "find their dot".

4 Understanding Delays from the Flow Diagnostic

Figure 17.2 demonstrates the four quadrants and the relationship between capacity, demand and delays in each of the quadrants, in addition to the 90 turn and 90 % utilization goals.

Figure 17.3 portrays the expectation of delays. Quadrants 2 and 4 because of the high utilization should expect delays. Quadrants 1 and 3, because of the low utilization, should be surprised by delays. The utilization is given a window, as shown by the dotted lines.

Using Figure 17.4, we see quadrant 1 has a good balance between capacity and demand, but certain hospitals may still have significant inefficiency in utilization of

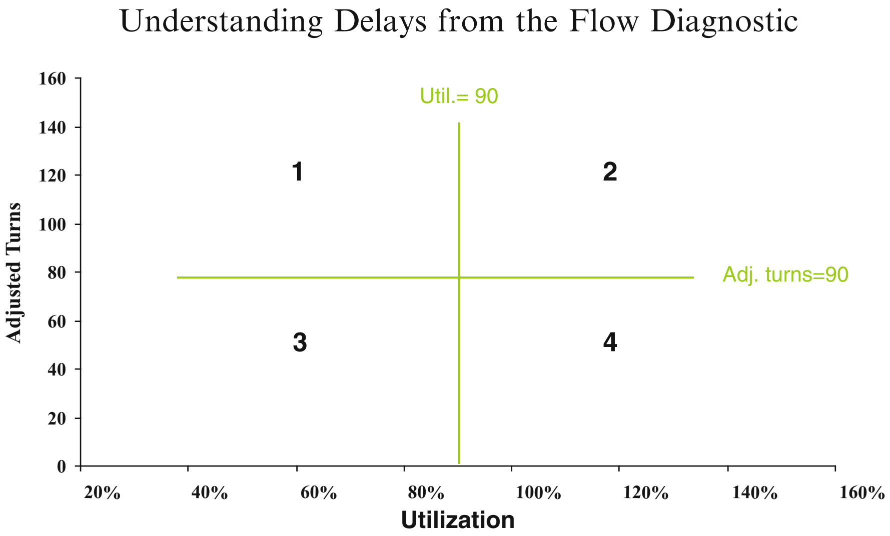


Fig. 17.2 Flow diagnostic: understanding delays

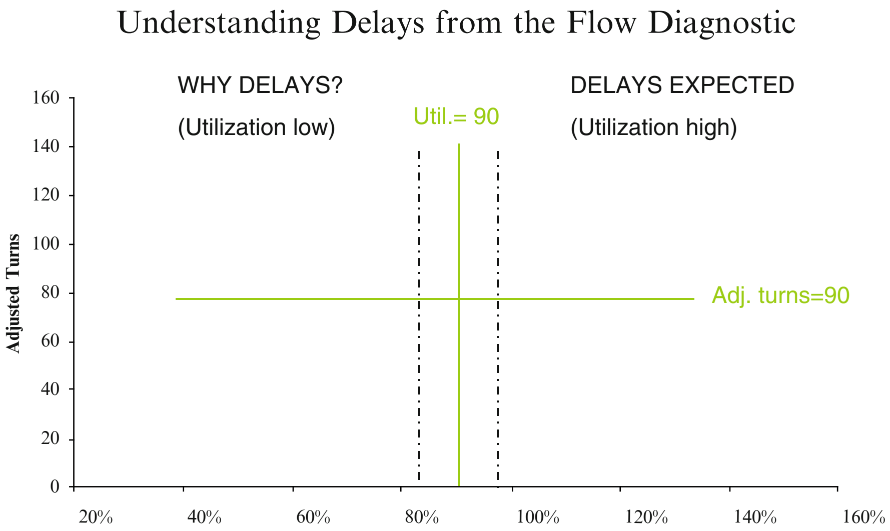


Fig. 17.3 Understanding delays from the flow diagnostic

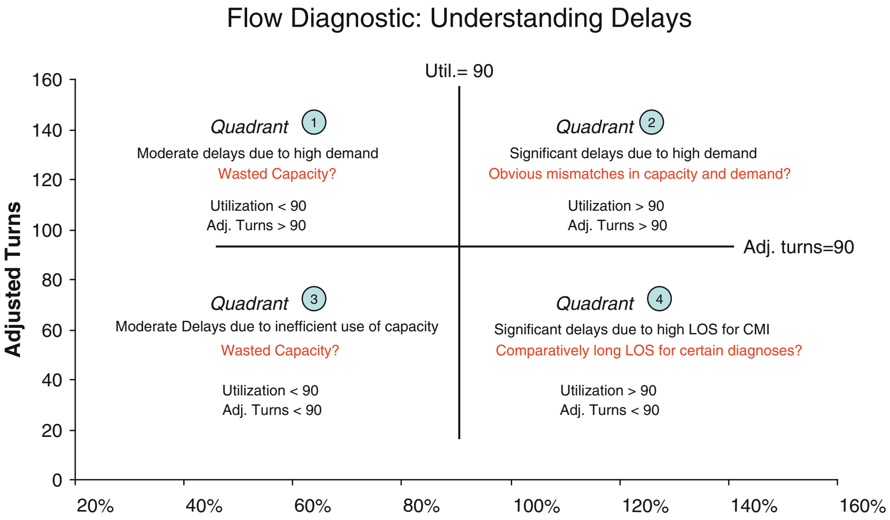

Flow Diagnostic: Understanding Delays
Util.= 90
Quadrant 1
Moderate delays due to high demand
Wasted Capacity?
Utilization < 90
Adj. Turns > 90
Quadrant 2
Significant delays due to high demand
Obvious mismatches in capacity and demand?
Utilization > 90
Adj. Turns > 90
Adj. turns=90
Quadrant 3
Moderate Delays due to inefficient use of capacity
Wasted Capacity?
Utilization < 90
Adj. Turns < 90
Quadrant 4
Significant delays due to high LOS for CMI
Comparatively long LOS for certain diagnoses?
Utilization > 90
Adj. Turns < 90
Adjusted Turns
0
20
40
60
80
100
120
140
160
20%
40%
60%
80%
100%
120%
140%
160%

Fig. 17.4 Flow diagnostic: understanding delays

varying degrees; all in our series still have delays. Quadrant 3 is an inefficient hospital in terms of utilization and also has low turns and delays. Both of these quadrants need to question delays from the viewpoint of “why”. With low utilization, one should expect no or minimal delays.

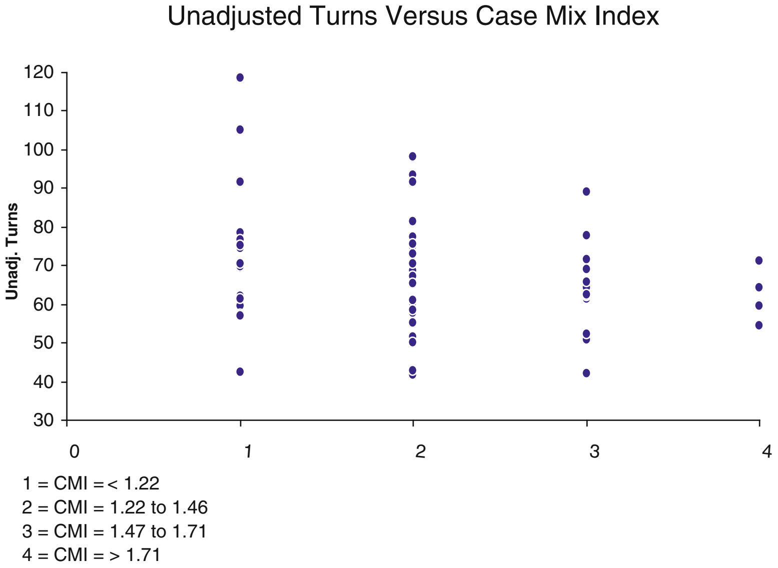


Fig. 17.5 Unadjusted turns versus case mix Index

Again using Fig. 17.4 Quadrant 2 hospitals are highly functional systems, but with delays. Quadrant 4 also high-functioning, in terms of utilization, but in both quadrants we should not be surprised to find delays—in fact they might be expected.

No one quadrant is without its own set of problems that need to be solved, but from a business case, being in quadrant 1 or 2 would be preferable, with quadrant 1 in the long run best from a queuing theory point of view. Participants in the flow collaborative—because of the nature of the hospitals that joined a collaborative on flow—all experienced significant delays and lost opportunity in one way or another, no matter in which quadrant their dot was placed.

Figure 17.5 shows hospitals in the IHI flow community. Each hospital is represented by a dot and the *y* axis is unadjusted turns. The *x* axis is the case mix index divided into four categories. The finding of wide variation within any given case mix grouping is extremely interesting. In category 1, for example, the unadjusted turns range from 40 to 120 turns/bed/year. Explaining this variation is very difficult. The assumption is that competent managers and executives work in each of these hospitals. In addition to the variation within a grouping, how can we explain a hospital with a case mix index of 1–1.22 and bed turns/year of 60, when another hospital with a case mix of 1.71 and obviously much sicker patients has the same number of bed turns? The answer may well be the proportionality in how specific hospitals handle streams of flow. The defining of the flow streams and the change concepts to improve the flow stream efficiency can be extrapolated in part from understanding the flow diagnostic.

5 Steps to Using the Hospital Diagnostic

This section provides a step-by-step explanation of how to apply the diagnostic.

Step 1

Collect the OIIB (remember these are any outpatients using inpatient beds), admissions (excluding newborns), case mix index for all payers, functional beds (see methodology below) and LOS. The data should be collected for a full year. If multiple years of data are available, you may wish to use this to establish a trajectory.

SideBar on methodology to determine functional beds. Most hospitals have licensed beds, staffed beds and even functional beds as different numbers. Usually the actual beds in use on a daily basis are less than the licensed beds. Functional beds can best be described as over a 3 or 4 month average how many beds are actually staffed. Difficulty arises when there is a great fluctuation with seasonality. Since seasonality usually is a recurring phenomenon, the 3 or 4 month average should include both high and low season. (Many hospitals could just average out the staffed beds for a whole year and get the functional bed number). Absolute accuracy is less important than stability in this measure. If stable over time, the change in bed turns can then be measured to reflect actual change.

Step 2

Calculate bed turns with and without adjustment for case severity using the case mix index, and calculate utilization.

SideBar Since the functional beds affects the bed turns fairly dramatically, the leeway within any quadrant allows this to be an average over months and then allows the comparison over time as a trajectory to be helpful.

SideBar: An alternative method to measure utilization, rather than actual bed turns divided by potential bed turns, is to directly measure the bed utilization. Each bed in the hospital is numbered. A random number generator for the total number of beds is created, and each hour for 1 week a bed number is generated and a determination is made as to whether the bed is filled or not. The resulting utilization is then calculated as a simple total filled over the total number of observations. A few rules apply, however:

1. Beds held for a patient in surgery is not a filled bed.
2. If patient is no longer alive, bed is not a filled bed
3. Patient not in room but down for procedure is a filled bed, but patient needed to be in the room prior to the procedure or test
4. Room being cleaned is not a filled room

Step 3

Place your hospital into a quadrant based on your calculations. Be careful if you are at the dividing lines since there is some overlap.

Step 4

Evaluate your current flow improvement strategy and see if it is consistent with the quadrant recommendations

Step 5 (Optional)

Evaluate previous years and determine the trajectory of your change based on bed turns and utilization. Is it going in the right direction? In this case, is it moving to the left upper quadrant as improvements are being made in both bed turns and delays?

6 Strategies

Although the aim of the chapter is not to provide a detailed discussion of the specific strategies associated with flow improvement, the appropriate use of the diagnostic still demands an understanding of the approaches to be taken for the specific strategies. In that light, the specific strategies need to be at least understood. The overall aim is to increase throughput (as measured by bed turns) and minimize delays (as measured by time) while assuring that high performance in flow is not at the expense of poor quality. The specific delay measurements commonly used in the Flow Collaborative are:

- Time from entering the Emergency Department to the floor (Using a door to floor concept)
- Diversions from the Emergency Department (Or alternatively hours on diversion)
- Left without being seen from the Emergency Department

In those hospitals with turns already high, the goal is to maintain or increase turns but now with minimal delays. In those hospitals with low bed turns the goal is to increase turns by capturing deflected demand or reducing capacity. Figure 17.6 presents strategies for quadrants 1 and 3

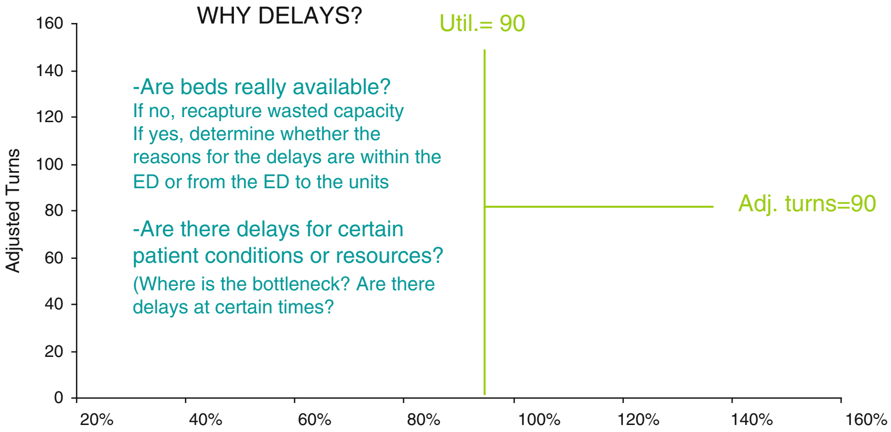


Fig. 17.6 Understanding delays from the flow diagnostic

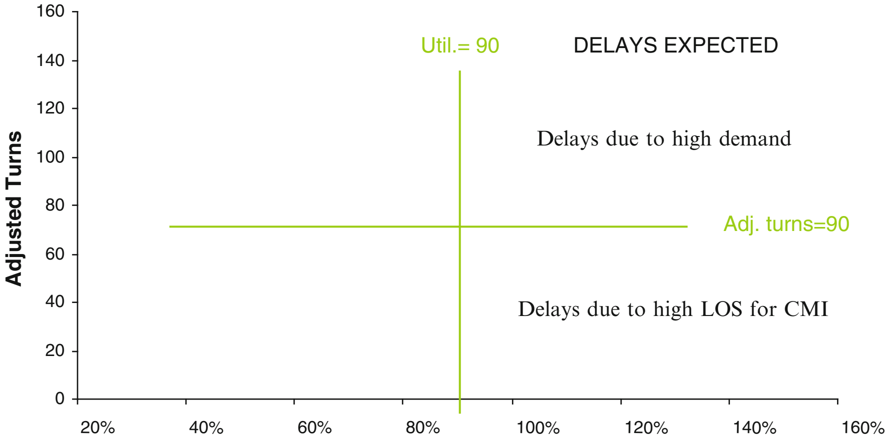


Fig. 17.7 Understanding Delays from the flow diagnostic, Quadrants 2 and 4

Quadrants 1 and 3 have poor utilization in common. The best quadrant is 1 but without delays. The worst quadrant situation is 3. Both need to determine whether beds are available and since in all likelihood they are the change concepts need to be related to recapture of wasted capacity. Large volume flow streams are essential starting points. These might include the cardiac service or the orthopedic service as examples. Ultimately those in quadrant 3 will need to increase turns and this might be done by increasing utilization and closing unneeded capacity. Figure 17.7 presents strategies for quadrants 2 and 4

Flow is a property of the entire system and can only be optimized at a system level. Understanding this fact has led to the idea of streams of flow that are then optimized by the specific change strategies.

In order to start the work on flow, a sound administrative system that manifests a bed management process that incorporates planning based on predictions of capacity and demand should be an overarching design in any improvement strategy. The chapter by Linda Kosnik demonstrates an hour by hour evaluation of capacity and demand with explicit processes designed to specific changes in demand or capacity. Some organizations are using bed huddles several times a day in addition to trying to understand larger issues of variation such as seasonality to adjust capacity and demand. A bed administrative system may or may not incorporate an electronic bed board product.

The process of where to go from the diagnostic can best be described in three steps:

1. Determine the major flow stream that will be subjected to the flow change strategies.
2. Rather than working on a given unit look at the project as a system problem crossing multiple silos of interest. A stream of flow should manifest a large volume, exhibit significant delays and flow from at least the admission through the discharge, although the work done by Mark Lindsay would suggest moving this out to the pre-admission and the post-discharge

3. Once a major flow stream has been identified, the specific flow improvement strategies now need to be applied, particularly those that are demonstrated as useful for the hospital's quadrant.

6.1 Quadrant 1 (Left Upper Quadrant)

The quadrant has, in general, adequate turns and reasonable utilization, but still has moderate delays due to high demand. Adjusted bed turns are greater than 90 and utilization is less than 90 %. The primary problem is waste of capacity. Any bed use variant could contribute to the wasted capacity. The wasted capacity would be easiest to pinpoint by looking at several large volume flow streams. Examples of a large volume stream might be orthopedic surgery or heart disease requiring medical telemetry. Obviously, the large volume stream work should focus on reducing wasted capacity by identifying and remove bottlenecks in the flow stream. The bottlenecks are commonly admissions, transfers or discharges and specific change concepts for each of these interfaces. For example, the use of discharge appointments could be linked to either transfers or admissions. Detailed information on this and other concepts can be found at www.ihi.org.

6.2 Quadrant 2 (Right Upper Quadrant)

The quadrant has adequate turns, but the utilization is so high that staff burnout and safety issues need to be considered. These are super high functioning hospitals. Adjusted turns are greater than 90 % and utilization is greater than 90 %. Significant delays are seen due to mismatch of capacity with a high demand, more than wasted capacity. Hospitals in this quadrant may truly have a need to add more capacity, but only after significant efforts have been made to correct the mismatch between capacity and demand. The mismatch is significant enough that use of electronic bed tracking systems may have tremendous utility for organizations in quadrant 2. Poor habits need to be purged from the system such as holding beds, lack of good discharge planning and lack of pre-admission planning for elective surgery. Again, the emphasis should be to start with larger flow streams and apply these change concepts.

6.3 Quadrant 3

Hospitals in quadrant 3 present a very unique opportunity. They experience moderate delays due to inefficient use of capacity. In most cases they have more beds than necessary, but due to the inefficient use of the capacity staff keep too many beds open. The result is adjusted turns less than 90 and utilization less than 90 %.

Again, the concept of isolating several major flow streams and applying some of the change principles would be the strategy. Those change concepts include eliminate bed holds, developing bed cleaning and turnaround strategies, reducing internal transfers, decreasing use of inpatient beds by outpatients or decreasing capacity. Most hospitals have an initial response of shock when a suggestion is made to reduce capacity, but the reduction in capacity reduces the staffed beds and immediately increases turns per bed and utilization of the beds remaining.

6.4 Quadrant 4

Quadrant 4 has significant delays due to a high LOS for the case mix index. Adjusted turns are less than 90 and utilization is greater than 90. Almost all the hospitals studied have a comparatively long length of stay for multiple high volume diagnoses. The utilization of flow streams would necessarily need to focus on one of those high volume diagnoses. The focus should be on conditions that have a long LOS compared to other hospitals. The strategy needs again to look at the bottlenecks. A commonly found bottleneck is the inability to work outside the hospital for certain chronic disease states, such as chronic ventilator use. An easy measurement is to look at the LOS for all patients admitted to a nursing home. Hospitals average 5–7 days around the country. Longer than 7 days suggests that flow efforts need to extend outside the walls of the hospital.

The business case for increasing turns can be illustrated in the following example. The data are from a real hospital.

- Admissions: 16,704 with 4,246 observations
- Case mix index: 1.49
- LOS: 4.7
- Average functional staffed beds 432
- Average revenue (actual collections) from an admission = $7,525
- Unadjusted Turns = (Admissions + Observations)/Average Functional Beds = 48 Turns/year
- Adjusted Turns = (Admissions × CMI + Observations)/Average Functional Beds = 67 Turns
- Potential Bed Turns = 365/LOS = 77 Turns
- Hospital efficiency = Adjusted Turns/Potential Turns = 74 %

High performing hospitals have 90 or more adjusted turns and an efficiency of 90 %. Adding 10 turns per year per bed produces $32,000,000 in revenue per year, assuming same average revenue per admission. Increasing turns to the level of high performing hospitals (89 turns) produces $60,000,000 in revenue per year.

Figure 17.8 is an example of using the diagnostic retrospectively to determine whether the appropriate flow improvement strategies were selected and the impact on turns. Initially the hospital was near the intersection of quadrant 2 and 4. Turns were reasonably high and LOS was actually low for the case mix index. Since the

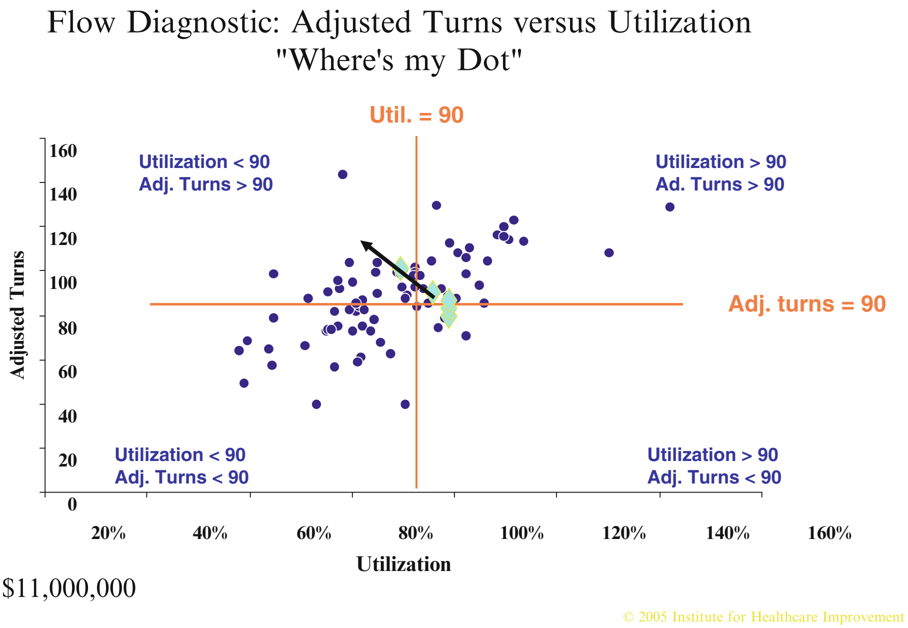


Fig. 17.8 Flow diagnostic, adjusted turns versus utilization

demand was high existing capacity needed to be optimized. An effective administrative system was added with an hour-by-hour response system. Discharge appointments were synchronized. Extensive collaboration with nursing homes was highly developed to provide continuum of care. Over the 4 years of measurement turns increased by 20 per bed, while at the same time the case mix index edged up slightly. At the same time the utilization decreased into a more reasonable sustainable pace from a little over 100–86 %. The financial benefit was approximately $11 million to the organization.

7 Conclusions

The hospital diagnostic can be used to initially determine a hospital's status in regard to utilization and throughput. General high level conclusions regarding these measures will set up the next step, which is the determination of high volume flow streams and the application of specific change concepts. The diagnostic in itself is not a change strategy. It allows for an initial evaluation and a methodology to follow a trajectory over time. The business case for increasing turns (throughput) is very compelling.

Acknowledgments The author would like to acknowledge the contribution of the IHI Innovation Team particularly the flow design team: Tom Nolan, Carol Haraden, Kirk Jensen, Diane Jacobsen, Eugene Litvak, and Kevin Nolan.

References

Institute for Healthcare Improvement. Patient Flow, http://www.ihi.org/IHI/Topics/Flow/PatientFlow/

Institute for Healthcare Improvement. Patient flow, Discussion, http://www.ihi.org/ihi/forums/ShowPost.aspx?PostID=1691

Joint Commission on Accreditation of Healthcare Organizations. (2004). Managing Patient Flow. Strategies and Solutions for Addressing Hospital Overcrowding. Joint Commission Resources.

Chapter 18
Improving Patient Satisfaction Through Flow

Kirk Jensen

Abstract Patient satisfaction has emerged as a key driver of health care service, quality and even patient safety. The increased prevalence of and reliance on patient satisfaction surveys and the patient experience has permeated health care practices ranging from solo physician offices to tertiary health care centers. The recent deployment of the HCAHPS (*Hospital Consumer Assessment of Healthcare Providers and Systems*) Survey and "pay for performance" initiatives have served to highlight an increased emphasis on national standards, public reporting, and hospital comparisons for patient satisfaction.

Patient satisfaction and patient flow are inextricably entwined aspects of patient care. This chapter describes factors behind the patient experience: what is assessed on surveys, the role of structure, process and teamwork, and the contributions of smooth patient flow. A number of strategies and tactics for enhancing and sustaining an optimal patient experience are discussed. Finally, for situations when waits are inevitable, this chapter reviews what can be done to manage the processes and expectations of the patient and the health care team.

Keywords Customer service • Customer service diagnosis • Communication • Motivation • Patient flow • Patient satisfaction • Psychology of waiting • Service recovery • Teams

1 Introduction

Is someone entering an emergency department or hospital a patient or a customer? My colleague Thom Mayer answers this question with an equation: the more horizontal the person, the more that person is a patient; the more vertical the person,

K. Jensen (✉)
BestPractices, 10306 Eaton Place, Ste. 180, Fairfax, VA 22030, USA
e-mail: kjensen@best-practices.com

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_18, © Springer Science+Business Media New York 2013

the more that person is a customer (Mayer and Cates 2004). A 17-year-old boy entering an emergency department with a sprained wrist is vertical; he does not need a bed and he will not require much time or many resources to treat. For the most part, he is a customer. An 85-year-old woman with severe abdominal pain is horizontal; she will need a bed and will likely require considerable time and multiple resources to be diagnosed and treated. She is primarily a patient.

Understanding the difference between customers and patients is critical for providing patient care *and* customer service. Someone who is vertical most values speed and convenience in the health care experience. These are customer service values. Someone who is horizontal most values good care—saving these persons' lives or limbs, heading off a critical diagnosis, keeping them safe. These are patient care values. Health care team members would do well to remember that both values matter, and that responding to the desire for speed and convenience is an important part of treating vertical patients. Accordingly, an Emergency Department (ED) or hospital process needs to include not just a *clinical diagnosis* of the patient's condition but a *customer service diagnosis* as well. The customer service diagnosis evaluates to what degree this person is a patient and to what degree a customer, treating the person in light of that diagnosis.

2 The Importance of Patient Satisfaction Surveys: Implications for Customer Service

There has been resistance among health care providers to placing patient satisfaction surveys high in evaluations of the quality of a department's care. There have been concerns about the methodology, validity, and value of patient satisfaction surveys. Welch et al. (2010) acknowledge these concerns but argue that EDs should take these surveys seriously, work to improve scores, and use results to improve the quality of the department's service. They also assert that patient satisfaction enhances clinical effectiveness: satisfied patients are more likely to be compliant with instructions and thus respond better to treatment (pp. 6–7). A Duke University study echoed this point, finding that patient satisfaction scores related more closely to quality of medical care than clinical performance measures (Privett 2011).

Welch et al. note that patient satisfaction also correlates with staff satisfaction (2010, p. 7). Finally, they point to lower incidence of malpractice suits against providers in facilities with good patient satisfaction records.

Once an ED or hospital accepts the importance of patient satisfaction and the validity of satisfaction surveys, and determines it will work to improve patient satisfaction with its services, achieving this goal is not difficult. At BestPractices (BestPractices is an emergency physician and hospitalist staffing and management group) we refer to such improvement as an "open-book test."

Table 18.1 Survey items and their correlation to patient likelihood of recommending ED to others

Priority rank	Survey item
1	How well you were kept informed about delays
2	The degree to which the staff cared about you as a person
3	How well your pain was controlled
4	The nurses' concern to keep you informed about your treatment
5	The waiting time in the treatment area before you were seen by a doctor

2.1 What Patients Want

Quality improvement is an open-book test because the customers have told us what they want. All we have to do is pay attention. As an illustration, Press Ganey Associates has compiled a list of patients' top ten priorities when they go to an ED, based on patient satisfaction surveys. The Press Ganey 2010 Emergency Department Pulse Report (Press Ganey 2010) sets out this list:

1. How quickly their pain was treated?
2. How well informed they were about delays?
3. How much the staff cared about them as persons?
4. Their overall rating of the care they received.
5. How long they waited to see a doctor?
6. How well their pain was controlled?
7. How sensitive the staff was to their pain?
8. Their likelihood of recommending the facility to others.
9. How adequate was the information to their family and friends?
10. How courteous the staff was to their family and friends?

Just as a student can ace a test when knowing the questions in advance, health care team members can ensure priorities are met when designing their processes and carrying them out. The Priority Index provides more insight for EDs and hospitals in regard to what specific items make patients more likely to recommend an ED to others. Table 18.1 shows these items and their priority ranking.

A health care staff that values customer service and patient satisfaction highly focuses on meeting and then exceeding patients' expectations. "Patient expectations" may seem nebulous, but knowing what they are is not difficult: providers merely need to ask. Our physicians at BestPractices ask their patients, "What is the one thing we can do today to exceed your expectations?" Not surprisingly, their expectations may not always be what appropriate treatment requires. Here are some responses that health care staff might encounter:

- "My head hurts. I should have a CAT scan."
- "My chest hurts. I need an EKG."
- "My throat is sore. I need an antibiotic."

Table 18.2 Frequent complaints and feared causes by patients

Chief complaint	Fear
Chest pain	Heart attack
Numbness, weakness	Stroke
Headache	Brain tumor
Pediatric fever	Meningitis
Abdominal pain	Appendicitis
Sore throat	Strep infection
Twisted ankle	Ankle fracture

Thus, providers need training in negotiating those expectations with patients. Physicians and nurses should listen to patients and communicate with them patiently and clearly, making sure they understand why the provider is following a particular course of treatment and not following another. Providers should also keep patients informed throughout the course of treatment.

2.2 *Seeing from the Patient's Perspective*

Because they get immersed in technical details or encounter conditions every day, health care professionals may view patients' conditions as a problem to be solved or as routine. Emergency departments and health care institutions need to instill in their staff the necessity of seeing a health care experience from the patient's point of view, and thus anticipating what the patient is likely to think or fear. Anticipating in this way can lead to more sensitive interaction between provider and patient and better communication. Table 18.2 lists some typical effects and projected causes.

Anticipating in this fashion and understanding what patients are thinking and feeling helps bring power and control into the interaction on the patient's side. Patients often feel lost and helpless in chaotic emergency departments. Empowering them brings some stability into their experience; it is also the humane thing to do in providing care.

2.3 *Empowering Patients*

"Empowering patients" may sound grandiose, but it is simple. Essentially, empowering patients is communicating with them and providing them the opportunity to communicate back, whenever they need to. Having easy access to the call button and being shown how to use the telephone in a room, for example, enable the patient to get in touch with providers whenever that patient wants more information or simply reassurance that everything is on track. Giving patients treatment options involves them in the process and gives them a sense of control within the chaos around them.

Empowerment of patients in this way has implications beyond a particular visit. For example, providers can give patients their business cards and tell them, “You can come back at any time you are getting worse and can’t get in to see your physician or specialist. Ask for me by name.” This kind of action may help establish patient loyalty that has long-term effects in the form of both return visits and referrals to others the patient knows.

3 Raising the Scores: Working on the Upper End

Aiming at excellence in satisfying patients—or customers—is an effective intention, for more reasons than one. Our survey at BestPractices scores patient feedback from 1 to 5 (with 1 being the worst complaint and 5 being the highest level compliment). These scores are then compared among EDs by percentile. For instance, at the 99th percentile, 66 % of feedback scores a 5 and just 1 % of feedback scores a 1. EDs aim to raise their percentile across all five scores. Table 18.3 reveals an interesting dynamic. The difference between hospitals widely separated by percentile ranking is very slight at the lower scores. But even a small difference in the scores in the 4s and 5s makes a big difference in the percentile ranking. Moreover, the difference between facilities in the 99th and 64th percentiles on scores of 4s versus 5s, though not overwhelming in percentage, is dramatic in hospital rank. So working to move 25 % of those who perceive the hospital as “very good” to the category of perceiving it as “excellent” is a strategy that will pay huge dividends. The key driver in improving patient satisfactions scores is converting the “very good” assessments into “excellent,” not focusing on the lowest scores. For most health care practitioners this is a counter-intuitive and often unappreciated aspect of patient care surveys.

In working to move from being “very good” to “excellent” more often, the health care team can take many steps to improve customer service, including making adjustments in processes and design of the facility. Keeping it clean and comfortable, with temperature controls, new stretchers, pillows and blankets available, is one obvious action. Ensuring privacy in the registration area and separate areas for fast track, observation, and pediatric treatment is another. But the most important area to focus on has nothing to do with design. Among the top ten patient priorities are being kept informed and being treated courteously. These are issues of communication.

4 Communication: And How to Improve It

Effective communication leads to effective service. A Medicare survey found that the strongest factor in patient satisfaction is good communication with nurses (Privett 2011). Ineffective communication, on the other hand, leads to problems.

Table 18.3 Sample distribution of inpatient satisfaction scores by percentile ranking of hospital: the impact of moving 4s to 5s

Percentile rank of hospital	1s	2s	3s	4s	5s
99th	1 %	2 %	7 %	24 %	66 %
64th	1 %	2 %	8 %	34 %	55 %
35th	2 %	2 %	8 %	36 %	52 %

Source: Press Ganey and Associates 2007

At the heart of human friction is almost always a lack of effective communication, or a breakdown in it. Communication is also involved in risk management; effective communication reduces the risk of malpractice. One study of patients who had sued their health care providers found that 71 % of patients cited a poor relationship as a main reason; 32 % felt deserted; 29 % felt devalued; 26 % thought information was delivered poorly; and 13 % perceived a lack of understanding by the provider (Beckman et al. 1994).

Effective communication in the ED and elsewhere begins with understanding what patients want in general, particularly in being kept informed regularly on their own treatment and on reasons for delays. Listening to patients' specific concerns continues effective communication in each case. Beyond that beginning, a very good approach is the use of scripts. Press Ganey has found this to be one of the most effective ways to raise patient satisfaction scores. Scripts are easy to learn and can be implemented with little effort and cost.

Scripts can be geared toward the specific questions on patient surveys that supply data for the Priority Index. For example, for a question on the physician's concern for the patient's comfort, here are a few short scripts for doctors:

- "I want to make sure you are as comfortable as possible during your stay in our department."
- "I want to be sure you have everything you need while you are here. The staffs here (and I) are always concerned about our patients' comfort."
- "I realize there will be a lot of down time waiting during your stay. I want things to be as comfortable as they can for you (and your family and friends) while you wait."

Here are similar scripts nurses can use in regard to concern for patients:

- "I want to try to make you as comfortable as possible during your stay here. Please call me if you need anything or if you just get scared and need to talk."
- "This place can be intimidating . . . but now you know me and Doctor X. We will make sure you get the care you need. We all work together here, so if you can't find us right away, just ask someone to get us."
- "I want to make sure you get the attention you need. This place can get crazy sometimes, but you are very important to us. If you feel forgotten somehow, please come get me. We don't want anyone to ever feel that way."
- "Do you have any questions? I have plenty of time."
- "How can I help you?"

Scripting can cover a wide variety of situations in the ED and the hospital, at any point in their processes. Here, for instance, is how a script applies to preparing to discharge a patient:

- “Now that we know you can go home I just want you to be sure to understand what you can do at home to continue to improve. Let’s go over the treatment plan for home. I especially want you to know when to return and what to look out for with this particular illness.”

“Scripting” may sound elaborate and perhaps intimidating, but as these examples show, scripts are short, ordinary statements that make obvious sense in interacting with patients. And they also make sense in helping raise patient satisfaction scores.

5 The Fundamental Motivation Behind Improving Patient Satisfaction

People in health care should ask: why work to make patients more satisfied? The reason seems obvious: higher scores give the facility a better reputation and please the board. Higher scores also imply better treatment and safer care. These, of course, are all true. But many health care professionals fail to think of another reason, one that is in fact the most effective motivator of actions to improve patient satisfaction:

Greater patient satisfaction makes the jobs of health care staff members easier.

Thom Mayer argues that institutions that tell health care providers that (as an additional part of their already difficult jobs) they need to get patient satisfaction scores higher, “should not be surprised when [their] staff members not only are not on board and not truly invested but in fact revolt. To many care providers customer service is just one more thing they have to do in the middle of a busy and increasingly busy job” (Mayer 2010, p. 66). Presenting an initiative to improve customer service in order to increase scores is thus unlikely to persuade staff to change the way they operate.

Intrinsic motivation, on the other hand, is a powerful force, and if staff members understand that changing the way they work makes their job easier, they are much more likely to embrace the change. Effecting changes in service behavior within the health care system is extraordinarily difficult unless programs aimed at changing behavior help physicians and nurses do their job. Helping them do it better is the best way to get them to change.

Self-interest, as noted, is intrinsic motivation; directors will invest much less effort and time in trying to achieve change when their doctors and nurses are seeking the change on their own. So convincing them that changes will make their lives easier should be the number one goal behind change initiatives. This principle also makes recognizing effective customer service easier. As Mayer puts it, “if a service excellence program does not make your job easier, it is not really customer service but something masquerading as service” (2010, p. 69).

Convincing staff that change will make their jobs easier is half the work of achieving excellence. The other half involves teams, specifically two different teams.

6 Two Kinds of Teams

A competitive sport may take two teams, but medicine is not a competitive sport, and a health care facility wants only one team: what we can call the “A team.” The other kind is the “B team.” This distinction is not between a first string of more talented clinical performers and a second string of not as gifted but still good performers. Members are on these teams because of their behavior and attitude.

Having A team members is essential for putting into place the intrinsically motivated changes to improve customer service. Being on a shift with them brings a natural emphasis on excellent service. Having B team members, however, puts a damper on service excellence. An important point about this kind of difference is this:

It only takes one B team member to wreck a shift for everyone on it.

A team members and B team members are easy to spot from the character traits listed below. Seeing these traits makes it easy to grasp why anyone would want to serve with A team members and not with B team members.

A team characteristics	B team characteristics
Positive	Negative
Proactive	Reactive
Confident	Confused
Compassionate	
Competent	Lazy
Effective communicator	Poor communicator
Team player	Frequently late
Trustworthy	
Willing and able to do whatever job requires	Constantly complains
Possesses sense of humor	Has victim mentality

Mayer notes that, while not typically described this way, A team characteristics are actually good customer service traits and B team characteristics are customer disservice traits—which in fact make everyone’s job harder (2010, p. 69). So embedding A team characteristics in a staff will, by definition, improve customer service and thus patient satisfaction will increase.

6.1 Forming and Molding an A Team

From an overall perspective, forming an A team begins with hiring—and ends in the opposite for B team members who cannot or will not drop the B team traits and pick up A team traits. B team members must ultimately be replaced if that is the only solution. One should not assume, however, that B team performers can never transform into A team members—many can and will with proper encouragement. Often they are unaware of the effect of their behavior on other staff members on their shift and on patients. They may have never been directly and honestly informed about examples of such behavior and pointed to more effective behavior.

On an ongoing basis, encouraging, communicating, and coaching A team behavior is essential to inculcate a culture of excellence that will produce first-class customer service, and in turn effective and safe clinical care. Coaching means pointing out characteristics that are not contributing to maintenance of an A team and suggesting characteristics that would. In a sense, this is reactive. But staff directors can be proactive by communicating examples of A team behavior that others can emulate—publically acknowledging and celebrating behavior that does establish an A team, even posting compliment letters or e-mails.

Delineating A and B team characteristics is not an absolute practice. A physician who is not a good communicator and is often late may still be clinically competent and compassionate. In such a case, while the director needs to coach the doctor on the two former characteristics, that director can publically give credit to the doctor for the latter two. Pairing a B team performer with a mentor from the A team can also be an effective way to show rather than tell the former how to transform into an A team member. Another technique is to pair two complementary performers, one of whom is good at certain desired behavior and the other who is good at a different effective behavior.

Patient ratings can be quite useful in coaching for A team behavior. Analyzing complaints and compliments and distilling the results into patterns provides concrete information that helps B team members see how they are perceived and what they must do to receive better responses. Patient satisfaction scores are a valuable tool in molding the A team. They can be used to reinforce A team behavior and eliminate B team behavior, while holding staff accountable.

6.2 Educating the A Team

Coaching individuals is not the only requirement for a facility that wants to produce an A team. Everyone on the staff needs to receive continuing education in A team behavior. In addition to highlighting the basic traits of A team performers, such sessions should focus on the importance of customer service diagnosis as well as clinical diagnosis, and how to conduct a service diagnosis; what effective customer service entails, including the use of scripts; and how to solicit patients' expectations, negotiate with them, and exceed expectations.

From experience, here are the top ten A team behaviors in the specific context of a patient visit, from start to finish. Education should emphasize and train staff on these ten behaviors:

1. Sitting down with the patient.
2. Reading nursing, triage, and emergency medical service notes.
3. Smiling genuinely.
4. Making respectful, friendly, and professional introductions.
5. Setting expectations for the visit.
6. Exceeding those expectations.
7. Making physical contact and taking measures for patient comfort.
8. Using scripts intelligently.
9. Providing updates and relevant information.
10. Offering a summary to reinforce the treatment plan and follow-up and making a graceful exit at discharge.

Education for A team behavior should be formalized, and not ad hoc. A study of a formal education program in which all ED staff who had any patient contact attended 4 or 8 h of customer service training found that patient complaints decreased by more than 70 %, from 2.6 per 1,000 ED visits to 0.6 per 1,000, and patient compliments increased by more than 100 %, from 1.1 per 1,000 visits to 2.3 (Mayer et al. 1998). The study also found that the most dramatic improvements in patient satisfaction scores were in assessments of physicians' and nurses' skills, likelihood of patients to return to that ED, and overall satisfaction.

7 Recovery

Exhibiting A team behavior from the outset is the most effective way to improve patient satisfaction. Anticipating patients' needs, expectations, and fears and being proactive in regard to them; exhibiting all the good customer service behaviors discussed here; and cultivating an atmosphere that produces an A team will lead to high-quality patient service. Sometimes, however, mistakes in service will occur. When they do, acting quickly and fully to rectify the problem is important.

The concept behind this sort of action is "recovery"—in other words, salvaging the patient's visit so that it does not become a failure. The basis, once again, is communication, which begins with listening carefully to the patient and his or her view of what has happened, or not happened. The other aspects of recovery also involve communication to a large degree. The first step is an apology. Then, the provider should ask, "what can I do to make this right?" Next, of course, is fixing the problem, and making sure to follow through and follow up, monitoring the process. Most people have likely experienced a visit to a restaurant where

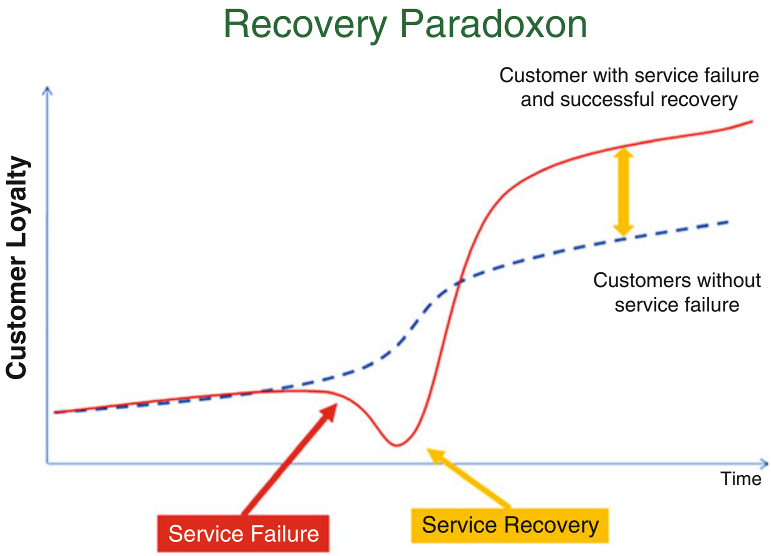


Fig. 18.1 The effect of recovery in customer service

the kitchen loses the order or the waiter "forgets" about the table; often, the management will offer a free dessert or something similar to make up for the error. Health care facilities, where feasible, can offer something extra in service to a patient who has experienced a lapse in service.

Recovery can pay off. Figure 18.1 illustrates the "recovery paradoxon," showing that recovery can even build a stronger level of customer loyalty than customers who have not experienced the need for recovery. Naturally, not having to recover is preferable, but the ability to recover gracefully and effectively is a strong complement to high-quality customer service in the first place.

8 Patient Satisfaction and Patient Flow

Working to improve patient satisfaction contains one more component. Improving patient flow in general in the ED and within the hospital will raise patient satisfaction scores. The truth of this principle is evident in two of the top ten desires of patients in the Priority Index: how quickly their pain was treated was number one, and how long they waited to see the doctor ranked number five. Often the greatest factor in patient satisfaction, and one that is measurable, is the "door-to-doctor" time—how long it takes from when the patient enters the ED to when that patient comes face to face with the treating physician (Boudreaux et al. 2004).

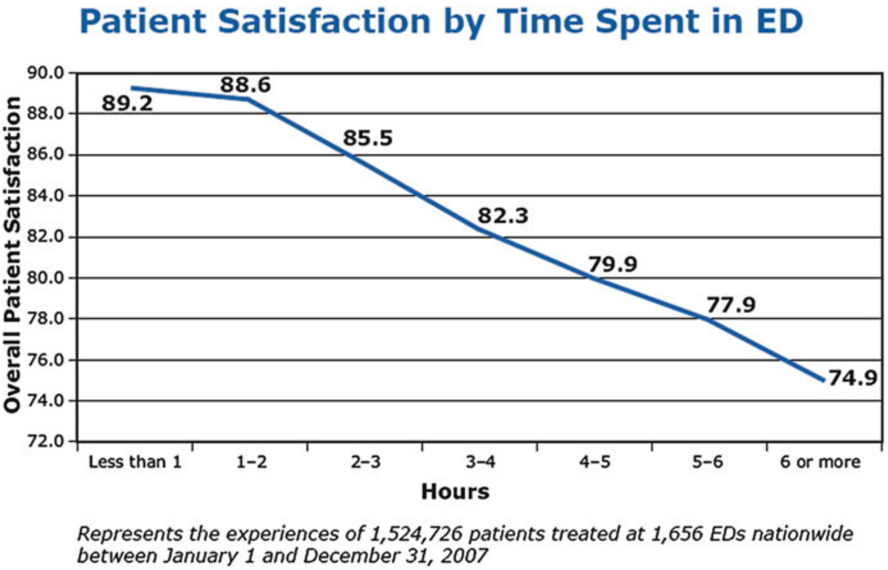


Fig. 18.2 The correlation between patient satisfaction and length of time in ED

EDs need to concentrate on refining and streamlining all aspects of patient flow through the ED. The longer the patients are in the department, the less satisfied they are likely to be. Figure 18.2 shows graphically how time spent in the ED correlates with patient satisfaction. If long waits are the norm in a facility, then patient flow is likely to be poor, and quality of care and safety correlate with effective patient flow.

Streamlining admission and discharge processes, for example by offering an appointment-based process and a centralized bed authority, can reduce crowded conditions and waits. Between admission and discharge, one effective method of improving flow and patient satisfaction at the same time is patient rounding. Patient surveys have shown that patients would like to be contacted every 20–30 min, so an ED staff that makes rounds every half hour is fulfilling patient expectations, as well as monitoring conditions and therefore smoothing flow. A study on rounding in ED reception and treatment areas found that using regular rounding protocols significantly increased patient satisfaction ratings for overall care and pain management (Meade et al. 2010). The study further found that rounding reduced:

- Patients who left against medical advice by 22.6 %.
- Patients who left without being seen by 23.4 %.
- Call light use by 34.7 %.
- Approaches to the nurses' station by 39.5 %.
- Falls by 58.8 %.

9 The Psychology of Waiting

Despite a department's or hospital's best efforts to smooth flow, sometimes it will become crowded. In crowded conditions, health care staff need to apply the lessons learned from study of the psychology of waiting to make delays less frustrating to patients (and staff). There is a psychology involved in humans' waiting, with specific principles that researchers in the field have identified, principles that lead to predictable behavior. Many businesses take advantage of these principles. As crowding continues and intensifies, health care systems need to intensify as well.

9.1 Eight Principles of the Psychology of Waiting

The classic study is Maister (1985). In that article, David Maister describes eight principles of the psychology of waiting, based on how people react mentally to having to wait:

1. Unoccupied time feels longer than occupied time.
2. Pre-process waits feel longer than in-process waits.
3. Anxiety makes waits seem longer.
4. Uncertain waits seem longer than known, finite waits.
5. Unexplained waits seem longer than explained waits.
6. Unfair waits feel longer than equitable waits.
7. The more valuable the service, the longer the customer will wait.
8. Solo waits feel longer than group waits.

It is not hard to find businesses that take advantage of these principles. Any good restaurant, for instance, will apply the first principle, aware that starting any kind of activity for the customer related to service will engage that customer psychologically, creating the sense that service has begun. So a waiter, or even an apprentice waiter, will come as soon as patrons are seated, handing out menus, indicating the waiter will be there soon to take their order, and asking if they'd like to order drinks.

Some well-known large national businesses are particularly good at applying these principles. One that excels is the Disney Company. When people wait in line for a ride at Disney World, videos or characters in costume may entertain them. People winding their way on the Star Tours ride, for instance, make their way through a futuristic android warehouse, complete with elaborate sets and props that keep customers fascinated as they move or stand still.

An ED or hospital can apply this same principle in various ways, many of them simple. Providing current issues of a variety of magazines as well as information on health that fits their conditions fills patients' time in waiting areas. Having televisions in waiting areas is another example—after all, people go to sports bars by choice to watch games. Having patients fill out registration forms as soon as they enter the waiting area quickly turns the wait into the start of a process, and makes the patient believe service has begun.

A couple of these eight principles stand out as particularly relevant to health care systems. "Anxiety makes waits seem longer" is one: a patient who comes to the ED or hospital is often worried and already having a bad day. Making that patient wait is unlikely to improve his or her mood, or patience. Letting patients know why they are waiting, what they are waiting for, and how long the wait will likely last helps assuage anxiety and responds to uncertainty, another clearly relevant principle in the ED and hospital. Yet this kind of communication is not difficult. Continuing to contact patients and keep them up to date, much as a waiter might check back to see if diners need anything, keeps the anxiety and uncertainty of patients from creeping back into their minds. ED patients, when surveyed, have expressed a preference for being contacted every 20–30 min—a fact ED staffs need to keep in mind, particularly when crowding increases.

These principles of the psychology of waiting are based on one underlying point, as the expression of them implies: waits "feel" and "seem." How long patients actually spend in the ED or hospital is less important than how long they perceive they wait. This fact has important implications for staff actions that seem inconsequential but which affect perception. For example, whether a physician sits or stands while talking with the patient molds that patient's perception. A survey found that patients overestimated the time the doctor spent in an interaction by an average of 1.3 min where both patient and doctor were seated (mean encounter length was 8.6 min). In interactions where the physician stood, patients underestimated the time spent by an average 0.6 min (Johnson et al. 2008).

In regard to communicating regularly with patients, one study found patients contacted every 15 min perceived an average length of stay in the ED of 92.6 min, compared with 105.5 min in a control group (Tran et al. 2002). This same study found that significantly more of the patients receiving regular communication rated the ED physician "excellent" or "very good" than those in the control group. So time is not the only perception affected by regular communication.

Giving patients specific figures when they ask how long something will likely take—receiving test results, for example—also improves perceptions of waits. A specific answer, such as 45 minutes, reducing uncertainty for the patient, is better than a vague and generalized response, such as "it shouldn't take long" or "we've had a lot of orders for tests today, so it may take a while." In addition, adding some time to the anticipated figure makes the patient think, if the result comes back sooner than that added interval, that the wait was shorter than it actually was. Again, Disney understands this principle; its rides often post a wait time of 40 min when the actual time, as the company well knows from experience, will be 30 min.

Something important in the principles as they relate to the ED is perceived fairness. Perceived unfairness is familiar in daily life; if someone enters a waiting area in a business or government agency and a later arriving person receives service earlier, the first person perceives the situation as unfair. The same perception arises in an ED or hospital when patients waiting in one area see those in another being treated before them. So keeping waiting areas for different functions (e.g., general waiting and fast track waiting) separated as much as possible helps mitigate this sense of unfairness.

Similarly, perception of value colors how long a wait seems. Once again, a fine restaurant is a familiar example: diners are much more willing to wait longer—often much longer—for a table at a restaurant widely perceived to be excellent than they would at one considered average. The same principle applies to health care facilities. When a facility is perceived as excellent, better than others in the area, patients are likely to tolerate waits they would not in one considered average. Creating a perception of value is clearly a larger task than putting magazines in a waiting room. But anything a health care system can do in general to create that perception will pay off in patient willingness to wait.

In the immediate, short-term context, two imperatives will help create a perception of value in patients. First, maximize benefits for the patients and their family and friends. Second, minimize burdens for the patients and their family and friends. Working toward those goals will give patients a sense of value in their treatment on this specific occasion, regardless of the wider context of reputation and perceived value of the facility.

Finally, that solo waits feel longer than group waits is probably obvious to most people. Not only allowing but encouraging friends and family to wait with patients throughout their ED or hospital experience, as much as possible, including when they are in beds in rooms, will make waits seem shorter.

9.2 Eight More Principles: Designing Lines

More recently, Donald Norman (2009) has focused on the psychology from a different perspective and proposed a new set of principles, also an octet. Some are similar to Maister's, but others focus on different aspects of the psychology. Norman's principles, for the most part, are more oriented toward action than observation:

1. Emotions dominate.
2. Eliminate confusion: provide a conceptual model, feedback, and explanation.
3. The wait must be appropriate.
4. Set expectations, then meet or exceed them.
5. Keep people occupied: filled time passes more quickly than unfilled time.
6. Be fair.
7. Start strong, end strong.
8. The memory of an event is more important than the experience.

Norman's last point underscores again the importance of perception. What patients actually went through during their stay in the ED and hospital is not what sticks in their minds as much as what they perceive—and memory is strongly colored by perception. His fourth principle carries the lesson about setting a specific figure into a broader context, the entire ED experience. If the staff can clearly

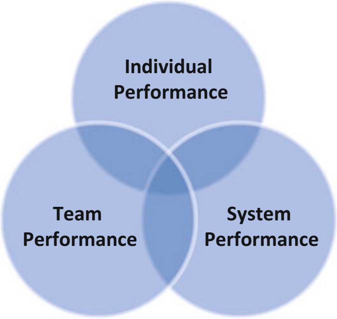


Fig. 18.3 Levels of ED performance in relation to patient satisfaction

convey how the patient will be treated, and then prove true to their word, or even better, patients' memories of their stay in the department will be positive.

All the other tools of enhancing patient flow, as described in the chapters in this book and in other books as well, will lead to increased patient satisfaction as they lead to smoother flow of patients through the ED. (For a detailed examination, see Mayer and Jensen 2009.) Refining processes is important. Creating a team that works well together by performing at an A team level is important. And focusing on improving individual performance through scripts, coaching, education, and the other means discussed here is important.

Figure 18.3 encapsulates how all these elements come together to create an optimal zone of patient service. Entering that zone is reaching the sweet spot that will boost both patient flow and patient satisfaction.

References

Beckman, H. B., Markakis, K. M., Suchman, A. L., & Frankel, R. M. (1994). The doctor-patient relationship and malpractice. Lessons from plaintiff depositions. *Archives of Internal Medicine, 154*, 1365–1370.

Boudreaux, E. D., d'Autremont, S., Wood, K., Jones, G. et al. (2004). Predictors of emergency department patient satisfaction: Stability over 17 months. *Academic Emergency Medicine, 11*, 51–58.

Johnson, R. L., Sadosty, A. T., Weaver, A. L., & Goyal, D. G. (2008). To sit or not to sit? *Annals of Emergency Medicine, 51*(2), 188–193.

Maister, D. (1985). The psychology of waiting lines. In J. A. Czepiel, M. R. Solomon, & C. F. Suprenant (Eds.), *The service encounter: Managing employee/customer interaction in service business* (pp. 113–123). Lexington, MA: Lexington Books.

Mayer, T. A., Cates, R. J., Mastorovich, M. J., & Royalty, D. L. (1998). Emergency department patient satisfaction: Customer service improves patient satisfaction and ratings of physician and nurse skill. *Journal of Healthcare Management, 43*(5), 427–440.

Mayer, T. A. (2010). Leadership for great customer service. *Healthcare Executive, 25*(3), 66–69.

Mayer, T., & Jensen, K. (2009). *Hardwiring flow: Systems and processes for seamless patient care*. Gulf Breeze, FL: Fire Starter.

Mayer, T., & Cates, R. J. (2004). *Leadership for great customer service: Satisfied patients, satisfied employees*. Chicago, IL: Health Administration Press.

Meade, C. M., Kennedy, J., & Kaplan, J. (2010). The effects of emergency department staff rounding on patient safety and satisfaction. *The Journal of Emergency Medicine, 38*(5), 666–675.

Norman, D. A. (2009). Designing waits that work. *MIT Sloan Management Review, 50*(4), 23–28.

Press-Ganey, March 2007. Taking 4s to 5s, as cited by Thom Mayer, MD at the ACEP Emergency Department Director's Academy, Phase ll, May 2011.

Press Ganey. (2010). *National emergency department priority index. Emergency Department Pulse Report*. South Bend, IN: Press Ganey Associates, Inc.

Privett, C. (2011). *Satisfied patients are the best measure of quality, Duke study finds*. Fuqua.duke.edu/news.

Tran, T. P., Schutte, W. P., Muelleman, R. L., & Wadman, M. C. (2002). Provision of clinically based information improves patients' perceived length of stay and satisfaction with EP. *The American Journal of Emergency Medicine, 20*(6), 506–509.

Welch, S. J., Hellstern, R. A., Jensen, K., Lyman, J. L., Mayer, T., Pilgrim, R., et al. (2010). Can't get no satisfaction? The real truth behind patient satisfaction surveys. *Emergency Medicine News, 32*(6–7), 26.

Chapter 19
Mayo Post Acute Care Program and Care Continuum

Mark Lindsay

Abstract The Mayo Post Acute Care Program highlights the opportunity for healthcare organizations to establish high quality post acute care pathways to match the demands in acute care hospitals. Key elements of the program include teamwork, staff empowerment, collaboration across care settings, and development of new pathways outside the hospital. The program eliminates the notion of discharge, and promotes high-quality programs, transitional and ventilator care as patients are most vulnerable as they transition from one setting to the next. Team centric processes, establishing key metrics, reliably implementing evidence based care are all important strategies for optimizing care across the continuum. There are also significant opportunities to apply improvement methods to maximize compliance of evidence based care in the ambulatory care setting as well which could have significant positive impact on reducing flow into the hospital.

Keywords Care continuum • Transitional care units • Ventilator units • Cost-effective • Critical access hospitals

1 Introduction

Much has been accomplished since the Institute of Medicine (IOM) report highlighting the concerns with our healthcare system related to errors and preventable deaths (IOM 1999, 2001). The Institute for Healthcare Improvement along with countless other healthcare organizations and teams has dedicated tremendous resources to quality improvement and patient safety. There continue to remain significant opportunities to positively impact healthcare quality and costs. Healthcare costs continue to rise, exceeding $2.4 trillion in 2009 with hospital

M. Lindsay (✉)
Mayo Clinic Health System, 1221 Whipple Street, Eau Claire, WI 54702, USA
e-mail: lindsay.mark@mayo.edu

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_19, © Springer Science+Business Media New York 2013

care comprising more than $750 billion, or 31 % of the total (National Healthcare Expenditures). Hospital care has accounted for the largest component of healthcare costs. This is significantly more than other industrialized nations. Healthcare improvement efforts are targeting more efficient and effective pathways for the most costly and vulnerable patient populations. Other efforts are focusing on improving hospital processes with greater reliability and emphasis on care coordination. Optimizing chronic disease prevention efforts with greater patient engagement, population management has the potential to avoid unnecessary costly hospitalizations on the front end.

1.1 Lessons Learned from Our Most Vulnerable and Costly Patient Populations: Prolonged Mechanical Ventilation

Patients requiring prolonged mechanical ventilation represent an extremely vulnerable patient population with excessive costs and high risk for poor outcomes. Ankrom and Barofsky reported outcomes from a nursing home ventilator unit with 15 % of patients liberated from the ventilator and 19 % were alive at 1 year (Ankrom and Barofsky 1998). Unroe and Kahn reported that more than two thirds of ventilator patients in their cohort had a hospital readmission and patients surviving to discharge had a median of four transitions of care following acute care hospitalization. Mean cost per patient was $306,135 (Unroe and Kahn 2010). There is an anticipated doubling of this vulnerable population by 2020 with anticipated costs of around $60 billion (Zilberberg et al. 2008). Inadequate post acute care options for these patients results in excessive costs for acute care hospitals. There are important opportunities to optimize the pathways for these highest risk patients. This chapter will attempt to highlight the challenges in caring for this unique patient population.

1.2 Case for Transitional Care

The literature would point out the needs for better alternatives for post acute care. Patients admitted to skilled nursing facilities are at high risk for urinary and lower respiratory tract infections, decubitus ulcers, etc. The high prevalence of these infections and complications may be related to high staff turnover, lack of attention to infection control practices as well as other factors (Garibaldi et al. 1981). Nursing staff turnover may be as high as 100 % (Decker and Gruhn 2003). Clinical outcomes that reflect the transition of high-risk patients from the acute care setting to skilled nursing facilities demonstrate results that are less than desirable. Cook and Martin found that 45 % of patients discharged from a surgical Intensive Care Unit (ICU) to an extended care facility died within 2 years (Cook and Martin 1999).

Carey and Parker reported that over 13 % of CABG (coronary artery bypass graft) patients in his report that were transferred to other healthcare facilities died (Carey and Parker 2003).

Elderly patients are a particularly vulnerable patient population. They account for a significant proportion of hospital admissions and prolonged hospital stays. Readmission rates may be as high as 66 % within 6 months post discharge. Hospital admissions have inherent risk for elderly patients, which include delirium, nosocomial infections, malnutrition, pressure ulcers, and adverse drug events (Callahan and Thomas 2002). Hospital readmissions account for billions of dollars and represent one of the targeted areas for cost containment by CMS (Affordable Care Act Update 2010). In addition, there is clear evidence that Transitional Care and strategies that optimize care coordination for these high risk patients can improve outcomes and reduce readmissions (Jackson and Trygstad 2013; Coleman and Smith 2004).

1.3 Chronic Disease and Primary Care Team Opportunity

Approximately one in two adults lives with chronic disease, which affects more than 130 million. Heart disease and stroke account for more than 30 % of deaths in the USA each year. The majority of healthcare costs and deaths each year are due to chronic conditions and disease (CDC). Prevalence of obesity and diabetes is growing at an alarming rate. Hospital costs associated with diabetes alone is staggering, accounting for $58 billion in 2007, 50 % of total direct medical expenditures for diabetes (Wang and Imai 2009). Cardiovascular disease and hypertension account for the vast majority of deaths in patients with diabetes. Unfortunately the percentage of patients that are at blood pressure goal is less than 30 % (Suh and Kim 2009).

We will highlight one example of a multi-site primary care Mayo Clinic Quality initiative focusing on improving hypertension in diabetes utilizing a care bundle, engaging patients in their improvement, and applying a team-based order set (Lindsay and Hovan 2013). Hypertension is the leading preventable cause of mortality in the world (Ezzati and Lopez 2002). One of healthcare's greatest opportunities to positively impact hospital flow and reduce costs in a meaningful way will be continuing the accelerated emphasis and resources on chronic disease prevention efforts, tapping into the expertise and dedication of primary care teams, and proactively avoiding unnecessary costly hospital admissions and readmissions on the front end.

1.4 Care Continuum

As population ages and more people suffer from chronic disease and as the projected number of primary care physicians do not match the anticipated demands, it will be more critical that primary care providers and care teams implement work flows that effectively and reliably apply evidence based practices to care for patients with chronic disease. Not all hospitalizations or readmissions can be avoided. The ability to manage patients across the continuum, coordinate their care, identify and mitigate risk factors and provide the "right care in the right place at the right time" will provide the highest value.

Hospitals continue to look for ways to reduce excessive acute care hospitals costs. A small percent of patients account for a high percent of bed days and resources. The ability to identify these outliers and develop pathways to better optimize resources, reduce waste, and improve outcomes is a priority for hospital leaders and administrators. One of the greatest opportunities for acute care hospitals is the establishment of post acute care pathways to match the acute care demands. This can be accomplished by partnering with post acute care partners. Medicare bed days beyond the mean geometric length of stay tied to diagnoses and DRGs identifies the business case and opportunities to establish post acute care pathways to match the demands in acute care hospitals.

The Medicare DRG reimbursement method provides a fixed payment for most diagnoses regardless of the length of stay (LOS) when excluding outlier payments. Certain transfer DRGs will result in a per diem rate when transferred to a post acute care provider within a designated length of stay (Law Watch). The ability to reduce length of stay for Medicare DRG patients, including Transfer DRG patients, beyond the designated time frame will result in reduced variable costs and will free up hospital beds.

In one report from a medical surgical ICU, patients who stayed for 14 or more days accounted for 7.3 % of the admissions but accounted for more than 40 % of the total patient days (Weissman 2005). Patients on mechanical ventilation account for less than 10 % of admissions to intensive care units but they account for over 30 % of bed days and resources, accounting for a very strong business and quality case for establishing high quality post acute care pathways (Lindsay and Bijwadia 2004).

2 Wisconsin Ventilator Program

Chronic ventilator dependent units have been developed to optimize the care for chronic ventilator patients (Gracey et al. 2000; Scheinhorn et al. 1997; Bagley and Cooney 1997). Many of these reports have demonstrated reduced costs as well. Chronic ventilator dependent units, largely representing long-term acute care (LTAC) facilities, provide an important mechanism to improve flow through acute care facilities, thereby improving access for new admissions and associated revenues. One of the greatest concerns is the lack of pathways to care for chronic ventilator patients in the home and skilled nursing facilities. This creates a bottleneck and significant financial deterrent for LTAC facilities to accept patients, which may have poor weaning potential, resulting in difficulty in placing these patients. Medicaid ventilator patients are a very difficult subset of patients to place due to financial reimbursement.

2.1 Benefits in Establishing a High Quality Post Acute Care Ventilator Unit

Ventilator patients account for an important subset of hospitalized patients that contribute to very long lengths of stay. Long Term Acute Care Hospitals (LTACs) provide an important downstream discharge option for Medicare patients but not all regions have adequate access to LTAC discharge options. In addition, some Medicaid ventilator patients may be very difficult to place. The most cost-effective approach to caring for ventilator patients is successfully liberating them from the ventilator and discharging them home.

2.2 Key Features

A ventilator program was established in a skilled nursing facility in Chippewa Falls, Wisconsin. Key features to the program included emphasis on socialization, such as encouraging patients to get out of their room with portable ventilators, go outside on a nice day, or participate in activities with other residents in the common room (Lindsay and Bijwadia 2004).

Empowerment of staff was also strongly emphasized with implementation of respiratory therapy and nurse directed weaning protocols, which have been demonstrated in the literature to reduce weaning time (Marelich and Murin 2000). In-line use of talking valves was also encouraged, which promoted communication and reduced anxiety.

Standardized of equipment, training and education, and a multidisciplinary care team approach were also key features to the model. Bedside rounds with patient and family at a set time occurred weekly. This intervention promoted a leveling of the hierarchy, provided an opportunity to create a care plan for the patient, and provided an opportunity to answer patient, family, and care team concerns. Multidisciplinary rounds have been shown to improve communication and satisfaction, reduce hospital stay as well as other benefits (Dodek and Raboud 2003; Leape and Cullen 1999; Young et al. 1998).

Key features in the success of Nursing Home Ventilator program, which has cared for almost 70,000 ventilator patient days from 1997 to 2012, included an

Table 19.1 Comparison results with hospital based ventilator units

Vent unit	% Weaned	% Neuro
Wi Program[a]	67	27
Scheinhorn	56	7.8
Mayo (Gracey)	60	NR
Bagley 97	38	19

[a]Lindsay JCJQS 04

emphasis on training and education, empowerment of staff, multidisciplinary approach to care, standardization, protocol development, and emphasis on socialization. We believe social engagement had a significant impact on the outcomes for the ventilator patients (Lindsay and Bijwadia 2004). Most of the patients utilized their talking valve the majority of waking hours. Patients ate their meals around a common table, engaging in conversation, and participating in off-site outings. The majority of patient referrals had been residing in an ICU prior to transfer. Depression, sleep wake disruption, and loss of hope were common symptoms for new referrals. We always encouraged hospital referrals to visit the unit prior to admission. We believe the socialization significantly impacts outcomes, provides hope for patients and families, and increases acceptance of a nursing home vent unit transfer. Socialization may have been the single most important factor in the success of the ventilator model.

The pulmonary physician led weekly multidisciplinary bedside rounds focused on establishing the plan for the week. A nurse practitioner was dedicated to the unit and was available when patient condition changed. Respiratory therapy and nurse led weaning reflect the emphasis on empowerment of staff and best practices (Kollef and Shapiro 1997).

2.3 Results

Table 19.1 demonstrates outcomes of the Wisconsin ventilator unit relative to other reports in the literature. Sixty-seven percent of the patients admitted to the unit were liberated from mechanical ventilation. Twenty-eight percent of patients that presented to the Nursing Home Ventilator Unit had a neuromuscular diagnosis, which is significantly higher than other reports (Lindsay and Bijwadia 2004). This subgroup of patients is the least likely to be liberated from mechanical ventilation. Through the use of an innovative protocol, many of these patients were converted to noninvasive positive pressure ventilation, allowing them to be discharged home. This results in significant patient satisfaction, cost savings and maintains flow through the ventilator unit and acute care hospitals.

Table 19.2 compares skilled nursing facility ventilator units. The establishment of the nursing home ventilator unit has resulted in a cost savings to more than 20 referring facilities of approximately $18.5 million. More than 50 % of the admissions to the unit were from Mayo Clinic and Mayo Clinic Health System.

Table 19.2 Comparison results with nursing home based ventilator unit

NH vent unit	% Weaned	% Alive 1 year
Johns Hopkins Geriatric Center	15	19
Wi Program[a]	67	70

[a] Lindsay JCJQS 04

Table 19.3 Culture of safety survey skilled nursing home ventilator program

	Chippewa Falls	AHRQ data
AHRQ SNF culture of safety survey November 2011	WI SNF (%)	Base (%)
Overall perceptions of resident safety	95	86
Feedback and communication about incidents	90	84
Supervisor expectations and actions promoting resident safety	90	79
Compliance with procedures	74	64
Teamwork	72	64

$n = 56$, 86 % survey response return rate

These cost savings do not take into account the new revenues generated from new admissions as a result of opening up thousands of bed days at the referring facilities.

The differences noted in the two facilities in Table 19.2 may be attributed to a number of factors that include effective implementation of multidisciplinary rounds, respiratory therapy and nurse directed weaning, empowerment of staff (including certified nurse assistants), training and education of the staff, 24 h/7 days per week pulmonary physician staff coverage, nurse practitioner coverage, and emphasis on patient socialization. We believe the quality of the program significantly impacted the financial success of the unit including the volume of referrals. The unit grew from a census of five or six patients in 1997 to a 24 bed unit over the 15 year period.

Table 19.3 reveals a *culture of safety* survey conducted in November, 2011 at Chippewa Falls SNF Ventilator Unit. The culture of safety survey demonstrated an overall perception of safety of 95 %. Teamwork scores also scored positively. We believe that the bedside rounds, 24/7 pulmonary physician support, team communication, nurse practitioner support, and respiratory therapy leadership were significant contributors to the overall culture and teamwork in the unit.

3 Mayo Transitional Care Began as a Pilot Project in Bloomer, Wisconsin

We had already demonstrated success with the ventilator program, reducing the length of stay for those particular DRGs significantly. Back in 2000, we named our utilization management team, "Continuum of Care Committee" emphasizing our efforts to establish post acute care pathways as one our key strategies. In response the success of the Wisconsin Ventilator Program, a proposal was developed to establish a Transitional Care Unit in Bloomer, Wisconsin, utilizing the swing beds in the critical access hospital. The 1997 Budget Act gave rural hospitals with a critical access hospital designation the ability to receive cost based reimbursement for swing bed admissions to their facility (Greene 2002). Transitional care units in rural hospitals can provide higher staffing ratios than skilled nursing facilities, allowing the rural facility to care for more complex patients. Prior to mid 2001, Bloomer Hospital had ten consecutive quarters with net negative Net Operating Income (NOI).

3.1 Key Elements of Transitional Care Model

The Transitional Care Model emphasized socialization, music therapy, patient-centered and clinical outcomes, and the ability to be cared for close to home. Bedside rounds with patients, family and care team were essential elements to the model. Key stakeholders from the acute and rural hospital met regularly to establish pathways and new programs, develop outcome measures, optimize coordination of care, develop needed educational programs for the transitional care unit staff, and break down obstacles that interfere with the success of the program.

The Bloomer Transitional Care Unit was initiated in mid 2001 with an initial focus on stroke and neurology patients. There was more than a doubling of bed days in the next couple of years. The concept was spread with establishment of Osseo Transitional Care Unit (TCU), which opened in January of 2004. Osseo initially developed a specialized focus on post cardiovascular surgery patients in addition to other diagnoses. The Osseo TCU staff and Eau Claire Cardiovascular Surgery department collaborated in peer review and the development of new pathways.

3.2 Results

Administrative and financial data along with clinical outcome measures were tracked as the Transitional Care programs were established. Figure 19.1 demonstrates growth of the Osseo and Bloomer Transitional Care Units, the programs utilize the swing bed days that are reimbursed by Medicare at cost +1 %. Figure 19.2 demonstrates an example of Transitional Care Unit Dashboard with tracking of functional independence measure, overall rating of care, likelihood of recommending hospital, and admission categories.

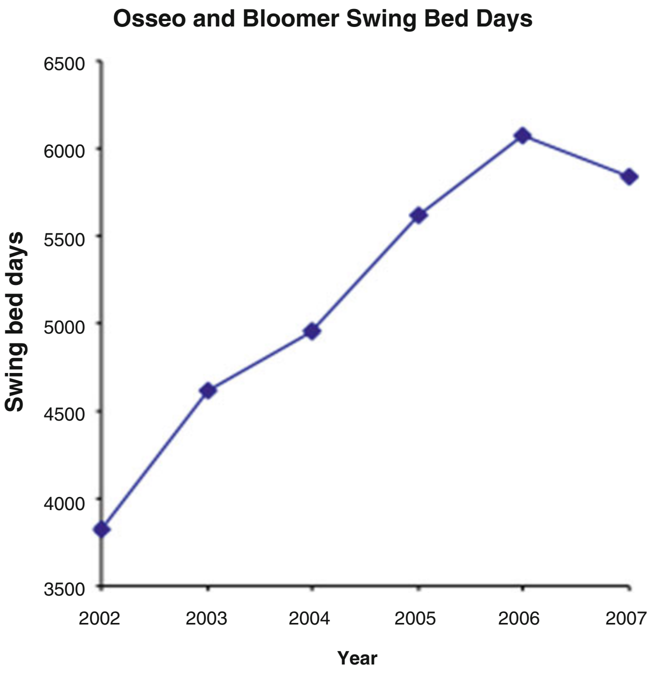


Fig. 19.1 Growth of Osseo and Bloomer transitional care unit

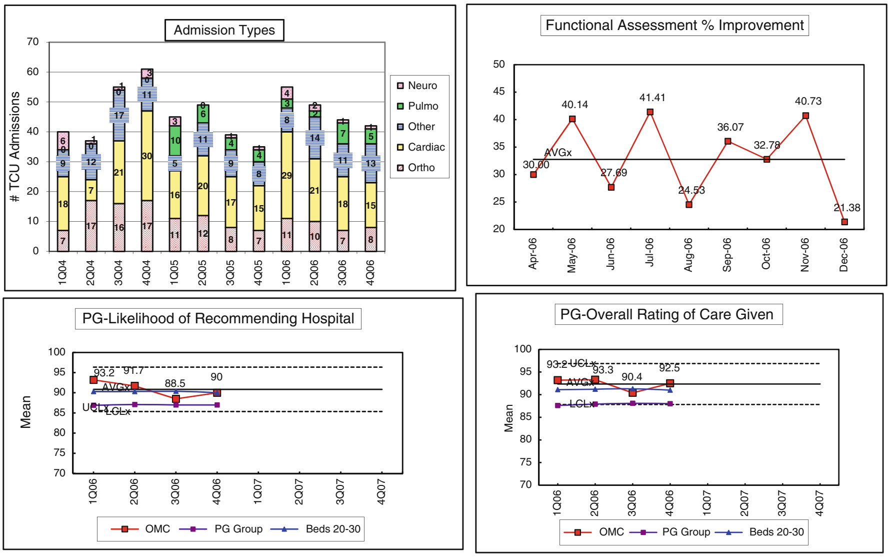


Fig. 19.2 Osseo and Bloomer transitional care unit dashboard

3.3 Financial Impact of Ventilator Program and Transitional Care Units

Transitional Care Units developed in Bloomer and Osseo had a significant financial impact for Osseo and Bloomer. Both facilities doubled their swing bed days. Bloomer had ten consecutive quarters with net negative NOI prior to mid 2001, followed by nine of the next ten quarters with positive NOI, linked to the doubling of the swing bed days.

The greatest impact is on the acute care hospital. Finance department performed an in-depth analysis looking at the financial impact of the Bloomer Transitional Care Unit along with the Chippewa Falls Ventilator Unit. The analysis focused solely on the financial impact of these programs on our acute care facility in Eau Claire, Wisconsin. The Bloomer Transitional Care Unit and the off-site nursing home ventilator unit had a positive financial impact to our Eau Claire Hospital of $2.99 million in 2003. This did not take into account the impact the ventilator program had on other complex respiratory patients, such as tracheostomy and Noninvasive Positive-Pressure Ventilation (NPPV) patients that were cared for in the ventilator unit.

4 Mayo Post Acute Care Program Proposal: Expanding the Wisconsin Transitional and Ventilator Care Program to 11 Facilities and Three States

Based on the successes of the Transitional Care and Ventilator programs in Wisconsin, a proposal was established to expand the Transitional Care and Ventilator program to 11 critical access hospitals in Minnesota, Wisconsin and Iowa (Fig. 19.3). The proposal included a detailed analysis of the Medicare bed days beyond the mean geometric length of stay in Mayo Clinic Rochester as well as analysis of underutilized capacity of the swing bed programs in the surrounding critical access hospitals. Clinicians participated in the analysis to determine whether outlier patients could have been discharged to post acute care facilities if adequate ventilator and transitional care capacity was available.

4.1 Culture Shift Requiring Collaboration of Mayo Clinic and Mayo Clinic Health System: Navigating a Proposal Through an Academic Medical Center

The Mayo Clinic Health System is a network of hospitals and clinics that are within an approximate radius of 150 miles of Mayo Clinic Rochester, both leadership

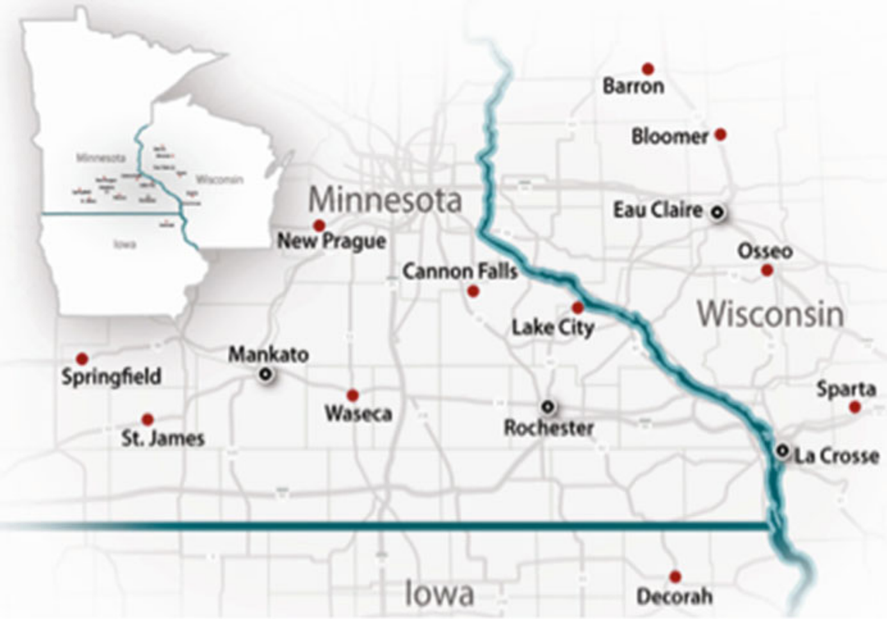


Fig. 19.3 Map of target region in Minnesota, Wisconsin and Iowa for Mayo post acute care program

teams needed to approve the proposal to move forward. The first step was contacting the leaders for each of the targeted critical access hospitals to determine level of interest in participating in the project. Their participation initially was strictly voluntary. All but one of the critical access hospital leadership teams opted to participate. The critical access hospital leaders recognized that this would require a culture shift for their administrative and clinical staff to be successful. A work team met regularly to finalize the detailed business plan. The Business Plan for Mayo Post was approved by leadership of Mayo Clinic and Mayo Clinic Health System in 2008.

4.2 Operations

The Mayo Clinic discharge planning team was instrumental in providing support and helping operationalize the program. Transitional Care and high quality off-site ventilator care provided new options for Mayo Clinic Rochester providers, patients, and families. Numerous presentations were provided to key stakeholders including key divisions and departments such as Surgery, Primary Care, Hospital Medicine, Primary Care, Surgery and Surgery Subspecialists.

Structured meetings occurred regularly with the critical access hospital Transitional Care and Ventilator Care teams. Meetings focused on ensuring that key elements of the program were established, outcome measures, training and education, staffing, bedside rounds, etc.

Centralized resources were essential to establishing the program. These resources included the Medical Director, administrative support, Respiratory Therapy Director, nurse educators, training and education, communication and marketing. These centralized resources included costs of approximately $900,0000 per year. Marketing materials were developed that included video descriptions of the ventilator and transitional care programs, brochures, Web marketing, and sharing of success stories through Mayo publications.

4.3 Key Elements and Benefits of Transitional and Ventilator Care

Key elements and expectations of the Transitional Care and Ventilator Care include:

- Multidisciplinary care team rounds
- Therapy offered 7 days per week
- Medical Director and Nurse lead positions identified
- Case Review
- Participation in data submission for dashboards
- Participation in all Mayo Post Acute Care training, competencies, and education

Benefits to Mayo Clinic and Mayo Clinic Health System critical access hospitals include:

- Improves outcomes across care settings.
- Reduces hospital readmissions.
- Provides high quality post acute care promoting shorter acute care lengths of stay, reducing excessive acute care hospital costs.
- Provides alternative high quality post acute care that addresses the quality gap that has been identified in the literature.
- Critical access hospitals are essential for ensuring access to healthcare for rural communities (well documented disparities identified in rural communities, fewer primary care physicians).
- Critical access hospitals are often times one of the largest employers in rural communities and are the centerpiece for promoting health and wellness.
- The ability to add resources to critical access hospitals such as respiratory therapists, and other key providers not only positively impacts the quality of Transitional Care but also increases capabilities for acute care and outpatient visits in these rural communities.

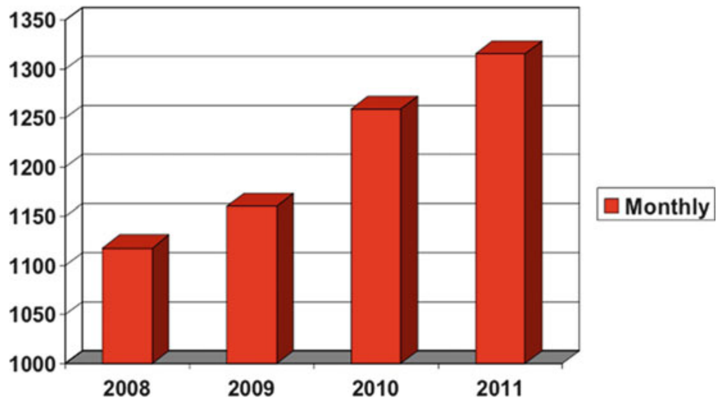


Fig. 19.4 Mayo post acute care bed days

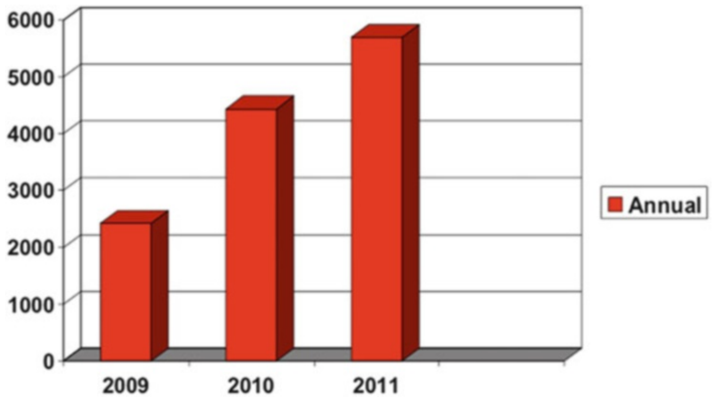


Fig. 19.5 Mayo post acute care bed days from Rochester admissions

- Transitional and Ventilator Care patients are high risk patients who benefit from higher hospital level staffing available in critical access hospitals that may not be present in skilled nursing facilities.

4.4 Results

Figures 19.4 and 19.5 demonstrate the growth of the Transitional Care Program and the increase in bed days coming from Mayo Clinic Rochester patients.

Table 19.4 demonstrates patient disposition outcomes with the vast majority of patients discharging to home or their previous care setting. The percentage of

Table 19.4 Outcomes: patient disposition

Discharged to	Percentage
Previous setting	72
SNF	14
Hosp. >30 days	2
Hosp. <30 days	6
Home	68
Rehab	5
Other	3
Asst. living	2

admissions to the hospital within 30 days was 6 %, well below other benchmarks in the literature.

Patient satisfaction for the program has been very positive with 94 % of patients rating the care very good and 92 % recommending the program. Many of the critical access hospitals participating in the program have also demonstrated very high employee satisfaction and teamwork scores.

The financial impact for the program was positive for the critical access hospitals and Mayo Clinic Rochester. The largest financial gains are realized by the acute care hospital through cost avoidance. As the majority of bed days that are impacted by moving patients out earlier from Mayo Clinic Rochester to Transitional Care are Medicare bed days beyond the mean geometric length of stay, cost avoidance for the acute care hospital is significant. The return on investment exceeded a ratio of 20, calculated as follows:

$$(\text{New revenue} + \text{Cost avoidance})/\text{Centralized resources} > 20/1$$

The tracking of administrative and outcome data was essential in gaining momentum for the program and gaining leadership support for moving the program from project into the long-term operations.

4.5 Collaboration of Acute Care Hospitals with Transitional and Ventilator Care

Mayo Post Acute Care Program has and will continue to benefit Mayo Clinic, Mayo Clinic Health System acute and critical access hospitals, and most importantly the patients. Mayo Post Acute Care Program was designed to optimize care as patients transition from acute care hospitals to post acute care with emphasis on care coordination, clinical outcomes, reducing hospital readmissions, and promoting value across the Care Continuum.

4.6 Care Coordination and Other Key Indicators

Mayo Post Acute Care Program, Transitional and Ventilator Care, has demonstrated value to key stakeholders. It is important to highlight additional best practices noted in the literature. Naylor developed a Transitional Care Model emphasizing care coordination for high risk elderly patients with chronic disease (Naylor et al. 1999, 2004). Interventions included transitional care nurse coordinating care across an episode of illness, home follow-up visits, identification and early response to health risks, and engagement of families, caregivers and providers with emphasis on communication and education. They demonstrated an overall reduction in healthcare costs, improved patient satisfaction and fewer re-hospitalizations. Dr. Eric Coleman and his team from University of Colorado developed The Care Transitions Program, demonstrating significantly reduced patient rehospitalization rates at 30, 90, and 180 days compared to controls (Coleman and Smith 2004). They emphasized patient centered record, follow-up with physician, knowledge of "red flags" or warning signs, and medication self management.

One of the key benefits of the Mayo Transitional Care program has been the hospital level staffing with a high nurse to patient ratio. Poor nurse staffing levels have been linked to increased mortality, poor patient outcomes, and poor staff satisfaction (Aiken and Clarke 2002; Needleman and Buerhaus 2002).

5 Quality and Patient Safety Perspective

We have already highlighted that elderly patients are at high risk for poor outcomes, hospital readmissions, and mortality. We have also shared some of the strategies applied with the ventilator and transitional care programs that are focused at improving quality and patient safety. Team centric, patient centered, processes are essential to ensure patient safety and quality outcomes. Ventilator patients are at very high risk for multiple transitions, morbidity and mortality. We will highlight one hospital's example of the importance of team centric processes in achieving top performance on quality indicators.

The medical literature would strongly support efforts to ensure high compliance of evidence based care. Malone reported that when patients were not treated with established pneumonia guidelines, they had a more than fourfold increase in mortality and significantly increased length of stay (Malone and Shaban 2001). Patients with severe congestive heart failure treated with enalapril had a 31 % reduction in mortality at 1 year (Consensus Trial Study Group 1987). Pfeffer reported that patients with left ventricular dysfunction after myocardial infarction treated with captopril had a 21 % reduction in cardiovascular deaths (Pfeffer and Braunwald 1992). RAND research conducted one of the largest studies on healthcare quality in the USA (Rand Health 2006). They reported in the largest study performed that only 55 % of participants received the recommended care. These findings were consistent in all geographic areas studied. They determined that diabetics received only 45 % of the recommended care, hypertensive patients less than 65 % of recommended care, coronary artery disease patients received 68 % of recommended care, and pneumonia patients 39 % of recommended care. The report concluded that there were significant gaps in the medical knowledge and the actual care provided.

5.1 Ventilator Bundle

It is well recognized that ventilator patients are at high risk for complications including deep vein thrombosis, stress ulcers and nosocomial pneumonia. Resar reported that hospital units that obtained high compliance on four prevention measures (stress ulcer prophylaxis, deep vein thrombosis prophylaxis, holding sedation, and head of bed elevation), had a significant reduction in nosocomial pneumonia (Resar and Pronovost 2005). He reported units that had 95 % compliance on the four prevention measures had a 59 % reduction in nosocomial pneumonia. Interestingly, improving the measures themselves should not directly reduce nosocomial pneumonias. Resar has postulated that those facilities that were highly compliant (95 %) were also very likely to be compliant on other elements that are important to caring for ventilator patients and preventing pneumonia. Resar highlights that it requires significant teamwork to accomplish 95 % compliance and that teamwork is also a likely factor in providing overall high quality care for those ventilator patients and achieving the reduction in nosocomial pneumonia.

When we looked at our initial results at Mayo Clinic Health System Eau Claire on ventilator bundle performance, consistently less than 50 % of our patients had documented evidence that all four measures were implemented. The poor results led to a concerted team approach to design a process to achieve the desired results. A standardized protocol was developed that allowed the nursing staff to implement the prevention measures even if the physician did not write the order. Additional checks, balances and redundancies were implemented utilizing the respiratory therapists and nursing to optimize the compliance. Figure 19.6 demonstrates very high compliance on the four prevention efforts for ventilator patients in our ICU. These outcome measures were displayed in our ICU for families and staff to view. This transparency is another key feature that is quite powerful.

5.2 Luther Midelfort Performance on Core Measures

Dr. Darren Lokkesmoe provided leadership for Luther Midelfort's Core Measure performance and was also instrumental in sharing and spreading important lessons to other sites across Mayo Clinic Health System. Dr. Lokkesmoe promoted a team

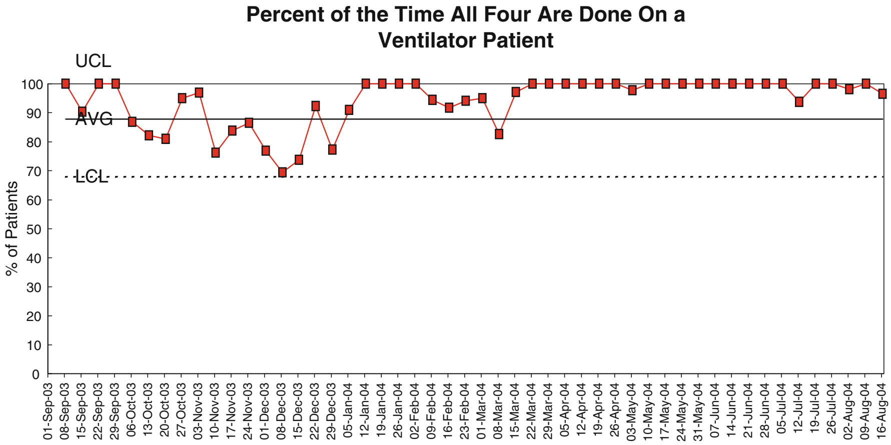


Fig. 19.6 Ventilator bundle compliance

centric approach that included standardization, necessary checks, balances, and redundancy. There was also a strong focus on transparency at the provider and system level. Dr. Lokkesmoe and his team promoted a frontline focus that empowered those people closest to the work to design and redesign the care processes. Luther Midelfort (Mayo Clinic Health System Eau Claire) achieved top 1 % performance on 22 process-of-care measures, which was third highest hospital in the country out of approximately 2,000 hospitals (Edwards 2008).

6 Chronic Disease Prevention: Avoiding the Acute Care Hospitalization on the Front End

We have already highlighted that chronic disease affects more than 130 million Americans with significant impact on healthcare costs, acute care hospitalizations, readmissions, morbidity and mortality (CDC). Obesity, diabetes and hypertension are growing rapidly in prevalence. Despite advances in chronic disease management most patients with diabetes do not have their blood pressure at goal.

6.1 Hypertension and Diabetes Multi-site Project

We provide an example of a multi-site quality improvement project that applies a care team bundle, team centered approach, patient engagement and transparency as key principles to improve performance of patients with hypertension and diabetes

(Lindsay and Hovan 2013). With the projected shortages of primary care physicians in the future, care team based efforts that empower nurses and other care team members to engage patients and promote evidence based guidelines will be critical to best manage the growing chronic disease and aging population.

6.2 Care Bundle

One of the greatest challenges for our project team which included four primary care teams from three states (Minnesota, Florida, and Arizona) was the fact that we had very different processes at each of the four sites. We tapped into the wisdom of Dr. Roger Resar's care bundle concept as a key principle that we believed could be tied to our outcome measure (Resar and Griffin 2012). Our goal was to improve the proportion of patients with diabetes that have their blood pressure at goal.

We pulled together the talent and expertise of the care team providers at the four sites and worked toward identifying three to four evidence based bundle elements that we believed could positively impact our outcome measures if we were able to achieve 90 % compliance on the all or none bundle elements. The three bundle elements selected by the care teams included:

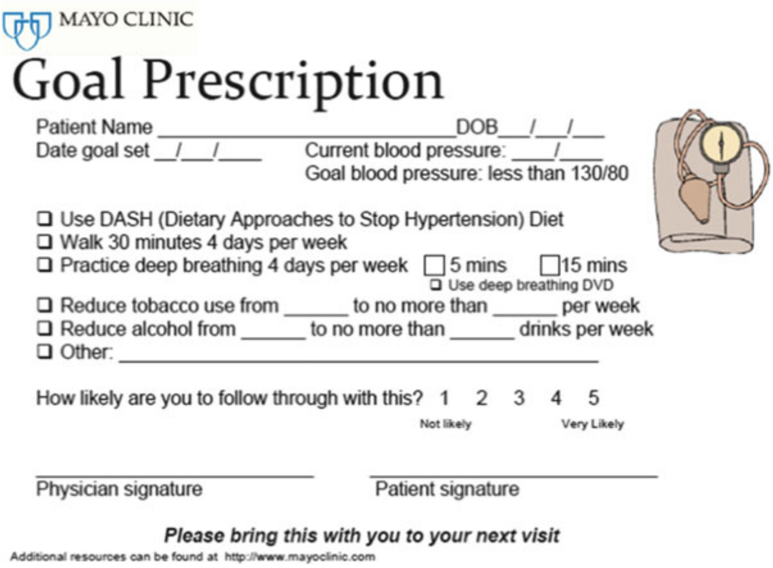
1. Standardized blood pressure process
2. Team based order set
3. Patient identified goal

6.3 Key Principles

One of the key elements of the team based order set is that it applied rational medication management with timely follow-up. It also allowed for nursing staff to make adjustments of the medications when patients returned with their timely follow-up as physician availability was sometimes an obstacle.

Standardized blood pressure process was adopted by all four sites, utilizing uniform technique, recording up to three readings. It is clear in the literature that poor blood pressure process technique can lead to inaccurate values.

Patient identified, evidence based goal was also selected as a bundle element to engage the patient in their care and emphasize the value of lifestyle changes that could positively impact their blood pressure. Figure 19.7 provides an example of a tool that was provided to promote patient engagement in selecting a goal.



MAYO CLINIC

Goal Prescription

Patient Name ______________________ DOB__/__/__
Date goal set __/__/___ Current blood pressure: ___/___
Goal blood pressure: less than 130/80

- □ Use DASH (Dietary Approaches to Stop Hypertension) Diet
- □ Walk 30 minutes 4 days per week
- □ Practice deep breathing 4 days per week □ 5 mins □ 15 mins
  □ Use deep breathing DVD
- □ Reduce tobacco use from ______ to no more than ______ per week
- □ Reduce alcohol from ______ to no more than ______ drinks per week
- □ Other: ______________________________

How likely are you to follow through with this? 1 2 3 4 5
Not likely Very Likely

Physician signature Patient signature

Please bring this with you to your next visit

Additional resources can be found at http://www.mayoclinic.com

Fig. 19.7 Patient identified goal

6.4 *Data Collection and Transparency*

We collected data for 12 weeks to identify a baseline for the proportion of patients with their blood pressure in control (less than 130/80) followed by a phase where we implemented the process changes. We then measured results after the process changes and bundle were fully implemented. We strongly encouraged transparency of data and each site committed to sharing their results in a public area with patient friendly language describing the intent of the project. Figure 19.8 shows examples of the transparency at the four primary care clinics in Minnesota, Florida and Arizona.

6.5 *Results*

After implementation of the bundle, proportion of patients with uncontrolled blood pressure decreased in three of the four sites ($p < 0.0001$) (Lindsay and Hovan 2013). This project demonstrated the value of care bundle in ambulatory care setting focused on improving outcomes of hypertension in diabetes. The strategy of applying a care team bundle to improving other chronic disease outcomes has merit.

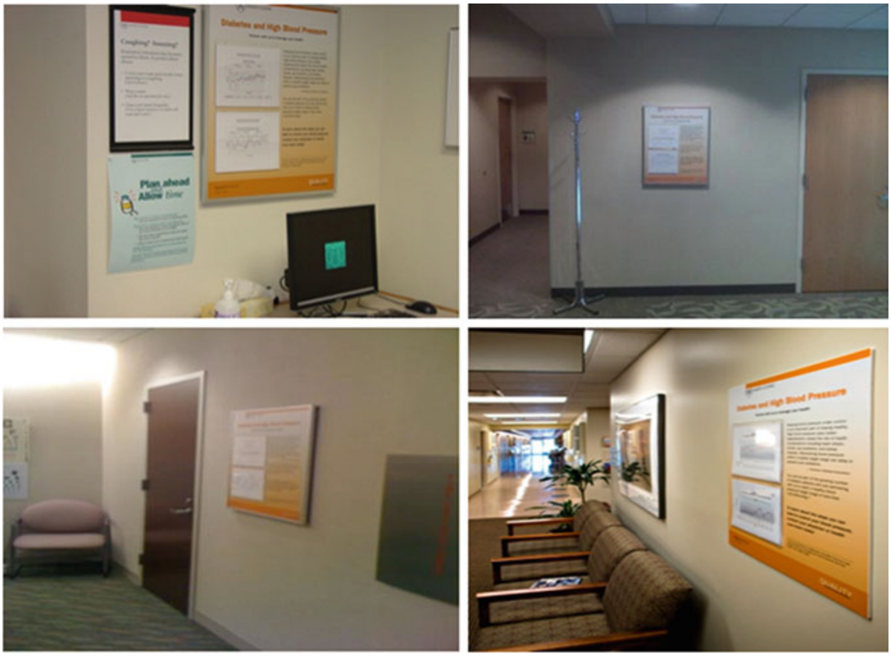


Fig. 19.8 Transparency of data in patient care area

6.6 Future Opportunities

There are significant potential gains in positively impacting flow as healthcare organizations continue to pursue effective strategies to care for chronic disease. Those efforts that engage patients and families in their own care, empower care teams, reliably apply evidence based care, promote transparency will be most successful. There are tremendous opportunities to positively impact flow through chronic disease improvement efforts.

7 Care Continuum

The purpose of this chapter has been to provide examples across the care continuum to positively impact flow. The Mayo Post Acute Care Program example provides strategies that can clearly be applied to other acute care hospitals. Hospitals have the opportunity to partner with post acute care entities such as skilled nursing facilities or other post acute care providers to optimize the transitions, reduce costly readmissions, and improve patient care. Hospital based team centric processes, such as successfully applying the ventilator bundle and improving compliance of the core measures, will reduce unnecessary complications and avoid delays, morbidity and potentially mortality.

The concept of care continuum acknowledges that there are opportunities to optimize care in ambulatory care, hospital, and post hospital settings. We have the potential to better coordinate care as patients transition from ambulatory care to hospital and post acute care settings. Team centric process improvement, empowered care providers, transparency of data will all be essential to optimizing care across the continuum.

8 Cases to Demonstrate the Strength of the Care Continuum

8.1 Case 1: Avoiding the ICU and Improving Outcomes

A patient with Chronic Obstructive Pulmonary Disease (COPD) and pulmonary edema was seen in the emergency room at 8 a.m. in severe respiratory distress. Arterial blood gas studies at 8 a.m. demonstrated a pH of 7.23, pCO_2 of 79 (on 15 l of oxygen). The respiratory therapist initiated the respiratory therapy directed NPPV protocol. The patient clinically improved with a follow-up ABG at noon, demonstrating a pH 7.41, pCO_2 36, and pO_2 78 (8 l of oxygen). The patient was admitted to telemetry, avoiding the ICU.

8.2 Case 2: Second Case in 24 h

A patient was admitted with COPD and an occipital stroke. The patient developed respiratory distress at 2 a.m. Arterial blood gas studies demonstrated hypercapnic respiratory failure with a pH of 7.26, pCO_2 of 92. The patient was placed on the respiratory therapy directed NPPV protocol on the floor. Follow-up blood gas studies at 6 a.m. were much improved with a pH of 7.39, pCO_2 of 66. The patient was not transferred to the ICU.

Pearls

1. Cases 1 and 2 are two patients who were cared for in the same 24 h period that were in respiratory failure. Both patients were started on NPPV protocol initiated by respiratory therapy.
2. Both patients not only avoided intubation but also avoided the ICU.
3. NPPV use reduces the need for intubation, improves outcomes, and avoids unnecessary ICU care and associated costs (Plant et al. 2000).
4. NPPV protocol also empowers respiratory therapy staff and improves job satisfaction (Lindsay and Schauer 2005).

8.3 Case 3: Best Practices Can Provide Very Effective "Pull Systems"

A patient with severe kyphoscoliosis and restrictive and obstructive lung disease. The patient was admitted to Luther Hospital for acute respiratory failure. The patient was hospitalized for several weeks and was transferred to the off-site nursing home ventilator unit. After several months, the patient was successfully converted to NPPV and discharged home. The Medical House Call advance practice provider participated on team rounds at the ventilator unit and provided home follow-up for several months post discharge. This patient was successfully cared for at home for more than 1 year without readmission.

Pearls

1. This case demonstrates the benefits of investing in each of the care settings. This patient had very high risk factors for readmission. The Nurse Practitioner who cared for the patient in the Nursing Home Ventilator Unit also cared for the patient at home.
2. This case also demonstrates the benefits of working with a single durable medical equipment provider who provided the equipment and supplies in the nursing home ventilator unit and the home supplies.
3. The ability to convert patients from mechanical ventilation to NPPV increases the likelihood of home discharge, maintains flow through the ventilator unit and acute care hospitals, and dramatically reduces the cost of care.

8.4 Case 4: Invest in All the Care Settings as Patients Transition from One Setting to the Next

An octogenarian was admitted to Luther Hospital with pneumonia and myocardial infarction. He had a prolonged hospitalization and was transferred to the Osseo Transitional Care Unit where the patient received extensive rehabilitation. The patient was subsequently discharged home with follow-up from the Medical House Call advance practice provider.

Pearls

1. This case demonstrates emphasis of the care continuum with investment in each of the care settings.
2. The Transitional Care Unit has a much higher nursing ratio and collaborates with the Cardiovascular Surgery program at the acute care hospital.
3. The mortality rate for skilled nursing facilities caring for post CABG patients has been reported as high as 13 %. The mortality rate for post CABG patients in the TCU program is very low (Carey and Parker 2003).

4. Nurse practitioner saw the patient in the home setting 3 days after discharge from the TCU program.
5. Nurse Practitioner programs that have evaluated patients in the hospital with follow-up in the home have resulted in a significant reduction in readmissions and in one report a 50 % reduction in Medicare dollars spent (Naylor and Brooten 1999).

8.5 Case 5: Advance Directives and Advance Care Plans

The patient is an elderly male with cardiomyopathy and progressive deterioration of his health. The patient has had increasing dyspnea, need for multiple procedures to withdrawal fluid from his lung, and decreased mobility. The patient's daughter viewed a video called "Choices" that described the benefit of advance directives and advance care planning. The development of the video was a community effort to improve community education on advance directives, advance care planning and end of life care. As a result of watching the video, the daughter initiated discussion with her father regarding his wishes and an advance directive and care plan was discussed and implemented. Within a couple of weeks of filling out the advance directives, the patient was involved in a car accident and was brought to the emergency room. The daughter respected her father's wishes and no heroic measures were implemented. The daughter shared her sense of relief with the physician that she knew her father's wishes.

Pearls

1. This case demonstrates the importance of advanced directives and advance care planning, respecting patient wishes and the impact it has on allowing patients to have a dignified death and avoid unnecessary and costly intensive care in patients that have serious chronic illness.
2. The studies have demonstrated that the majority of patients with chronic illness have not filled out an advanced directive. There are opportunities to emphasize the importance of advanced directives and advance care planning (Lynne and Goldstein 2003; Morrison and Meier 2004).
3. Tools such as video education of important topics such as advanced directives "Choices Video" which educates patients and families on advance directives and advance care plan can be developed.

Acknowledgments It is important to acknowledge the commitment of Mayo Clinic Health System and Mayo Clinic leadership and care teams who have strongly supported these programs. Special thanks to Dr. Terry Borman, Dr. Randy Linton, Dr. Wiechmann, Dr. Carmody, Dr. Lokkesmoe, Karl Palmer RN, Steve Underdahl, Jerry Ehn, Dr. Jeff Rabatin, Lisa Carter RN, Whitney Schauer RT, Gregg Thomas, Dr. Mike Rock, and countless others who have contributed to this work.

References

Affordable Care Act Update. (2010). http://www.cms.gov/apps/docs/aca-update-implementing-medicare-costs-savings.pdf

Aiken, L., & Clarke, S. (2002). Hospital nurse staffing and patient mortality, nurse burnout and job dissatisfaction. *JAMA, 288*(16), 1987–1993.

Ankrom, M. A., & Barofsky, I. (1998). What happens to patients in a nursing home-based chronic ventilator unit: A five-year retrospective review of patients and outcomes. *Annals of Long-Term Care, 6*(10), 309–314.

Bagley, P. H., & Cooney, E. A. (1997). Community-based regional ventilator weaning unit: Development and outcomes. *Chest, 111*(4), 1024–1029.

Callahan, E. H., & Thomas, D. C. (2002). Geriatric hospital medicine. *The Medical Clinics of North America, 84*(4), 707–729. Review.

Carey, J. S., & Parker, J. P. (2003). Hospital discharge to other healthcare facilities: Impact on in-hospital mortality. *Journal of the American College of Surgeons, 197*(5), 806–812.

Center for Disease Control and Prevention CDC. (2013) http://www.cdc.gov/chronicdisease/resources/publications/aag/chronic.htm

Coleman, E. A., & Smith, J. D. (2004). Preparing patients and caregivers to participate in care delivered across care settings: The care transitions intervention. *Journal of the American Geriatrics Society, 52*(11), 1817–1825.

Consensus Trial Study Group. (1987). Effects of enalapril on mortality in severe congestive heart failure. Results of the Cooperative North Scandinavian Enalapril Survival Study (Consensus). *The New England Journal of Medicine, 316*(23), 1429–1435.

Cook, C. H., & Martin, L. C. (1999). Survival of critically ill surgical patients discharged to extended care facilities. *Journal of the American College of Surgeons, 189*(5), 437–441.

Decker, F. H., & Gruhn, P. (2003). *Results of 2002 American Health Care Association Survey of Nursing Staff Vacancy and Turnover in Nursing Homes*. www.ahca.org

Dodek, P. M., & Raboud, J. (2003). Explicit approach to rounds in an ICU improves communication and satisfaction of providers. *Intensive Care Medicine, 29*(9), 1584–1588.

Edwards, J. (2008). *Luther Midelfort Mayo Health System: Laying tracks for success*. Commonwealth Fund pub. 1194 (Vol. 2).

Ezzati, M., & Lopez, A. (2002). Selected major risk factors and global and regional burden of disease. *Lancet, 360*, 1347–1360.

Garibaldi, R. A., Brodine, S., & Matsumiya, S. (1981). Infections among patients in nursing homes: policies, prevalence, problems. *The New England Journal of Medicine, 305*(13), 731–735.

Gracey, D. R., Hardy, D. C., & Koenig, G. E. (2000). The chronic ventilator-dependent unit: A lower-cost alternative to intensive care. *Mayo Clinic Proceedings, 75*(5), 445–449.

Greene, J. (2002). Rural renewal. As momentum builds, critical access hospital program shows signs of success. *Hospitals and Health Networks, 76*(4), 50–54. 2.

Institute of Medicine. (1999). In L. T. Kohn, J. M. Corrigan, & M. S. Donaldson (Eds.), *To err is human: Building a safer health system*. Washington, DC: National Academy Press.

Institute of Medicine. (2001). In L. T. Kohn, J. M. Corrigan, & M. S. Donaldson (Eds.), *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Jackson, C., & Trygstad, T. (2013). Transitional care cut hospital readmissions for North Carolina Medicaid patients with complex chronic conditions. *Health Affairs, 32*, 81407–81415.

Kollef, M., & Shapiro, S. (1997). A randomized controlled trial of protocol-directed versus physician-directed weaning from mechanical ventilation. *Critical Care Medicine, 25*, 567–574.

Law Watch. CMS Proposes expansion of post acute care transfer payment policy. www.folev.com

Leape, L. L., & Cullen, D. J. (1999). Pharmacist participation on physician rounds and adverse drug events in the intensive care unit. *JAMA, 282*(3), 267–270.

Lindsay, M. E., & Bijwadia, J. S. (2004). Shifting care of chronic ventilator-dependent patients from the intensive care unit to the nursing home. *Joint Commission Journal on Quality and Safety, 30*(5), 257–265.

Lindsay, M., & Hovan, M. (2013). A multisite quality improvement project that applies a 3-step care bundle to a chronic disease model for diabetes with hypertension. *American Journal of Medical Quality, 28*, 365–373.

Lindsay, M., & Schauer, W. (2005). Standardized protocols, empowered staff and NPPV improve hospital flow. *Advance, 14*(9), 19–20.

Lynne, J., & Goldstein, N. E. (2003). Advance care planning for fatal chronic illness: Avoiding commonplace errors and unwarranted suffering. *Annals of Internal Medicine, 138*(10), 812–818.

Malone, D. C., & Shaban, H. M. (2001). Adherence to ATS guidelines for hospitalized patients with community-acquired pneumonia. *Annals of Pharmacotherapy, 35*, 1180–1185.

Marelich, G. P., & Murin, S. (2000). Protocol weaning of mechanical ventilation in medical and surgical patients by respiratory care practitioners and nurses: Effect on weaning time and incidence of ventilator-associated pneumonia. *Chest, 118*, 459–467.

Morrison, R. S., & Meier, D. E. (2004). Clinical practice. Palliative care. *The New England Journal of Medicine, 3550*, 2582–2590.

National Healthcare Expenditures. http://www.census.gov/compendia/statab/2012/tables/12s0134.pdf

Naylor, M. D., et al. (1999). Comprehensive discharge planning and home follow-up of hospitalized elders: randomized clinical trial. *JAMA, 281*(7), 613–620.

Naylor, M. D., Brooten, D. A., Campbell, R. L., Maislin, G., McCauley, K. M., & Schwartz, J. S. (2004). Transitional care of older adults hospitalized with heart failure: a randomized, controlled trial. *Journal of the American Geriatrics Society, 52*(5), 675–684.

Needleman, J., & Buerhaus, P. (2002). Nurse-staffing levels and the quality of care in hospitals. *The New England Journal of Medicine, 346*, 1715–1722.

Pfeffer, M. A., & Braunwald, E. (1992). Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction. Results of the survival and ventricular enlargement trial. The SAVE Investigators. *The New England Journal of Medicine, 327*(10), 669–677.

Plant, P. K., Owen, J. L. & Elliott, M. W.(2000). Early use of non-invasive ventilation for acute exacerbations of chronic obstructive pulmonary disease on general respiratory wards: A multicentre randomized controlled trial. *Lancet, 355*, 1931–1935.

Rand Health. (2006). *Research highlights*. The First National Report Card on Quality of Health Care in America. www.rand.org

Resar, R., & Griffin, F. (2012). *Using care bundles to improve health care quality* (IHI Innovation Series white paper). Cambridge, MA: Institute for Healthcare Improvement.

Resar, R., & Pronovost, P. (2005). Using a bundle approach to improve ventilator care processes and reduce ventilator associated pneumonia. *Joint Commission Journal on Quality and Patient Safety, 31*(5), 243–248.

Scheinhorn, D. J., Chao, D. C., Stern-Hassenpflug, M. A., LaBree, L. D., & Hetsley, D. J. (1997). Post-ICU mechanical ventilation: Treatment of 1123 patients at a regional weaning center. *Chest, 111*(6), 1654–1659.

Suh, D. C., & Kim, C. M. (2009). Trends in blood pressure control and treatment among type 2 diabetes with comorbid hypertension in the United States: 1988–2004. *Journal of Hypertension, 27*, 1908–1916.

Unroe, M., & Kahn, J. M. (2010). One-year trajectories of care and resource utilization for recipients of prolonged mechanical ventilation: A cohort study. *Annals of Internal Medicine, 153*(3), 167–175.

Wang, J., & Imai, K. (2009). Secular trends in diabetes-related preventable hospitalizations in the United States, 1998–2006. *Diabetes Care, 32*(7), 1213–1217.

Weissman, C. (2005). The enhanced postoperative care system. *Journal of Clinical Anesthesia, 17* (4), 314–322.

Young, M. P., Gooder, V. J., Oltermann, M. H., et al. (1998). The impact of a multidisciplinary approach on caring for ventilator-dependent patients. *International Journal for Quality in Health Care, 10*(1), 15–26.

Zilberberg, M. D., de Wit, M., Pirone, J. R., & Shorr, A. F. (2008). Growth in adult prolonged acute mechanical ventilation: Implications for healthcare delivery. *Critical Care Medicine, 36*(5), 1451–1455.

Chapter 20
A Logistics Approach for Hospital Process Improvements

Jan Vissers*

Abstract This chapter proposes a sustainable logistics approach for hospital process management as an alternative to traditional quality management. The underlying concepts are discussed and illustrated with material from case studies regarding cardiology patient flows. The distinctions between unit logistics, chain logistics, and their combination network logistics are used to illustrate the focus of logistic improvement. A framework for hospital production control offers support for a systematic matching of supply and demand at different levels of planning and control. The concept of a focused factory and business unit is used to create a context for patient group management. These three concepts are used for a systematic and sustainable approach for hospital process improvement, as illustrated for an important patient flow in hospitals, i.e., cardiology.

Keywords Health operations management • Patient flow logistics • Patient group management • Cardiology

*The author would like to thank his colleague Geerhard de Vries for his work on objective setting for process management in hospitals and the hospitals in which the projects on cardiology process improvement took place.

J. Vissers* (✉)
Institute for Health Policy and Management, Erasmus University Medical Centre,
Rotterdam, The Netherlands

Prismant Institute for Health Management Development, Utrecht and Eindhoven
University of Technology, Eindhoven, The Netherlands
e-mail: jan.vissers@kiwa.nl

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_20, © Springer Science+Business Media New York 2013

1 Introduction

How can service improvements be both realized and sustained in hospital processes? Over the years many improvements have been started successfully but did not survive the pilot phase. These efforts were often part of a quality management improvement program, and they demonstrated clearly that considerable gain is to be made in hospital process management. However, these improvements were not sustained for a number of reasons: the innovation concerned only an isolated process for a specific group of patients, no consideration was given to the interaction between processes and resources, and process monitoring and process management was lacking. To make improvements in process management sustainable, a more fundamental approach is required.

In general, hospitals do not manage the patient processes but rather manage departments or units such as outpatient clinics, diagnostic departments, operating theaters, and wards. The process of the patient as such (referral, outpatient first visit to a specialist, diagnostic tests, second visit, admission to a ward, surgical procedure, rehabilitation, outpatient follow-up, referral back) is not yet mastered by hospital planning. One of the reasons is that no one is responsible for the whole trajectory of patients through the system. Therefore, no function assures that processes of individual patients are executed within the range of targets set for their patient groups.

This contribution will focus on an operations management approach to hospital process improvements, in which:

- Groups of patients are distinguished for which care delivery will be organized.
- Processes are described in main lines, using the expert knowledge of medical and other care professionals
- Care processes are organized in different ways, depending on their characteristics in terms of elective/acute, degree of predictability, use of bottleneck resources, etc.
- Responsibilities for process management are clarified, including objective setting for the performance of the process, and monitoring the performance of processes.
- Corrective measures are taken when processes are no longer within the bandwidth defined for its performance.

The chapter describes these steps, illustrated with examples for cardiology. This description is preceded with a section on principles of operations management, such as the distinction between logistics of units and chains and their combination, i.e., network logistics, and the systematic matching of demand and supply at different levels of planning, i.e., a framework for production control of hospitals.

2 An Operations Management Context for Hospital Process Improvement[1]

2.1 *Unit, Chain, and Network Logistics*

In our logistic approach we distinguish between unit logistics, chain logistics, and network logistics. Units carry out the same types of operation for different types of patients whereas processes/chains represent a series of different operations (undertaken in different units) for the same type of patient. Network logistics combines both perspectives. Figure 20.1 illustrates these different logistical perspectives for a hospital setting. It describes a hospital as a representation of units and chains.

As can be seen from the example, general surgery has its own outpatient facilities that are not shared with, for instance, general medicine. Diagnostic departments, such as radiology and pathology, are shared by all patient groups and specialties. Wards are shared by the different patient groups within a specialty, with sharing of beds between specialties limited to overflow. In addition, not all patient groups use the ward at a similar level. All surgical patient groups share the operating theater department, though some groups are treated without an

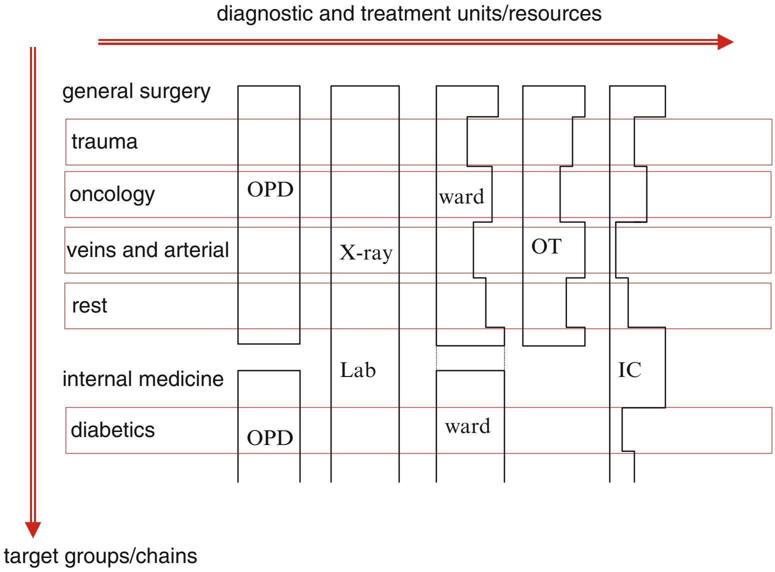


Fig. 20.1 Unit, chain, and network perspectives (OPD = outpatient department, OT = operating theaters, IC = intensive care)

[1] Based on Vissers and Beech (2005).

intervention. The IC is again shared by all specialties, though the use of the IC differs much between patient groups.

The unit perspective is represented by the units: OPD, X-ray/lab, ward, OT and IC. Managers of these units are responsible for the running of the unit, for the level of service the unit offers to physicians requiring a service on behalf of their patients, and for the efficient use of the resources available. This is regarded as a total responsibility. The unit's concern is the total flow of all patients requiring a service of the unit, as this determines the prime objective of the unit, i.e., reaching a high but balanced use of resources without peaks and troughs in the workload during the hours of the day and days of the week. High occupancy level or use is seen as an important indicator of the "efficiency" of the unit while balanced use is important not only for efficiency but also for the working climate of the personnel in the unit. Additional aims from the perspective of the unit are to produce the amount of output required with as less resources as possible or to produce as much output as possible with the amount of resources available (capacity management). The focus of unit logistics is therefore on the total flow of the patients using the unit, and on the effect of this flow on the use of resources and the workload of personnel.

The chain perspective is represented by patient groups, i.e., trauma patients, oncology patients, etc. The focus of this perspective is on the total process of the patient, using different units on their journey through the hospital. The chain perspective strives to optimize this process according to some targets, which all relate to the time dimension. Typical targets are: short access time, short throughput time, and short in-process waiting times. Short throughput time can be reached by combining operations in one visit to the OPD, instead of having to come twice, or by having finished the diagnostic phase before the admission. The prime objective of the chain perspective is to maximize the service level for patients belonging to a certain patient group. As the focus is on the one patient group considered, it is difficult to look at the efficiency of the chain, in terms of use of resources. Resources are, in general, not allocated to patient groups, but to specialties. Therefore, efficiency issues can only be considered at the level of flows from all patient groups belonging to the specialty.

Network logistics combines the unit and the chain perspectives. It draws on the notion that optimization of the service in the chains needs to be balanced with efficiency in the use of resources in the units. A network logistics approach will make explicit any trade-off to be made between the service level provided in the chains and the utilization of resources in the units: for example, a desire to improve patient access to diagnostic services by making these services available for 24 h per day might have a negative impact on the performance of diagnostic departments.

For a network logistics approach, ideally all chains and all units (i.e., the whole hospital) need to be included. However, this might be regarded as too complex, especially for a change to improve the performance of the process for a single patient group. There might be a desire to address such a change via a chain logistics approach. However, one should also strive for a network logistics approach by, for example, including all patient groups of the relevant specialty in the analysis. This would make it possible to look at the impact of the change on the use of resources available for both the specialty as a whole and for the other patient groups within it.

Table 20.1 Differences between the unit, chain, and network logistics approaches

Perspective item	Unit logistics approach	Chain logistics approach	Network logistics approach
Focus points	Resource utilization Workload control	Service level	Trade-off between service level and resource utilization
Strong point	Capacity management	Process management	Combination
Weak point	Not process oriented	Not related to the use of resources	More effort
Suitable for	Efficiency analysis of OT's, OPD's, etc.	(Re)design of a process	Redesign and efficiency

Consider a change that aims to improve patient access to physiotherapy by stroke patients. If there is a limited supply of physiotherapy services, improvements in the process of care for stroke patients might result in a reduced level of service for other patient groups both within neurology (the specialty that treats stroke patients) and within other specialties containing patient groups which require physiotherapy. These adverse consequences would go unnoticed if only a chain logistics approach is adopted. A network logistics approach therefore helps to avoid a situation where an improvement in one process goes unnoticed at the expense of a drawback for other processes.

The OM approach implies making explicit the choices to be made in a systems perspective. This serves also as a warrant for sub-optimization, i.e., an improvement in one part of the system goes at the expense of the functioning of the system as a whole. Table 20.1 summarizes the main differences between the unit, chain, and network logistics approach.

2.2 A Framework for Hospital Production Control

A further feature of our approach is the distinction between the levels of logistic decision-making, i.e., operational, tactical, and strategic. The framework we use as a reference in this contribution is a hierarchical framework for production control of hospitals that deals with the balance between service and efficiency, at all levels of planning and control. It shows analogies to frameworks used in industrial settings for manufacturing organizations. The framework is based on an analysis of the design requirements for hospital production control systems (De Vries et al. 1999) and builds on the production control design concepts developed in Bertrand et al. (1990). The design requirements are translated into the control functions at different levels of planning required for hospital production control. This translation is built on notions of the hospital as a virtual organization with patient groups as business units and a focused factory approach for the production control per business unit. In short we can distinguish a number of production control functions,

Table 20.2 Production control functions distinguished in the planning framework for hospitals

	Decision focus
1.	Range of services, markets and product groups, long-term resource requirements, centrally coordinated scarce resources; contracted annual patient volumes, target service levels
2.	Amount of resources available at annual level to specialties and patient groups, regulations regarding resource-use
3.	Time-phased allocation of shared resources, involving specialist-time detailed number of patients per period
4.	Urgency and service requirements, planning guidelines per patient group
5.	Scheduling of individual patients, according to guidelines at patient group level and resource-use regulations at resource level.

which can be positioned at different levels of planning in a framework (see Table 20.2).

At the highest level decisions have to be made on the range of services provided, the markets one wants to operate in and the product groups for each market. Also decisions have to be made on the long term resource requirements of the hospital, which scarce resources are centrally coordinated, what level of annual patient volumes one wants to achieve, what service philosophy will be used and what level of service one wants to target for. These are all longer term strategic decisions, which essentially do not belong to the domain of OM, but which have impact on the management of operations at shorter terms.

The next level focuses on the amount of resources that is available annually to specialties and patient groups, to ensure that the contracted annual patient volume can be realized. At this level the rules for using the resources also need to be established to ensure that the target service and efficiency levels are achieved. At the third level, the focus is on the allocation of shared resources in time, taking into account the availability of specialists and seasonal developments. This requires more insight into the detailed numbers per patient group per period within the year. At level four, the urgency and service requirements per patient group need to be established, and the planning guidelines per patient group. The fifth level regards the scheduling of individual patients, according to the planning guidelines for the patient group and the resource-use regulations for the resources involved.

Though the planning framework seems to be working only top down, the need for each level and the requirements for coordination are established bottom-up. At the lowest level, individual patients are coupled to resources in the day-to-day scheduling. This level in the framework is called *patient planning and control*. The way patients are operationally scheduled needs to be governed by rules established at patient group level. Oncology patients, for instance, have different urgency and service requirements from patients with varicose veins. Therefore, operational scheduling of patients needs to be governed by what we called *patient group planning and control*. To allow for the planning of a patient group resources need to be allocated, taking into account the availability of specialists and personnel. This level is called *resources planning and control*, and includes also the time-phased allocation of resources. The level of resources required results from the annual patient volumes contracted, and the service and efficiency levels targeted for. This level is called *patient volume planning and control*. Finally, the volume level is governed by the strategic planning level, where, for instance, decisions are taken about which resources need to be shared or not. This level is called *strategic planning*. At this level there is no control involved.

These levels of planning can be further elaborated (Vissers et al. 2001), resulting in the planning framework as shown in Fig. 20.2. The framework shows that every level needs a horizontal control mechanism to match patient flows with resources and that vertical control mechanisms are required to set the targets for lower levels (feed forward) or to check whether activities develop within the boundaries set by higher levels (feedback).

The framework can be used as a reference for improving the capability of the hospitals to deal with process development within the wider context of the hospitals as a whole.

2.3 Iso-process Patient Groups as Business Units

An important third feature of our approach is that iso-processes are considered as basis for production control of hospital processes. Iso-process grouping is a way of classifying patients according to the trajectory that patients follow through their patient journey. There are different ways to classify hospital products, depending on the focus of the classification.

Given that processes or chains generate a service for a client, the focus for product classifications is driven by the requirements of the client. In particular, clients want a service that is efficient (for example, unnecessary delays in treatment are avoided) and effective (for example, evidenced based practices are used). In turn, the achievement of these goals is likely to increase levels of client satisfaction. Hence, in processes or chains, the operations manager wants a product classification that allows them to plan and monitor the efficient and effective delivery of products.

This means that an iso-process perspective is required. Patient group in an operations management perspective can include different products/subgroups that are homogeneous in terms of market performance and process. Homogeneity in terms of market performance implies similar criteria for urgency, acceptable waiting times, etc. An example of such a subgrouping, based on market performance, could be that a product needs to be delivered on an emergency basis (e.g., process chain: attendance at emergency department, ward admission, outpatient follow-up) or on a scheduled basis (e.g., process chain: referral to outpatient department, elective admission, outpatient follow-up). The first subgroup will require a different planning approach then the second subgroup. Homogeneity in terms of process implies that the patients within the product group use the same constellation of resources. Patients requiring routine diabetes care (such as a one-off consultation) might be grouped with more complex patients who require more follow-up visits. This is because they are essentially using the same

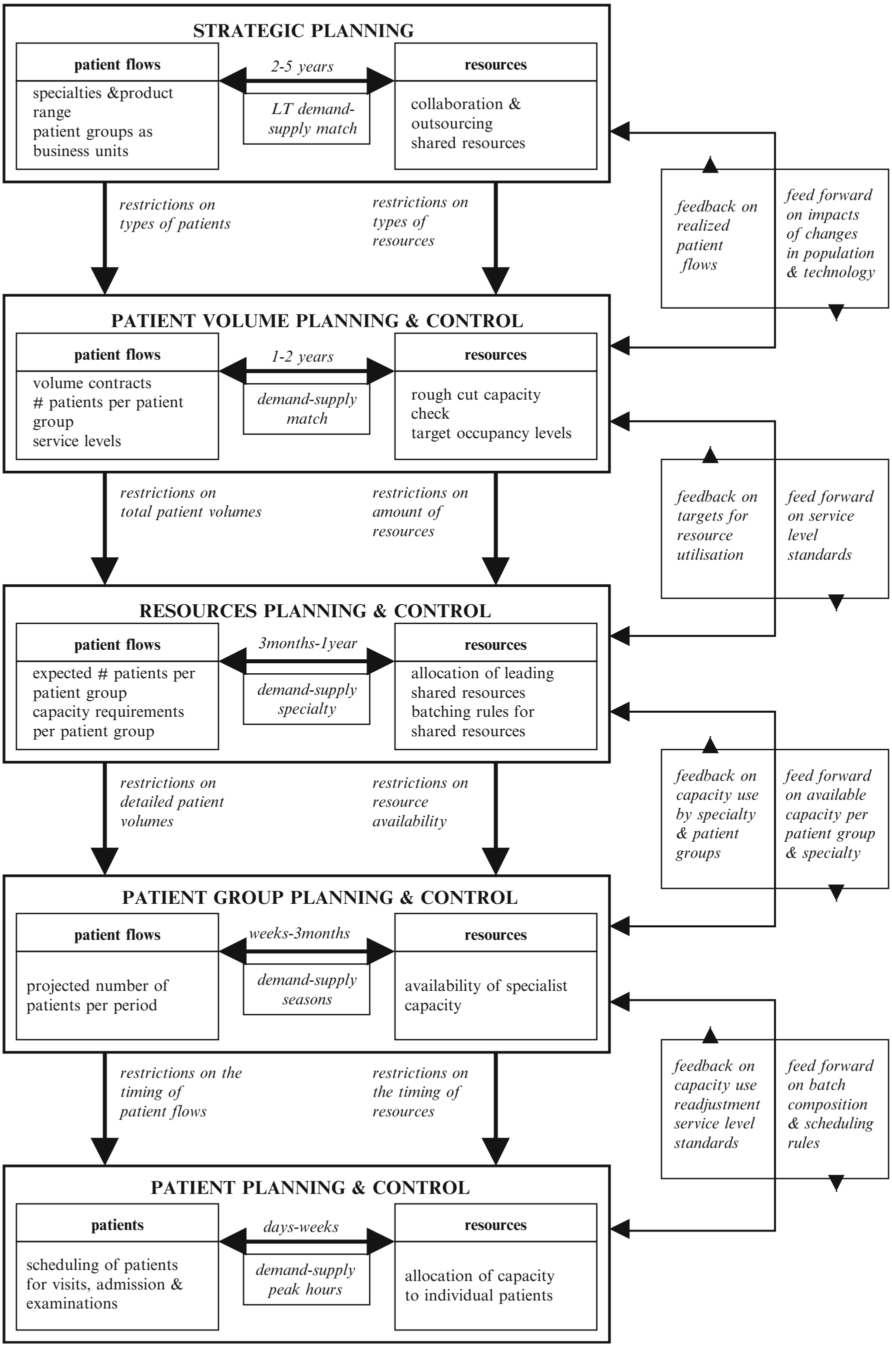

STRATEGIC PLANNING
patient flows
specialties &product range
patient groups as business units
2-5 years
LT demand-supply match
resources
collaboration & outsourcing
shared resources
restrictions on types of patients
restrictions on types of resources
feedback on realized patient flows
feed forward on impacts of changes in population & technology
PATIENT VOLUME PLANNING & CONTROL
patient flows
volume contracts
# patients per patient group
service levels
1-2 years
demand-supply match
resources
rough cut capacity check
target occupancy levels
restrictions on total patient volumes
restrictions on amount of resources
feedback on targets for resource utilisation
feed forward on service level standards
RESOURCES PLANNING & CONTROL
patient flows
expected # patients per patient group
capacity requirements per patient group
3months-1year
demand-supply specialty
resources
allocation of leading shared resources
batching rules for shared resources
restrictions on detailed patient volumes
restrictions on resource availability
feedback on capacity use by specialty & patient groups
feed forward on available capacity per patient group & specialty
PATIENT GROUP PLANNING & CONTROL
patient flows
projected number of patients per period
weeks-3months
demand-supply seasons
resources
availability of specialist capacity
restrictions on the timing of patient flows
restrictions on the timing of resources
feedback on capacity use readjustment service level standards
feed forward on batch composition & scheduling rules
PATIENT PLANNING & CONTROL
patients
scheduling of patients for visits, admission & examinations
days-weeks
demand-supply peak hours
resources
allocation of capacity to individual patients

Fig. 20.2 Framework for hospital production control

constellation of resources: for example, access to a clinician and a diabetic nurse. However, the overall amount of resources used by patients within the group may vary considerably, a fact that would need to be allowed for when planning capacity requirements. This iso-process grouping makes a logistics approach different from an economics approach (iso-resource grouping) and a medical approach (iso-diagnosis grouping).

The traditional way of classifying "individual" patient products in the acute hospital is according to their complaint or diagnosis. Iso-diagnosis groupings of patients, for instance, are based on well accepted international classification schemes such as the ICD (International Classification of Diseases). These classification systems can be very extensive: for example, the ICD-9-CM version of 1,979 counts 398 main groups and 7,960 subgroups. However, such product classifications are mainly used for medical purposes. The number of patient groups that they generate, and the fact that some may generate very few admissions during a planning period, mean that it is difficult and usually undesirable to use them to plan and schedule care from an operations management perspective.

Acute hospitals also traditionally group patients by specialty: for example, general medicine patients, orthopedic patients. However, these groupings are too aggregated from an operations management perspective, as the constellation of resources used by patient types within specialties are likely to be very different. For example, patients diagnosed with asthma or stroke might both be grouped under the specialty general medicine but the care that they receive will be very different.

Hence, from an operations management perspective, a product classification somewhere between these two "traditional" approaches seems to be required. The first attempt to define hospital products from a managerial perspective can be credited to Fetter (Fetter 1983). They developed the DRG-system (diagnosis related groups) to classify all diagnoses into groups of diagnoses that are recognizable for physicians and homogeneous in terms of use of resources. Up to then, X-rays, lab tests, medication, surgical procedures—in the DRG system seen as intermediate outputs—were considered as hospital outputs. Fetter developed 467 DRG's to describe the hospital's inpatient output.

Continuing lines of development have included Ambulatory Visit Groups (AVG's) for classifying ambulatory care products (Fetter et al. 1984), and a refinement of DRG's which take into account the stage of development of the disease with the patient (Fetter and Freeman 1986). Another line of development in The Netherlands—with many parallels to the DRG approach—is to define hospital products as combinations of diagnosis and treatment (Baas 1996). Similarly, in the UK, and again based on the DRG approach, Health care Resource Groups (HRGs) have been developed.

Product groupings such as DRGs were primarily developed to support the financial reimbursement of hospitals rather than to support the planning and management of health care chains. However, they have relevance to operations management as there will be a direct relationship between, for example, a hospital's DRG cost and the efficiency with which resources are used within a DRG. Hence, there are parallels between the analysis of DRG costs and the efficient planning of care within process chains.

Although specific groupings within, for example, the DRG system may be useful for operations management purposes, the overall number of groupings generated by such systems is again likely to be too large. In addition, products which use a similar amount of resources (iso-resource) will not necessarily use a similar constellation of resources (iso-process). For instance, a patient with a DRG/AVG profile of an admission of 5 days, five lab tests and three outpatient visits may represent a patient admitted on an emergency basis (with five tests during admission, and three outpatient visits to a specialist), as well as a patient admitted on a scheduled basis (with three preceding outpatient visits always using the same constellation of resources, i.e., the specialist, a specialized nurse, and the lab). Finally, the boundaries of health care chains may stretch beyond, for example, DRG boundaries. For example, the care chain for a patient who has suffered a stroke will include follow-up care in the community. However, the DRG(s) to which such patients are assigned will only embrace their care within the acute hospital.

Alternatively, it might be possible to generate product groups because the care of the patients covered can be regarded as being delivered in a "focused factory": a business unit concept. De Vries et al. (1999) specified the requirements for a "focused factory": a clear relationship between the product group and the resources required; the volume of activity is large enough to allow the allocation of dedicated resources; and it is possible in advance to identify the level of specialization required.

Some "focused factory" product groups might contain the same types of patient. For example, dedicated facilities and units for patients requiring treatment for cataracts have been established. In other "focused factory" product groups, different types of patient might be clustered so that the volume of activity justifies the provision of dedicated resources. An example might be patients requiring day surgery. In the UK, the development of dedicated diagnostic and treatment centers will further increase the relevance of patient groups based upon the principles of a focused factory.

Finally, regardless of concerns about the volume of activity and clarity of resource requirements, client concerns about the continuity and coordination of existing services within a care chain might be the main driver for the creation of product groups. Such client concerns tend to be most evident for illnesses with a relatively long duration and/or which require contact with a range of professionals or agencies. Hence, in the UK, National Service Frameworks have been developed that map out the desirable care pathways and services required for patients receiving treatment for conditions such as diabetes and stroke. To some extent, product groupings driven by a desire to promote continuity and coordination mirror developments in clinical protocols and pathways. However, the variety of processes and agencies involved means that planning and controlling the care of patients within such multidisciplinary patient groupings is extremely complicated.

Although the above discussion has outlined a range of product classifications and groupings, it should be noted that there are some process characteristics that have a strong impact on the predictability of resource use by patients within product groups. An awareness of these characteristics is therefore helpful when developing product groups:

- Treatments for well-defined complaints with almost 100 % certainty about the processes required and the outcome (e.g., a bone fracture) should be distinguished from treatments for ill-defined complaints with no routine treatment-path available and no certainty about results. We call these the routine and nonroutine processes (see also Lillrank and Liukko 2004).
- For routine processes, it is possible to define a treatment path, often based on a clinical guideline or protocol, which defines the different operations in the process and their timing. Still the variability in resource use for these routine processes can be quite high due, for example, to practice variations, different modes of treatment, and the consequences of the interaction between doctor and patient. Nevertheless, process patterns can be recognized.
- For non-routine processes, the specialist will proceed in a step-by-step way, checking the patient's reaction on a treatment and deciding on the next step from there. There is no guarantee on the outcome, and there is no in advance layout of the process the patient will follow. Naturally, the predictability of resource use is much lower here than with routine processes.

Of course, these are the extremes on a continuous scale; there is much variation between specialties and within a specialty. However, the variation between specialties is dominant. For a surgical specialty with many protocol patients, such as orthopedics, the number of routine processes may be very high but for a non-surgical specialty, such as internal medicine, it may be much less.

Using iso-process patient groups as business units implies that patient groups are distinguished fulfilling the criteria (homogeneous in terms of process and market performance), that the volume of the patient flow is sufficient to allow for a specific production control, that the trajectories within these groups are described, that production control for each of the patient groups is defined and that the responsibility for managing the patient group is clarified. These steps will be illustrated in the next section with data of an application of such an approach for cardiology.

3 Steps in an OM Approach to Hospital Process Improvement

Based on the logistic concepts discussed in Sect. 2 we have performed a number of case studies with different specialties: cardiology, pulmonology, neurology, general surgery, orthopedics, and dermatology. As the emphasis in these case studies is on process improvement, they can be positioned on the second level of the planning framework discussed in Sect. 2.2. The steps followed are:

- Defining the iso-process patient grouping
- Mapping and analyzing the patient processes

- Defining production control for each patient group
- Setting of objectives and monitoring performance
- Managing the process

These steps result in an organization of patient flows for patient groups that have each a production control fitting the characteristics of the process, and that are run as a business unit. We will illustrate these steps below with data from a case study with cardiology with six cardiologists in a medium sized hospital. They were interested to know how to improve their practice from the perspective of patient groups, and wanted to develop more insight into issues such as:

- How large are the patient flows per patient group?
- What are the average throughput times of patients?
- Do we have enough clinic capacity available for each patient group?

3.1 Iso-process patient grouping

The first step in an OM approach to hospital process improvement is the definition of patient groups fulfilling the criteria of being iso-process. Often these groupings already exist in the practice of clinics, i.e., general clinics and clinics for specific patient groups. Table 20.3 provides information on the main patient groups for cardiology.

The average patient flow on a weekly basis counts 140 patients. Important to notice is that at the main level of patient groups complaint-based labels are chosen as to allow for a good linkage with referring general practitioners or physicians from another specialty. The next step is that within these main groups subgroups of patients are defined that follow a similar trajectory. Trajectories can be distinguished for invasive or noninvasive treatment paths, for simple or more complex problems, etc. Often at this level of the classification of patient groups it is possible to make the link with a medical diagnosis. It is also important to quantify the size of patient flows within patient groups and trajectories to be aware of the number of

Table 20.3 Patient groups for cardiology

Patient group	Average number of patients per week	Percentage of total
1. Chest pain	59.5	42.5
2. Short of breath	35.0	25.0
3. Heart rhythm	24.5	17.5
4. Heart murmurs	4.0	2.9
5. Risk analysis	7.0	5.0
6. Preoperative screening	10.0	7.1
Total	140	100 %

Table 20.4 Patient groups and trajectories for cardiology (based on mean weekly averages)

	Patient groups			Trajectories		
Inflow cardiology (per week)	Name	Cases per week	% of total	Name	Cases per week	% in patient group
140	Chest pain	59.5	43 %	No angina pectoris (ap)	6.0	10 %
				Stable angina pectoris	11.9	20 %
				Unstable ap (outpatient)	11.9	20 %
				Unstable ap (intervention)	3.0	5 %
				Acute coronary syndrome	26.8	45 %
	Short breath	35	25 %	Diagnostic test	7.0	20 %
				Heart failure	14.0	40 %
				Chronic condition	7.0	20 %
				Acute	7.0	20 %
	Heart rhythm	24.5	18 %	Simple	7.4	30 %
				Serious	7.4	30 %
				Chronic	7.4	30 %
				Collapse	2.5	10 %
	Heart murmurs	4	3 %	Valve defect	3.6	90 %
				Valve procedure	0.4	10 %
	Risk analysis	7	5 %	Dyslipidemic	6.7	95 %
				Chronic dyslipidemic	0.4	5 %
	Preoperative screening	10	7 %	General	10.0	100 %
	Heart murmurs	4	3 %	Valve defect	3.6	90 %
				Valve procedure	0.4	10 %
	Risk analysis	7	5 %	Dyslipidemic	6.7	95 %
				Chronic dyslipidemic	0.4	5 %
	Preoperative screening	10	7 %	General	10.0	100 %

patients concerned. Table 20.4 shows the breakdown of the cardiology flow to the patient groups and trajectories.

The different trajectories can also be used for labeling a multidisciplinary collaboration with another specialty or a multi-site collaboration with a more specialized hospital where part of the treatment is performed. Therefore, the defining of patient groups and trajectories has a strategic dimension, as the patient groups and treatments offered need to fit within the strategic profile of the hospital.

The data shown regard mean weekly averages of number of patients flowing in each of these patient groups and distributing over trajectories. A week is chosen as time period of analysis as later on the match needs to be made with resources (outpatient clinic hours and staff, diagnostic department facilities, etc.) available to handle the flow of patients. The data are partly derived from administrative systems, and partly based on expert knowledge of professionals. These expert estimates on distribution of patients over trajectories will be used as a starting point as routine information systems will not be able to deliver these data. However, it is possible to check later on the distribution chosen by small tests with data from routine information systems.

3.2 Mapping and Analyzing Patient Processes

The explicit description and mapping of the process is an extremely important step as this will allow exchanging the different views of medical specialists, nursing staff, clerical staff and patients. It is obvious that medical specialists play an important role in the describing of the process as it regards their core activities. There are more reasons that we resolve to use expert knowledge as the basis for describing processes. First of all, the data we need will not be available in routine information systems as these systems are most of the time based on departmental data collection and have no possibility to provide in a systematic way the information on processes followed by patient groups. Of course, DRG's will be able to provide much of the data but not according to the more aggregated grouping chosen in our description for improving hospital processes. Also the information delivered by DRG's will not be able to provide the time dimension for the process description. DRG's contain the ingredients of the process (numbers of visits, diagnostic and therapeutic procedures, admissions, etc.) but not the process as such. For a description of a health care process we need to be able to follow the steps taken by patients in their journey through the hospital, which includes for instance:

- Time waited by patients for an appointment with a medical specialist (access time),
- Number and timing of visits to outpatient clinics and diagnostic departments in the diagnostic phase,
- Waiting time before admission or before a surgical procedure,
- Diagnostic and therapeutic procedures during admission,
- Number of follow-up visits to outpatient clinics after discharge, and the timing of all these steps.

The second reason to use expert knowledge is to involve the key players in the process of improvement. Often they have to get used to this way of looking at their practice, because it differs from the way they look as medical decision makers to the care process. Once they have appreciated the value of such a non-medical description of the process of the patient, they will be able to use it to improve their practice from the perspective of developing patient centered services.

We have developed a computer model "processor" that allows us to patient flow of a specialty, in close interaction with the professionals involved in the process. The structure of the model—used for mapping the information on processes and resources—is illustrated in Fig. 20.3.

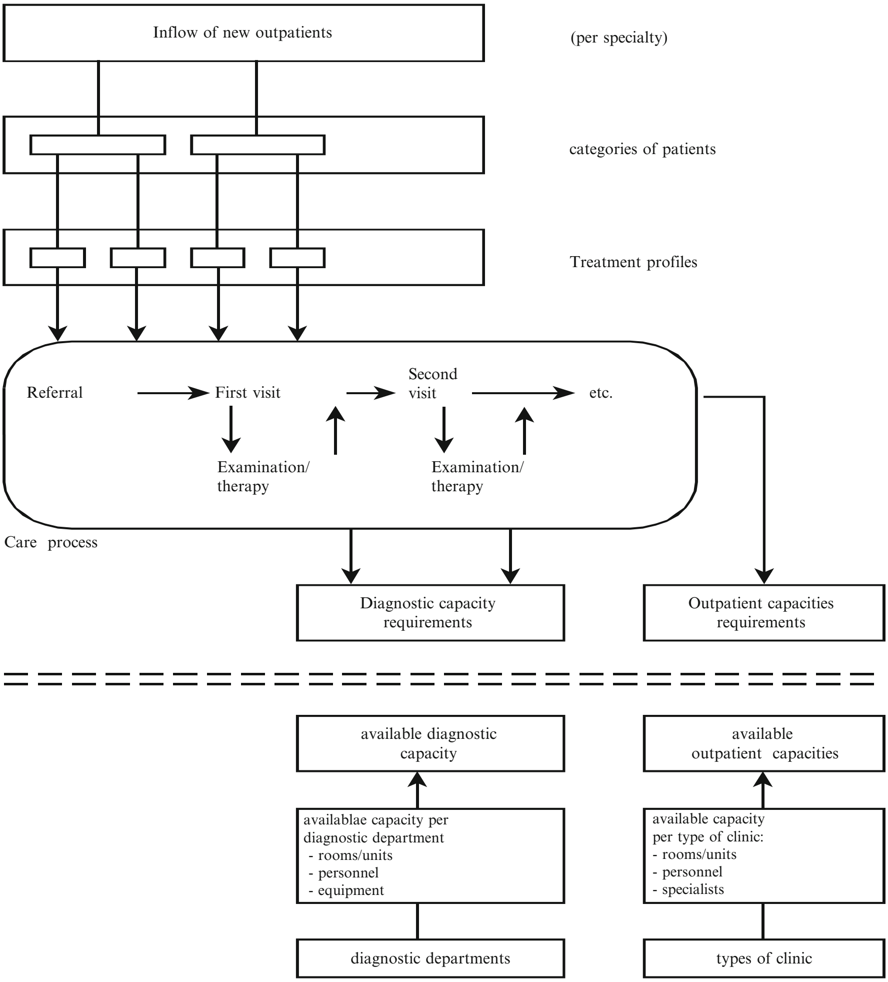


Fig. 20.3 Structure of the demand–supply model for hospital processes

Figure 20.3 shows the demand side of the model (upper part) and the supply side (lower part). The demand side starts with the inflow of new outpatients and the distribution over groups of patients that can be distinguished in the inflow and used for streamlining services and flows. Within these patient groups different treatment profiles are distinguished, representing the modes of treatment available for a patient group and the trajectories followed by patients. At the level of an individual trajectory, the process of the patients from this patient group following this trajectory is described, using input from the medical specialist and information on examinations and treatments. The inflow and distribution of patients over patient groups, combined with the treatment profiles, allows for a calculation of resource requirements in the outpatient department and the diagnostic departments.

Table 20.5 Information on weekly clinic schedules

Clinic	Specialist	Hours	First visits	Duration first visit	Follow-up visits	Duration follow-up visit
General clinics by cardiologists	Specialist 1	15.33	12	20	68	10
	Specialist 2	15.33	12	20	68	10
	Specialist 3	15.33	12	20	68	10
	Specialist 4	15.33	12	20	68	10
	Specialist 5	15.33	12	20	68	10
	Specialist 6	15.33	12	20	68	10
	Total	92	72		408	
Clinics by other professionals	Lipid control	15	3	30	54	15
	Heart failure	29	6	30	107	15
	After care	9	9	60		
	Diagnostic service	6	8	30		
	Pacemaker control	36			56	40
	Total	95	26		217	

The supply side consists of a description of the available capacity in the outpatient and diagnostic departments, in terms of rooms, personnel and equipment. In the results part of the model demand and supply are matched, which allows for visualizing throughput times of processes and calculation of resource utilization.

We will now illustrate how the model is used for our case study setting of a cardiology practice with six cardiologists. The numbers of patients treated in 2001 are: 7,000 first visits (including patients seen in the emergency department), 14,000 follow-up visits, and 1,300 inpatient admissions. Information on the clinics held each week is given in Table 20.5.

Each specialist has five sessions a week in which they see all types of patients. For each session the number of time slots reserved for first and follow-up visits is given as well as the average time available for first and follow-up visits. Apart from the clinics held by the six cardiologists there are a number of clinics held by nurse practitioners and other professionals dealing with patients with chronic conditions. To be able to determine the number of time slots available for each of the patient groups we need to know the mix of patients per clinic. Investigation of the type of patients seen in each clinic provided information on the mix of patients per type of clinic (see Table 20.6).

As can be seen from Table 20.6, the mix of patients in a general clinic reflects the range of patients seen in the cardiology practice, while the other clinics have their dedicated patient group. So the amount of time available for each patient group is indeed not an obvious issue.

We have seen before that the average flow of patients in the practice is 140 each week, with about 85 % of patients in the first three patient groups. Most patients enter cardiology practice via the outpatient department, but a considerable part

Table 20.6 Mix of patients per type of clinic

	Patient group					
Type of clinic	Chest pain	Short of breath	Valve defects	Heart murmurs	Risk analysis	Preoperative screening
General	40 %	22 %	8 %	5 %	5 %	5 %
Lipids control					100 %	
Heart failure		100 %				
After care	100 %					
Diagnostic service		100 %				
Pacemaker control			100 %			

enters also via the emergency department. Though it seems that we concentrate on the use of resources in the outpatient department, we have to take all the inpatient work of the specialists into account. Many people seen in emergency departments, or as inpatients, will continue as outpatients and want to use the time slots shown in Table 20.5. It demonstrates that even if the focus is the outpatient department, the complete workload of the cardiologists must be understood because of the links between the different strands of work and activity generated in OP departments. The six main patient groups are broken down into a number of subgroups that sometimes refer to a specific complaint or sometimes to a different treatment path. At this level a link can be made to the process of the patient. The processes of all these patient subgroups are described based on expert opinion from cardiologists. This information is structured with the help of the computer model that also returns the information to the specialists in the format of a process chart. Figure 20.4 shows one of these processes, i.e., a stable angina pectoris patient.

Figure 20.4 shows that the process for patients with stable angina pectoris consists of on average eight steps. All steps take place at the outpatient department. In the first step some diagnostic tests are performed (ECG, upward and downward arrow) and some tests require the patient to return at another occasion (e.g., ergometrics). After the first step everybody will return for a follow-up visit at the outpatient department in 30 days time (not shown in this figure, but see Table 20.7). With step two 33 % of all patients seen are discharged. In step three again 10 % are discharged and 5 % of patients have—depending on the results of tests—a chance to be admitted immediately. The rest of the patients are following a trajectory with annual follow-up visits, with a chance of being discharged, admitted, or referred for a PTCA. The information that is used to describe the different steps in the process is summarized in Table 20.7.

For each step in the process information is given on the content of the step (in this case, first and follow-up visits, but it could also regard an admission or a procedure), the percentage of patients with diagnostics that are performed instantly (as a direct follow-up of the encounter with the specialist) or with an appointment on another occasion, the next step in the flow (percentage of patients that will return for a next step, or require immediate admission, or continue in another profile, or are discharged), and the reappointment interval for patients that return.

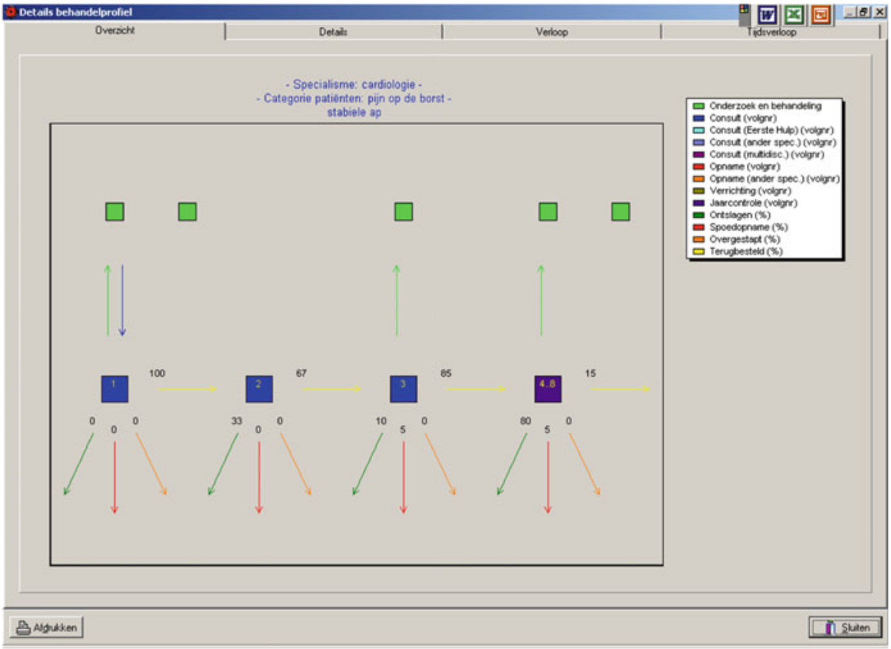


Fig. 20.4 Graphical illustration of care process: "stable angina pectoris" patient

The diminishing patient flow by the chance of discharge in each step can be visualized by the model's output (see Fig. 20.5). The model also provides insight into the match between available and required slots in outpatient clinics per patient group (see Table 20.8).

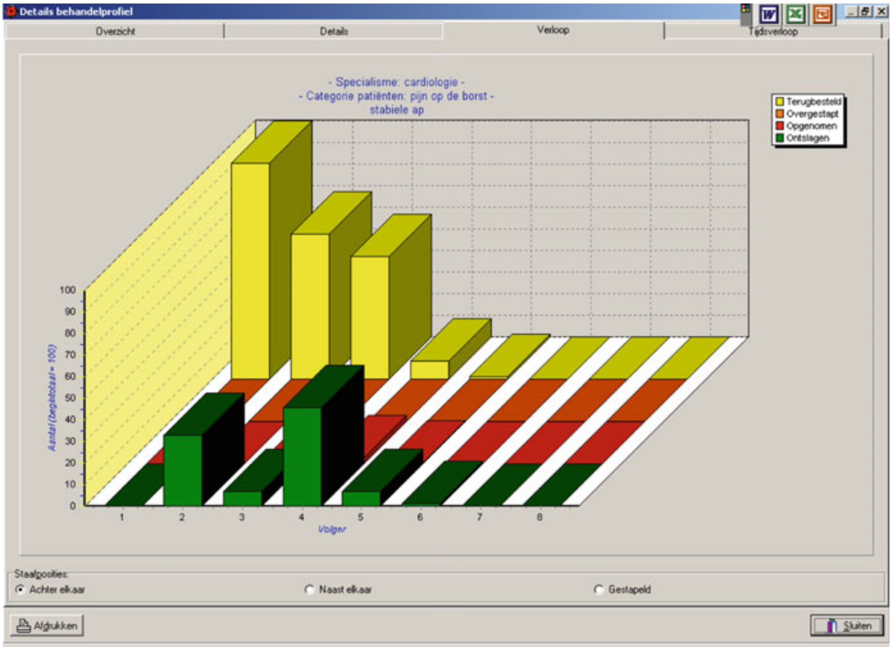
Table 20.8 illustrates that the match at the level of the total of clinics is inadequate (too few slots for first visits, too many slots for follow-up visits). Also the match at the level of individual patient groups can be improved. Another type of output is the throughput times of processes. Table 20.9 provides information on the number of activities per type and on the throughput times to complete all the steps in the process. For patient groups with a chronic condition, the process has been artificially cut off. The throughput times can also be visualized, as shown for the "chest pain" patient group in Fig. 20.6. The difference between the current throughput time and the "minimum" throughput time shows the gain to be made by better planning of patients in relation to the diagnostic tests.

3.3 Defining Production Control

We now have insight into the different patient groups and trajectories, the size of the patient flows, the process of patients and their resource requirements. The next

Table 20.7 Summary of steps in the care process of stable angina pectoris patients

	Patient Group: chest pain			Trajectory: stable angina pectoris				
		Diagnostic/therapy		How data				
				Percentages				
Step	Content	Direct	Appointment	Continue	Urgent admission	Other trajectory	Discharge	rcapp. interval
1	First visit	ECG (100 %)	Ergometrics (100 %) lab tests (100 %) CAG (10 %) thallium test (10 %)	100 %				30
2	Follow-up visit			67 %			33 %	90
3	Follow-up visit	ECG (100 %) lab (100 %)		85 %	5 %		10 %	180
4–8	Follow-up visit (5×)	ECG (100 %) lab (100 %)	PTCA (5 %)	15 %	5 %		80 %	360



Staafposities:
● Achter elkaar ○ Naast elkaar ○ Gestapeld

Fig. 20.5 Graphical illustration of the diminishing patient flow in the care process "stable angina pectoris" patients

Table 20.8 Match between demand for and supply of time slots in cardiology clinics

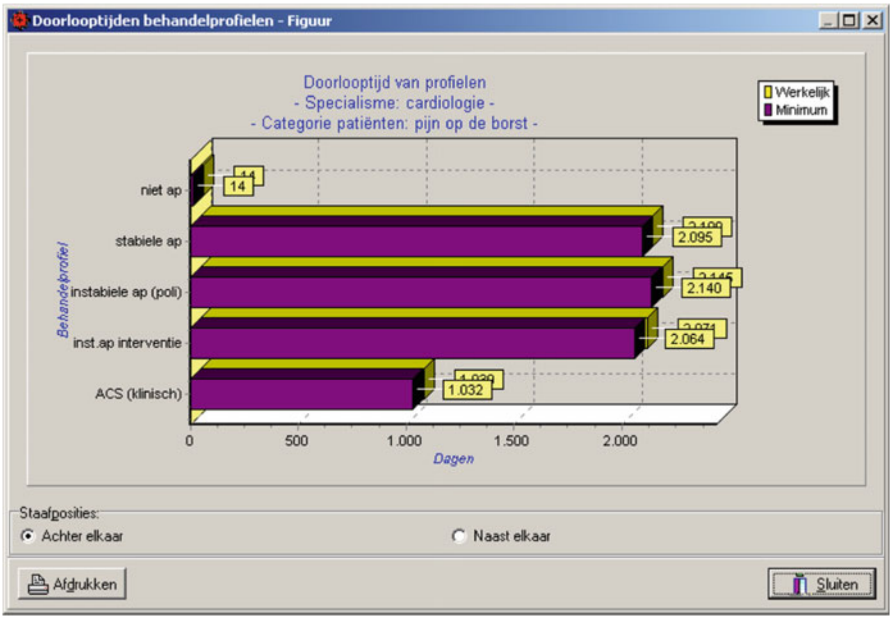
	First visits			Follow-up visits		
Patient group	Demand	Supply	Occ.	Demand	Supply	Occ.
Chest pain	33	38	87 %	117	163	72 %
Short of breath	28	30	94 %	243	197	123 %
Valve defects	16	14	114 %	63	89	71 %
Heart murmurs	4	4	100 %	12	20	56 %
Risk analysis	7	7	100 %	28	74	38 %
Preoperative screening	6	4	150 %	10	20	49 %
Total	98	87	112 %	475	564	84 %

step in the OM approach to process improvement is to define production control for each patient group. To evaluate which type of control fits best with the process looked at, we need to know the characteristics of each of the processes. An evaluation of these characteristics is provided in Table 20.10. For this purpose some of the trajectories were combined, such as angina pectoris (ap) and heart failure (hf).

The characteristics considered are: number of steps in the process (whether it regards a short or a long process), the complexity of the process (due to diagnostic tests performed or due to consultation with other specialties), the chronicity of the

Table 20.9 Match between demand for and supply of time slots in cardiology clinics

		Visits cardiologists						
Patient group	Trajectory	First	Follow-up	Annual visits	Nurse visits	Admit	Admit other hospital	Throughput time
Chest pain	No ap	1	1					14
	Stable ap	1	2	5				2,100
	Unstable ap (outpat.)	1	4	5				2,145
	Instable ap (intervention)	1	3	5			1	2,071
	Acute coronary syndrome	1	6			1		1,039
Short of breath	Diagnostic service	1	5		1			1
	Heart failure	1	5		12			467
	Chronic	1	8					610
	Acute	1	3			1		312
Valve defects	Simple	1	6		8	1		591
	Serious	1	6			1		769
	Chronic	1	9		7	1		1,865
	Collapse	1	1	4	9	1	1	1,630
Heart murmur	Valve defect	1	4					1,454
	Valve procedure	1	5				1	862
Risk analysis	Dyslipidemic	1			6			350
	Chronic dyslipidemic	1			11			1,970
Preoperative screening	General	1	1					14



Staafposities:
● Achter elkaar ○ Naast elkaar

Fig. 20.6 Graphical illustration of the throughput times for "chest pain" patients

process (whether it regards a chronic process with no defined end), the predictability of the process (regarding number of steps, duration of steps and the routing of steps), and the use of resources in the process (whether use is made of a shared resource or a bottleneck resource). For instance the patient group "chest pain with ap" is characterized by a long process that is complex and chronic, with not good predictability of the process on duration and routing, and with use of a resource that is shared with other specialties (ergometric test).

We now can use these characteristics to define a way of production control that fits the process considered. Though the best fit is a matter of tailor-made decision making for a specific patient group, the rules used to derive to a solution are:

- When predictability is high, and the process is not too long or complex, and there are no critical resources involved, the process can be scheduled with a longer planning horizon;
- When predictability is high, but the process is long or complex, and critical resources are involved, scheduling can be per phase, i.e., diagnosis, therapy, aftercare;
- When predictability is low, scheduling can only be performed on a short planning horizon, i.e., per step of the process.

For the example of the patient group "chest pain," we would suggest a production control per phase of the process, i.e., scheduling all steps in phase in advance.

Table 20.10 Characteristics of patient groups

	Patient group									
	Chest pain		Short of breath		Heart rhythm		Heart murmurs			
Characteristics	No ap	ap	No hf	hf	Not relevant	Relevant	Not relevant	Relevant	Risk analysis	Pre operative screening
Number of steps	−−	+	−	+	−−	0	−−	+	−	−
Complexity	−−	+	−	+	−−	0	−−	0	−−	−−
Chronic	−−	+	−	+	−−	+	−−	+	−	−
Predictability										
No of steps	++	+	++	+	++	+	++	+	++	++
Duration	++	0	++	+	++	+	++	0	++	++
Routing	++	0	++		++	+	++	+	++	++
Use of resources										
Shared	−−	0	++	++	−	0	0	0	0	0
Bottleneck	+	+	++	++	0	++	−	0	−	++

−− not relevant at all, − not relevant, 0 hardly relevant, + relevant, ++ very relevant

This will create optimal conditions for creating the best service for the client (by combining visits and diagnostic tests) and advance scheduling of resources.

3.4 Objective Setting and Monitoring

Having defined the production control setting for each patient group we now have to set objectives for the performance of patient groups. We have distinguished four areas of performance:

- Quality: conformance to specification
- Delivery speed: time between the moment of demand and the moment of delivery
- Delivery reliability: degree of meeting the arrangements agreed
- Flexibility: degree of coping with unpredictable situations

These areas of performance are further operationalized with measures that are specified in Table 20.11. Tables 20.12, 20.13, 20.14, and 20.15 provide information on the scoring of these measures per patient group in the areas of the objectives. The scoring was based on the evaluation of these measures by the different cardiologists.

The overview of the performance on "quality" measures in Table 20.12 provides insight into areas where improvement is required. Results that were evaluated as "0" or lower were regarded as indicators for quality that needs to be looked at. For the example of the patient group "chest pain," action is required in these areas:

- Consistency between cardiologists in handling patients with angina pectoris: protocol for treatment of angina pectoris patients was discussed within the group of cardiologists to deal with the differences in treatment between them.
- Instruction and expertise: instruction material for guidance of angina pectoris patients improved and better qualified nursing staff were selected to deal with these problems.
- Inflow and outflow: problems in these areas required attention for the links with general practitioners. Improvement in the referral process and in the phase of aftercare could be realized by putting this patient group on the agenda for discussion in meetings with general practitioners.
- Tuning: Improvement of the tuning with other specialists dealing with the same patient group required discussion with other specialties on the multidisciplinary treatment of patients.

Table 20.13 also provides information on the scoring of "delivery speed" performance by patient groups.

From the information in Table 20.13 one can see where action is required to improve the performance on the speed of delivery of cardiology services to patients. For the example of "chest pain" patients timely access for different urgency levels was OK, but the time required for a timely diagnosis was clearly indicated as an

Table 20.11 Objectives and measures of performance (based on De Vries et al. 2004)

Objective	Measure	Description
1. Quality	Right at one go	No rework due to complications
	Consistency	No difference between different providers
	Adequacy	Patients are treated adequately
	Instruction	Patients receive adequate instruction
	Expertise	Expertise of professionals is up to standard
	Inflow	Correct referral and referral information
	Outflow	Continuity of care with follow-up providers
	Tuning	Tuning with other professionals in multidisciplinary treatment
	Patient satisfaction	Satisfaction of patient with treatment
2. Delivery speed	Access time	Waiting time for patients to get access to cardiology services, for different levels of urgency
	Waiting list	Waiting time on waiting list in case of elective admission
	Wait time in clinic	Waiting time in the outpatient department
	Throughput time	Lead time required for diagnosis, treatment and aftercare
	Consultation	Availability of consultation by other specialists
	Diagnostics	Availability of diagnostic facilities
3. Delivery reliability	Canceling	Level of cancelation of appointment due to
	Appointments	Rescheduling of clinics
	Canceling	Canceling of admissions due to, for instance, not
	Admissions	Availability of beds
	Output arrangements	Levels of output required to fulfill contracts with purchasers
4. Flexibility	Service	Time required to introduce a new type of service
	Mix	Dealing with shifts in the mix of patient groups
	Volume	Dealing with changes in the amount of patients in a patient group
	Delivery speed	Dealing with changes in the arrangement for access of patient groups

area where improvement was required. The low score on item 6 indicated that this was due to the availability of diagnostic facilities. Also the throughput time in the phase of treatment and aftercare could be improved.

Table 20.14 provides information on the scoring of "delivery reliability" performance by patient groups, indicating where improvement is required in the reliability of the delivery of services, in terms of canceling appointment or admissions, and in terms of meeting the contract arrangements with the purchasers of hospital services, such as health insurance organizations. For "chest pain" patients, attention was required for the fact that when clinics had to be canceled due to non availability of a specialist, appointments for "chest pain" patients had to be rescheduled at a short notice.

Table 20.15 provides information on the scoring of "flexibility" performance by patient groups, giving insight into the performance of the organization of cardiology services to deal with unpredictable changes. For the patient group "chest pain,"

Table 20.12 Objective "quality": measures of performance per patient group

	Chest pain		Short of breath		Heart rhythm		Heart murmurs			
Items	No ap	Ap	No hi	hf	Not relevant	Relevant	Not relevant	Relevant	Risk analysis	Preoperative screening
1. Right at one go	++	+	+	+	+	+	++	++	++	++
	+	–	+	0	+	+	+	+	+	++
2. Consistency	++	+	++	+	++	+	++	++	+	++
3. Adequacy	0	0	0	0	0	+	+	+	+	+
4. Instruction	0	0	0	+	0	0	+	+	+	+
5. Expertise	0	0	0	0	0	0	0	0	+	+
6. Inflow	+	0	0	0	+	+	+	+	+	+
7. Outflow	na	0	na	0	na	na	+	+	0	––
8. Tuning										
9. Patient satisfaction	++	+	0	0	0	+	++	++	+	0

–– very bad, – bad, 0 not good, + good, ++ very good; na not applicable

Table 20.13 Objective “delivery speed”: measures of performance per patient group

	Chest pain		Short of breath		Heart rhythm		Heart murmurs			
Items	No ap	ap	No hf	hf	Not relevant	Relevant	Not relevant	Relevant	Rise analysis	Preoperative screening
1. Access time:										
Acute	++	++	++	++	na	++	++	++	+	++
Urgent	++	++	+	+	na	++	+	+	+	++
(Semi-) elective	++	++	0	0	+	+	+	+	+	++
2. Waiting list	na	+	na	na	na	–	na	+	+	na
3. Clinic wait	+	+	0	0	+	+	+	+	+	+
4. Throughput time:										
Diagnosis	––	––	–	–	––	–	+	0	0	+
Treatment	na	0	na	–	na	–	na	+	+	+
After care	na	–	na	+	na	+	++	+	+	++
5. Consultation	na	+	na	+	na	na	+	+	+	++
6. Diagnostics	0		0		0	0	+	+	+	+

–– very bad, – bad, 0 not good, + good, ++ very good; na not applicable

Table 20.14 Objective “delivery reliability”: measures of performance per patient group

	Chest pain		Short of breath		Heart rhythm		Heart murmurs			
Items	No ap	ap	No hf	Hf	Not relevant	Relevant	Not relevant	Relevant	Risk Analysis	Preoperative screening
1. Canceling appointments	–	–	na	0	+	+	+	+	0	0
2. Canceling admissions	na	++	na	++	na	++	na	–	na	na
3. Output volume										
Visits	+	+	+	+	+	+	+	+	+	+
Admissions	+	+	+	+	+	+	na	+	+	na

–– very bad, – bad, 0 not good, + good, ++ very good; na not applicable

Table 20.15 Objective “flexibility”: measures of performance per patient group

	Chest pain		Short of breath		Heart rhythm		Heart murmurs			
Items	No ap	ap	No hf	hf	Not relevant	Relevant	Not relevant	Relevant	Risk analysis	Preoperative screening
1. Service	+	+	+	+	0	0	+	+	+	+
2. Mix patients	+	+	0	0	+	+	+	+	+	+
3. Volume	+	+	–	–	+	+	+	+	+	++
4. Delivery time	+	+	–	–	+	+	+	+	0	+

–– very bad, – bad, 0 not good, + good, ++ very good

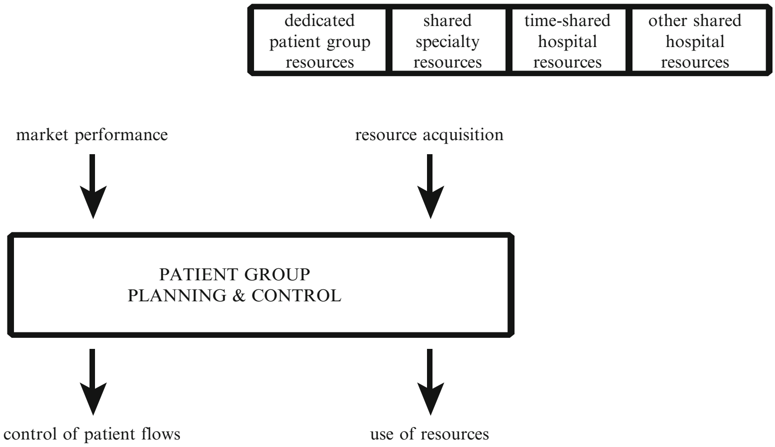

dedicated patient group resources
shared specialty resources
time-shared hospital resources
other shared hospital resources
market performance
resource acquisition
PATIENT GROUP PLANNING & CONTROL
control of patient flows
use of resources

Fig. 20.7 Responsibilities of patient group management

no areas were found to require attention. The low scores on “volume” and “delivery time” flexibility for “heart failure” patients referred to the capacity of diagnostic facilities to deal with unexpected larger patient flows or increased delivery speed.

3.5 Process Management

We have shown in Sect. 3.4 that a number of indicators need to be monitored per patient group to assure that the performance of cardiology services is up to the agreed standards. It is very important that someone is responsible for and interested in the performance of processes. Unless the responsibility for process management is clearly defined, process improvements will not be sustained. Therefore it is necessary to discuss who will be responsible for process management in hospitals.

Process management involves a number of responsibilities. Most of them can be derived from the function process management fulfills in the framework for production control discussed in Sect. 2.2. Process management was discussed here under the heading of patient group planning and control. At the patient group level there must be a management function responsible for defining the market performance to be delivered (patient mix, urgency criteria, acceptable waiting times, etc.), the acquisition of the resources required for the patient group, the control of the patient flow and the utilization of resources by the patient group. The different tasks of patient group management in terms of planning and control are illustrated in Fig. 20.7.

The detailed control at the level of the patient group involves checking that the day-to-day scheduling of patients is in line with the service requirements specified for the patient group as a whole and the regulations on resource use imposed on the patient group. We will concentrate first in Fig. 20.7 on the resource acquisition and resource use regulations for the patient group. Some of these resources may be dedicated for use by the patient group only, some may be made available from the resources that are dedicated at the level of a specialty but shared by the patient groups served by the specialty, others must be contracted for specified periods as these are shared between specialties, and the remaining resources are supposed to be generally available for different patient groups. The four categories of resources distinguished in Fig. 20.7 show a decreasing extent of influence that the management of the patient group can perform on their availability for the patient group.

In summary, the responsibilities of process management come down to defining the market performance to be delivered, acquiring the resources required for the patient group, controlling the flow of patients to assure that the process of individual patients fits within the bandwidth defined for the patient group considered, and controlling the efficient use of resources by the patient group.

Process management is clearly a managerial responsibility. Nevertheless, we would advocate a strong involvement of medical specialists in the running of patient groups. Ideally, a small management team should be defined for a patient group, consisting of a medical specialist, and for example possibly a nurse practitioner or other key professionals, supported by a manager. The medical specialist should be acting as the chairperson of the management team, with responsibilities for defining, monitoring and evaluating programs offered to patient groups. A nurse practitioner could assist in collecting the data required for the monitoring of the performance of the patient group. A manager could support the team by organizing the resources required for the patient group and by providing information on the use of resources. By putting the medical specialist in the position of chair of the management team for a patient group, we strive to make him/her the owner of the process organized for the patient group. Furthermore, running a patient group would also imply that he/she takes the lead in discussing medical practice with colleagues, if colleagues are performing outside the bandwidth agreed for the patient group.

We believe that this task can be very well performed in combination with clinical work, and would not require much time from medical specialists, as they receive many signals during each day if processes are running smoothly or not. Process management by medical specialists is a form of management participation by specialists that is much closer to clinical practice than the (co) responsibility of running a department. After all, professionals in health care want to spend more time on patients and less on managerial matters. The form of process management we are looking for will help them to have more influence on the way hospitals are organized to meet the demand of patients.

4 Conclusions and Future Work

We have discussed in this chapter an operations management approach to hospital process improvement. First we have argued that a more fundamental approach for hospital process improvement should be used to make improvements sustainable and part of a new routine to be developed in hospitals, in which the process of patients is the basis for organizing hospitals. This is what operations management is about. We have introduced a few concepts that can be used to create such an operations management context for looking at hospital process improvement. The first concept is the distinction between unit and chain logistics and its combination, i.e., network logistics. We also need to include unit logistics and chain logistics in our approach, because we want to avoid sub-optimization by only looking at an isolated process, and we want to account for the impacts of process improvements on the efficient use of resources. The second concept is a framework for production control of hospitals in which decisions regarding patient flows and resources are organized at different levels of planning. This will help to position process improvements in a wider context of production control of hospitals, and make those involved aware of the interactions between the different parts of the hospitals and the creation of conditions at higher levels of planning to make process improvement sustainable. The third concept is the definition of iso-process patient groups, based on the trajectory of the patient through the system and the use of similar constellation of resources. These patient groups can then be considered as business units that require a production control system fitting the characteristics of the process of the patient group. This will help to identify patient groups with a very predictable process that can be organized as a focused factory, and other patient groups with a less predictable process, which can be planned per phase of the process (diagnosis, therapy, aftercare) or at a more aggregate level of a program for patient groups.

The operations management approach to process improvement was then applied to the cardiology patient flow in hospitals. The different steps in the approach were elaborated and illustrated with cardiology patient flow examples. The first step was to identify iso-process patient groups, for which a specific organization of services is developed. Six patient groups were identified. Together with the different trajectories followed by patients within patient groups, a total of 18 processes were sufficient to describe the patient flow of cardiology. The second step was to describe these processes in a way that allows analysis of the service and resource use impacts of processes. We used a computer model developed for this purpose, which uses as a basis the expert knowledge of professionals on the characteristics of processes and the way processes are organized. The computer model helps to make the description more consistent and allows analysis of processes to improve the insight into specialty practice. The third step was to define a production control per patient group, taking into account the characteristics of the process considered. Characteristics taken into account are whether it regards a short or a long process, whether the process is complex due to diagnostic tests performed or due to consultation with other specialties, whether it regards a chronic process with no defined end, whether the process is predictable and whether the process makes use of a shared resource or a bottleneck resource. The fourth step involves the setting of objectives for the performance of the process to enable its monitoring. Objectives were defined for the quality of the service, the speed and reliability of delivery and the flexibility of the organization to meet changes of practice, such as an introduction of a new technology, a shift in the mix of patients, an increased volume of patients or a shorter access time for a patient group. The fifth step regarded the responsibility for process management in hospitals. We argued that medical specialists need to take up the responsibility for process management in order to make improvements in hospital processes sustainable.

Though we believe that we have come far in developing a comprehensive approach to hospital process improvement, a number of issues need to be looked at more fundamentally. We will list a few of these issues below.

The first issue is to improve the method of patient grouping based on the iso-process concept. This would require investigation of the variations in the process of a patient group, as well as the type of variation (more or less steps, longer or shorter duration of steps, variation in routings of steps), and identify what part of the variation has a structural cause and what part of variation is due to stochastic behavior of variables in the system. This will provide insight into the level of standardization that is possible in hospital business processes.

The second issue is to look more fundamentally at the characteristics of processes and to their importance in finding the best fit for production control of a patient group. Is predictability of the process the most important factor, and what role do the other characteristics play?

The third issue is how to construct from the massive data on patient encounters reliable process descriptions for patient groups. Can process mining, in which logs of data of patient encounters are used to construct process descriptions, help?

The fourth issue is how to define band widths for the performance of patient groups on the objectives for process management.

These are a few fundamental questions that can be put forward to provide better support for the steps taken in our operations management approach to hospital process improvement. But, above all, we need more experience in applying this approach to develop a hospital organization that puts patients in the center by organizing processes in an effective and efficient way.

References

Baas, L. J. C. (1996). Producttypering medisch-specialistische ziekenhuiszorg (Dutch). *Medisch Contact, 51*, 356–358.

Bertrand, J. W. M., Wortmann, J. C., & Wijngaard, J. (1990). *Production control: A structural and design oriented approach*. Amsterdam: Elsevier.

de Vries, G., Bertrand, J. W. M., & Vissers, J. M. H. (1999). Design requirements for health care production control systems. *Production Planning and Control, 10*(6), 559–569.

de Vries, G. G., Buwalda, P., & van der Linde, M. (2004). Besturing van patientengroepen in een zorginstelling. *Kwaliteitin Beeld, 2*, 8–11.

Fetter, R. B. (1983). *The new ICD-9-CM diagnosis-related groups classification scheme*. HCFA Pub. No. 03167. Health Care Financing Administration, Washington: U.S. Government Printing Office.

Fetter, R. B., Averill, R. F., Lichtenstein, J. L., & Freeman, J.L. (1984). Ambulatory visit groups: A framework for measuring productivity in ambulatory care. *Health Services Research, 19*, 415–437.

Fetter, R. B., & Freeman, J. L. (1986). Diagnosis related groups: Product line management within hospitals. *Academy of Management Review, 11*, 41–54.

Lillrank, P., & Liukko, M. (2004). Standard, routine and non-routine processes in health care. *International Journal of Quality Assurance, 17*(1), 39–46.

Vissers, J., & Beech, R. (Eds.). (2005). *Health operations management. Patient flow logistics in health care*. London: Routledge.

Vissers, J. M. H., Bertrand, J. W. M., & de Vries, G. (2001). A framework for production control in health care organisations. *Production Planning and Control, 12*(6), 591–604.

Chapter 21
Managing a Patient Flow Improvement Project

David Belson

Abstract There are many opportunities for improvements in patient flow, but a successful project to do so requires proper organization and execution. The professional project management tools used in many industries are relevant, but the hospital environment places special demands that affect chances for success. A patient flow improvement project must have a well-understood scope and objective with measurable results. The project team must be multidisciplinary, and participation from staff members must be encouraged. Examples and specific relevant project management methods are discussed.

Keywords Project Management • Patient flow • Change

1 Introduction

In their attempt to reduce patient flow delays hospitals and other health care institutions invariably implement a project as a way to accomplish the desired change. Projects are defined as unique and one-time endeavors. They can be as large as a move to a new facility or as small as the change in a current procedure. (See PMI, 2013 and Gray and Larson 2011.) Various situations may impel a project—such as the need to replace an information system or the necessity to change current practices. Executives frequently request their managers to undertake a “project” to correct some problem or to facilitate some objective and managers often assign their staff to undertake various projects. Thus, hospitals often have many projects underway at any point in time. However, the managers and staff involved may not be knowledgeable as to what it takes to assure a project’s success

D. Belson (✉)
Epstein Department of Industrial and Systems Engineering, University of Southern California,
Los Angeles, CA 90089-0193, USA
e-mail: dbelson@thegrid.net

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3_21, © Springer Science+Business Media New York 2013

and sustaining the results. Some larger institutions have a Project Management Department that is responsible for projects.

Projects are also viewed as an opportunity. Staff may use them as a career opportunity to exhibit their skills to a range of people inside and outside the hospital. Projects can be a way to involve oneself in interesting new trends or to learn new skills or merely to socialize. Thus, projects may be an opportunity for growth and advancement for the individual.

Some projects result from new problems or crises. Few successful changes take place in a health care institution without a so-called project to organize the effort and to marshal the resources necessary to accomplish a goal that crosses organizational lines.

Projects go through a series of stages: approval and start-up, planning, execution, and termination. The start-up stage may be long and includes development of the objectives and analysis of the situation. The planning phase is critical to success and is the subject of most of this chapter. The execution phase uses the most resources in executing the elements of the project. The termination phase is relatively brief and its role is often misunderstood.

Although projects are a commonplace activity in a hospital, few health care professionals have training in project management. This is unfortunate since there is clear body of knowledge regarding how to plan and execute a successful project. Project Management is one of the few fields where there is a comprehensive and universally agreed upon body of knowledge (Project Management Institute 2013).

2 Project Management Methods in Health Care

A project is a series of work tasks that have a definite beginning and an end and leads to an outcome (Verzuh 2011). A successful project requires a clear plan with certain necessary components. The science of project management has existed for many years and the vast number of projects using these concepts has solidified the understanding of what is needed. Some the current project management tools were developed in the 1950s for large military and aerospace efforts (Kerzner 2004). Many aspects of projects to improve patient flow are similar to projects in other industries, which have projects involving changing work practices or procedures. Projects in health care generally require a multidisciplinary team and a focus on effective communication among the members.

First among the requirements for a project management plan is the definition of a project's objective and its scope. A project objective is a concise description of what the project hopes to achieve. If possible, the objective is defined in measurable terms, such as a specific and quantifiable change in the average patient discharge time or an increase in the number of patients served per day in a particular department. Without a written objective, which can be measured after the project is complete—there is little chance for success.

A project's scope is a definition of the elements that are included in a project as well as what is not included. Defining what is not included is important because some people may assume that an aspect is included when it is not. For example, a project to improve the patient flow through a radiology department may include speeding up the time patients spend at scanning equipment but *not* include speeding up the processing of scanning output data. By defining what is not included, as well as what is included, a project is less likely to disappoint the recipients of the project's results or disappoint the organization that is paying for the project.

	Example
Objective	The project will increase the total number of patients served per day in the hospital's CT function without increasing staffing costs or the number of machines in the unit. Ideally, the objective would be converted into a quantitative goal, such as scans completed per day.
Scope	The project will redesign the work procedures in the department based on an analysis of the current patient flow and identified bottlenecks. Accelerating the handling of the patient scanning results is not part of this project.

2.1 Project Management Concepts for Patient Flow

Project Management, as it applies to the health care industry, can rely on the extensive body of knowledge regarding project management generally. PMBOK describes the area of expertise as involving managing the following aspects of projects:

1. Integration
2. Scope
3. Time
4. Cost
5. Quality
6. Human resources
7. Communications
8. Risk
9. Procurement

In health care, each of these nine knowledge areas has a particular meaning and places specialized demands on the project manager.

1. *Integration*—it is necessary to coordinate the various elements of a project. In the case of a project to improve patient flow this can mean the need to coordinate an information system change to provide necessary reports while at the same time coordinate changes to clinical and administrative data entry procedures by the staff. The different organizational groups participating in a project may have individual priorities, which must be accommodated as a project progresses. For

example, nursing may focus on patient convenience while administration may be more focused on costs and economic margins. The project manager must have the ability to deal with many different elements of a project, sometimes on a simultaneous basis.

2. *Scope*—often projects are ambitious. Hospitals wish to improve care and improve productivity, but often with limited resources. Thus, the challenges of the project must be clearly defined and the impact of potential changes well understood. Perhaps one of the most common causes of project failure is successive expansion to the initial project scope. Project team members or the project's client may want to accomplish more than was originally envisioned. "Project Creep" is a common ailment. An effort to reduce outpatient's waiting times might expand into building construction, information systems and outsourcing. If a scope change is deemed necessary, the project management should document the change and the related changes to the project's budget and time schedule as well as approval of the change. A formalized scope change control is necessary and a documentation of any changes is needed.
3. *Time*—the timely completion of a project is often an organization's primary measure of success. To assure a project's progress and meet time expectations, a number of items are necessary:

    (a) Task definition
    (b) Task duration estimation and extent of variability
    (c) Task sequence requirements
    (d) Scheduling based on task time, sequence and resource requirements
    (e) Schedule control managing the plan and progress

    If these are properly managed, then the project can be completed on time.
4. *Cost*—involves processes to assure that a project is completed within the planned budget. Sometimes hospitals will undertake a project without giving the project an explicit budget. This may seem to make the job of the project manager easier, but it may rob the hospital of a necessary control over its scarce resources. If a project is sufficiently large, the organization may institute project accounting whereby people working on the project report their time by task and project accomplishments are compared to costs.
5. *Quality*—involves making sure that a project meets the purpose for which it was undertaken. Health care places a high priority on quality and the same attitude should prevail in the execution of a project. Moreover, quality of care may be impacted by the changes generated by a project and there must be project resources allocated to verifying that no adverse impact has taken place.
6. *Human resources*—involves making effective use of the people involved in a project. Ad hoc groups are often drafted to do projects in health care. The job of the project manager is to develop this team and focus its efforts. Team members must be motivated to act together and to rely on each other within the context of the project's scope. This may require so-called team building exercises to create a feeling of cohesion and a willingness to extend effort in support of the project.

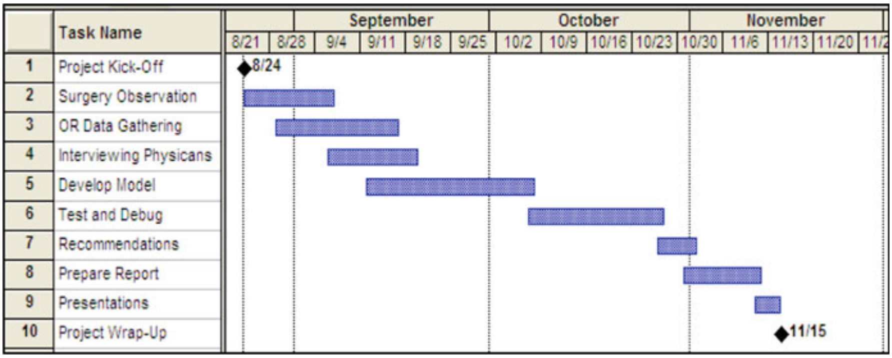
7. *Communications*—involves the distribution of information regarding the project. Since health care organizations involve many different specialties and work groups, this can be a challenge. Differences in terminology, work schedules and reporting relationships are barriers. Moreover, regulatory authorities may have an interest in a project and information they need must be managed.
8. *Risk*—involves identifying and planning how to respond to project risks. The possibility of failure exists for patient flow projects. Tools are available to identify, quantify and respond to project risk.
9. *Procurement*—goods and services sometimes are required, as part of a project and the project must manage these as well as the use of internal resources.

The Project Management Institute has long been a source for such knowledge. It publishes a body of knowledge book (PMBOK) in order to clarify the scope and important ideas of the field. It also administers a widely recognized professional certification program. It supports a special interest group (SIG) for health care. It includes subgroups interested in projects involving information systems, regulatory obligations, business process re-engineering, and other areas.

Project managers should be aware that all projects involve conflicting interests. The so-called "triple constraint" of project management is the desire to minimize a project's costs, maximize results and accelerate the time taken to do the project. These three aspects tend to work in opposition to each other (MacGregor 2005). For example, a project to conduct a patient flow analysis may want to minimize data collection in order to reduce its costs. However, better results may result from more data collection and more time taken to do the work. The best a project manager can do is to balance the achievement in each of the three constraints in regard to the expectations of the organization. A budget to approve the costs, an agreed upon timeline and a clear project objective and scope are the best way to deal with the problem of the triple constraint.

2.2 Project Management Tools

Projects can be made more effective through the use of the various software products made available to assist the project manager. Perhaps the most basic tool for the project manager is to develop a work plan using techniques such as a GANTT Chart or a PERT Diagram that provide a graphical picture of the project. These help clarify the relationships between activities and the relationship between activities and time. Scheduling a large project may prove impossible without such tools. They were, in fact developed specifically for large projects such as the Apollo space project and defense industry projects. Many computer software products are available for developing project plans by automating the charting and scheduling tasks. The most popular ones include the following:



	Task Name
1	Project Kick-Off
2	Surgery Observation
3	OR Data Gathering
4	Interviewing Physicans
5	Develop Model
6	Test and Debug
7	Recommendations
8	Prepare Report
9	Presentations
10	Project Wrap-Up

Fig. 21.1 Typical project Gantt chart

1. *Microsoft Project*, the project management software with largest market share
2. *Open Workbench* an open source alternative to MS Project, formerly Project Workbench
3. *Primavera Project Manager*, popular for large scale projects such as construction and public works and SureTrak® Project Manager from Primavera for smaller projects
4. *Artemis*
5. *FastTrack Schedule*
6. *CA Super Project*

Such software facilitates revising the project schedule as the project progresses and can help facilitate team communications by sharing project plans electronically.

The Gantt Chart or Bar Chart (Fig. 21.1) was first popularized by H. L. Gantt, an Industrial Engineer, during World War I, and it is useful for planning and describing a time schedule. Projects are broken down into activities and the duration of each of these is displayed against a timeline. The meaning of the resulting diagram is self-evident. Additional information can be displayed in the diagram to present task-by-task progress or the interrelationships between activities.

As projects become larger, it is necessary to do some "decomposition"—divide up the work into segments. Larger projects may involve a work breakdown structure (WBS), numbering of tasks for accounting purposes. The numbered tasks are then used for the reporting of time spent and determining the extent to which planned work is completed. Some patient flow improvement projects may be relatively small and not require much measurement of their project. However, once a project becomes complex or requires a significant expense by the hospital, then project accounting and a formalized work breakdown structure becomes worthwhile.

The WBS serves as a helpful device to track costs and progress. Typically a hierarchical numbering scheme is used. It divides the project into various components and levels. For example, the project is divided at a high level into functional levels and within those levels into tasks, subtasks, and work elements, such as the following:

1. Data Gathering
    - 1.1. Observing Daily Clinic Activities
        - 1.1.1. Housekeeping of patient rooms
        - 1.1.2. Admissions upon patient arrival
        - 1.1.3. Security
    - 1.2. Gathering Computer Reports
    - 1.3. Gathering Patient Records
2. Interviewing Medical Staff
    - 2.1. Interviewing Hospital Physicians
    - 2.2. Interviewing Clinic Physicians
3. Identifying Computer Resources

The WBS is a tool to define the entire components of a project in an organized way, but it generally is not a tool to determine the sequence or duration of tasks.

Once the WBS is set, the project can use its structure to develop the project work plan by defining the sequence and duration of each task. In addition, the WBS numbering scheme can be used for project team members to report their progress and to allocate their work time. Progress can be reported in terms of the percent each task is complete as well as estimates of the amount of remaining time to complete each task. Thus, the project manager has the information to determine the status of a project regarding whether it is being completed according to the planned schedule or if it may not be finished at the time planned.

A useful measurement is to compute "earned value" as tasks are partially or totally completed. For example, a patient flow project may represent tasks that total 500 h of work. Each task would have its individual hours also estimated as part of that total. As a task is completed, the completed or earned percentage can be completed. If 200 h of tasks were completed then the project is reported as having a 40 % earned value. If the project were planned to cost $10,000 then the earned value would be $4,000, which could be compared to the actual costs spent thus far on the project. Thus, administration can get a numerical picture of progress and cost as a project is underway.

Given the highly personal nature of health care services and the type of people they tends to involve, it is important to have good communications within the project and between the project team and the project's client. Many project managers have found the kick-off meeting to play a vital role. This is the initial meeting of the project, generally between the project team and the organization, which will be receiving the results of the project (the project's client). A kick-off event is a workshop type of meeting in which the beneficiaries of the project and the project team review the goals and objectives of the project, how it will be organized, etc. and who are then able to contribute to its planning, assignment of responsibilities, target dates, etc. (Wideman 2002). A project to redesign the surgery scheduling system, for example, might have a kick-off meeting with the surgery leadership and with the entire surgery staff as well as the project team. The meeting might be part of a regular OR committee meeting or a special meeting set up for this purpose.

Several things should be accomplished in the kickoff meeting. All the participants should be introduced and how people can be contacted should be explained. Perhaps most important, the project's objectives and scope should be discussed and agreed upon and any differences resolved. Only if the objectives are known can support be expected from the recipients of the results of the project.

The kick off meeting can also accomplish other objectives. For example, at this time it may be appropriate to identify the resources that the project team will need, such as office space or access to medical staff.

2.3 Project Management Issues

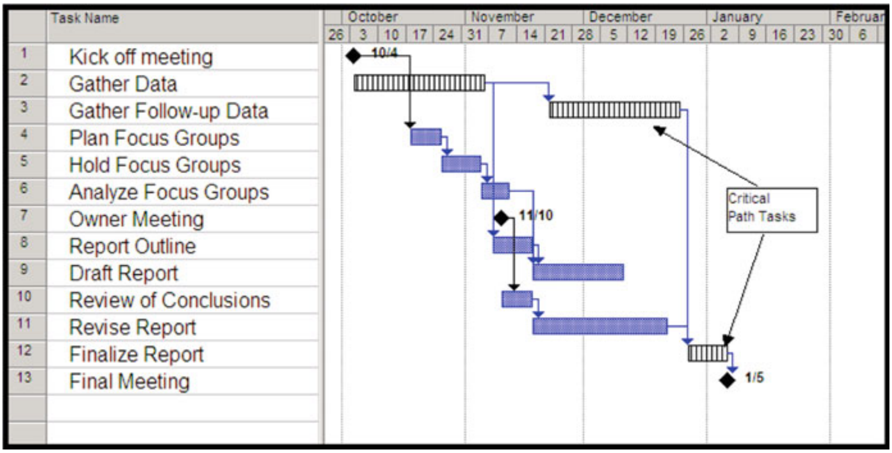
Patient flow committees are common in hospitals. The Joint Commission on Accreditation of Health care Organizations (JCAHO) virtually requires it. However, do they contribute significantly to improvements? Do they ask the right questions? Frequently, hospitals create a committee to address patient flow issues and these groups initiate individual projects to institute a particular change.

The first step must be a clear definition of what the hospital intends. There are often a number of alternative goals and each leads to a different improvement effort. Clarity is important. For example, a project to speed up patient discharge time may include the goal of a reduction in costs. Or, the goal of the same project may be to increase the number of patients served within the available resources. Still another goal might be to reduce average patient waiting time. Each of these requires a different allocation of resources, different tasks and a different measure of success even if they are all projects to speed up patient discharge time. Thus, a first step is a clear definition of goals and agreement by all of those involved.

Many organizations have found SMART goals a way to assure satisfaction with a project's success (Bovend'Eerdt et al. 2009). The letters reference the words: Specific, Measurable, Attainable, Relevant, and Time-bound. If all of these attributes are not included, the results may not meet expectations.

Other project management planning steps, in addition to setting objectives, are as follows:

- *Define the project scope*, including what you are *not* going to do. An explicit written scope becomes a contract between the project and the hospital. If the focus is on delays, then the scope should clearly define what particular delays are being addressed, which are not and to what extent are delays to be affected. The more this is clarified at the start—the better the chance for success. For example, a project may address waiting for an appointment but not address waiting after the patient arrives for the appointment.



	Task Name
1	Kick off meeting
2	Gather Data
3	Gather Follow-up Data
4	Plan Focus Groups
5	Hold Focus Groups
6	Analyze Focus Groups
7	Owner Meeting
8	Report Outline
9	Draft Report
10	Review of Conclusions
11	Revise Report
12	Finalize Report
13	Final Meeting

Fig. 21.2 Project Gantt chart with critical tasks identified

- *Secure the participation of the people* who significantly influence the area being studied. This includes doctors, nurses, and technicians. A variety of techniques exist to assure participation (Gent et al. 1998; Jones and Redman 2000).
- *Develop a project timeline* based on estimates from the project's participants. The elapsed time for each major segment must be estimated so that the project's total time can be understood. If projects are regularly late, management's confidence in a project manager will be lost. This time estimation step also considers the sequence of tasks and determines the feasible final completion date. Staff in patient flow improvement committees may optimistically take on projects that are unrealistic, particularly regarding total project time. Projects often seem to always be late. However, that is not necessary with explicit planning. Some of the causes for late projects in health care include the following:
  - Forgetting to include tasks that delay the start-up, such as the approval process, or tasks that delay the implementation of results. Often implementation and experiencing the benefits of a project take longer than expected.
  - Failure to consider the dependencies between tasks. Work on one task may depend on another, due to limited resources or the need for the output of one task to accomplish another. For example, IT may need to prepare a report before analysis of data can take place.
  - All projects have a critical path (Fig. 21.2). The critical path is the subset of tasks that constitute the longest time sequence for the project. These tasks are those that deserve the greatest attention to keep on time.
  - Health care workers often are burdened with many responsibilities but willing to take on new projects. However, their availability to work on a project may

be less than expected. Thus, the project manager must determine the time required from individual participants and if they cannot commit to the requirement—alternate plans must be made. If an individual is absolutely necessary for a project and is unavailable, the project must be rescheduled accordingly. Some project managers have found it helpful to get a signed commitment from project team members as to the amount of time they will devote to the project.

- Develop a budget for the project. People cannot accomplish significant changes when a project is merely added to their current duties. The staffs in hospitals are already fully occupied and projects such as patient flow improvement will receive a low priority unless a specific budget is allocated. Even if the project is to be done without a separate financial accounting, it is a good idea to establish how much time and resources are to be allocated to the project. Otherwise, the naive assumption that a project can be squeezed among other duties will be assumed and the project will fail.
- Implementation of intra project communications. This may include regular meetings, but periodic meetings are probably not sufficient. A project manager must regularly gather status from team members and report progress to stakeholders.

The Project Manager is a critical element in a project as the person who is responsible for the eventual successful execution of a project. If the project does not meet its objectives, the project manager is the person most responsible for any failure. The project manager is in a difficult position. In a sense, the project manager's fate is either a successful project whereupon he loses his job (because the project has ended) or an unsuccessful project where he also loses his job.

The project manager must play several roles:

- Communication with and among team. The team members may come from various departments or specialties, such as doctors, nurses, technicians or administrators. Each of these individuals has different backgrounds, training and priorities. Thus, the project manager must understand their concerns as well as get them to work together as a team with a specific set of goals and objectives. Sometimes it becomes helpful to have a "kick-off" meeting (see above).
- Communication with project's sponsors. Those interested in a project's results may include regulators, accrediting agencies and hospital administration. In some cases, patient's themselves need to be informed.
- Gathering resources needed to execute the project, such as a budget, authorization, or information systems support.
- Monitoring project progress and reporting progress to management.

Depending on the requirements of a project, the project manager may not be responsible for completing particular work plan tasks but rather serve only the leadership role.

3 Patient Flow Projects

Patient flow improvement projects typically involve a sequence of phases to first understand the current processes, analyze them and then to determine opportunities for improvement. Also typically involved in patient flow projects is the need to measure and report on the ingoing impact of a project. There may be a tendency in a health care institution to revert to practices that were in place before the improvement project took place and this tendency must be managed until the improvement becomes permanent.

Completion is also an important phase of any project. A meeting may be held to review the results and to define any ongoing effects. In the case of patient flow projects, a successful project often means a change to past practices. The success of a project should be determined regarding whether or not it made an improvement to patient flow commensurate with the cost of a project. A well-planned project would have been set up so such measurement of results will occur. For example, a project to improve radiology throughput should be able to point to the number of patients served per shift before the project was done and the number of patients served per shift after the project's recommendations were implemented.

3.1 Performance Measures in Patient Flow Projects

A key factor of success in any project is the ability to measure results. In the case of patient flow projects, here are a variety of choices. Besides the obvious measure of patients passing a certain point, waiting time is an indicator of good patient flow as is variability and total visit time per outpatient. In terms of projects to improve patient flow, the use of "SMART" goals is suggested (see Sect. 2.3).

A project to improve patient flow for a particular function, such as radiology, might set a target at the outset, such as an increase of a specific percentage above current levels or a reduction of patient wait time by a specific amount of time. Measurement of targets aids the project team in identifying where they must focus their efforts. Measurement also rewards the team at the end of a project when they can point to specific accomplishments.

4 Case Study: Identifying Patient Flow Bottlenecks

A project was requested to study patient flow and identify related problems at a large county owned teaching hospital (LAC/USC), which was discussed in Chap. 1. The county's department of health services felt that they needed an outside review of their problems and a determination of where and how improvements in patient flow could be made using the tools of systems and industrial engineering.

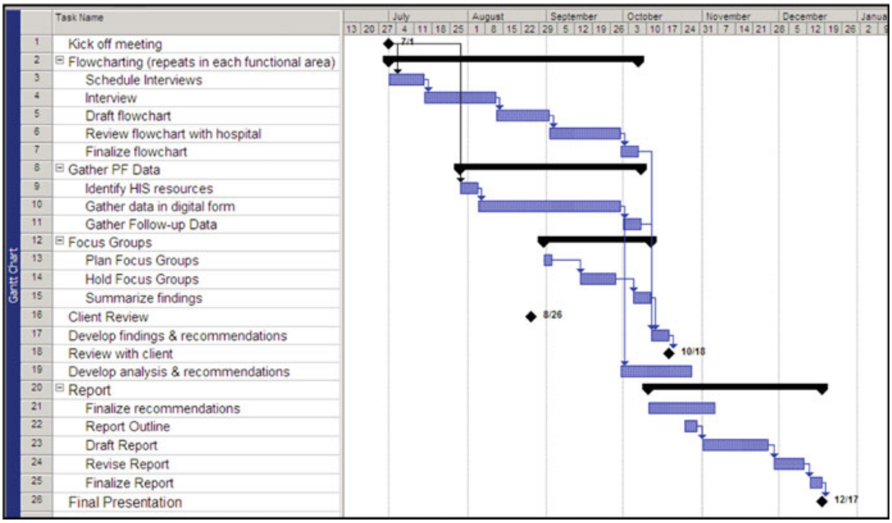


Fig. 21.3 Patient flow project bar (Gantt) chart

As noted above, it is critical to have a clearly defined objective and scope. In this particular patient flow project they were:

> To identify specific bottlenecks in the movement of patient flow in the General Hospital that could be improved by systems changes within the existing main facility and make specific recommended short-term and long-term actions. Excluded from the scope were information systems changes and clinical practices.

The project was then broken down into a group of tasks with subtasks categorized in outline form. Normally a work breakdown structure (WBS) identifying accounting categories for the tasks would precede this. However, since the project was relatively small, involving four people, and no accounting of tasks costs was necessary, a formal WBS was unnecessary. A task-by-task timeline was developed with sufficient detail so that progress could be tracked and the likelihood of delays in completion determined. The initial Gantt chart for this project is shown in Fig. 21.3.

During the course of the project, progress was tracked to assure that the agreed upon completion date would be met. There was an initial kick-off meeting before the project work began. This meeting was attended by the hospital's senior management (CEO, CMO and CNO), the county's head of health care services and members of the county's health care services department. The meeting identified individuals at the hospital that would serve as liaison and who would provide office resources to support the project. Thus, the project team knew how to get office space and services such as parking and telephones. It is not unusual for delays in such simple arrangements to delay a project's completion or reduce the quality of the results. The comprehensive nature of this initial meeting, particularly the discussion of roles and responsibilities, was critical to the project's eventual success.

The project proceeded by executing the individual tasks as planned. The most time consuming tasks were flowcharting the individual processes (summary task #2 in Fig. 21.3) and extracting patient flow statistical data from the hospital's information system (task #10). The flowcharting tasks were demanding because of the wide diversity of the general hospital's services and the simple logistics of identifying and meeting with appropriate individuals from each department. In some cases the department head was the best source of workflow information, but in many cases the best source was a nurse of technician who had years of familiarity with the department's activities. This step involved interviewing, flowchart diagramming, reviewing the initial diagrams with the department, and revising the diagrams based on their feedback. Flowcharts were also helpful when discussing changes in the workflow since the chart provided an explicit description that everyone could view. We found the best approach when flowcharting a department, such as the pharmacy or surgery, to begin at a high level and then proceed to a more detailed set of charts. The initial flowchart may contain only a few overall processes, but when the department is fully charted there may be dozens or hundreds of processes worth noting.

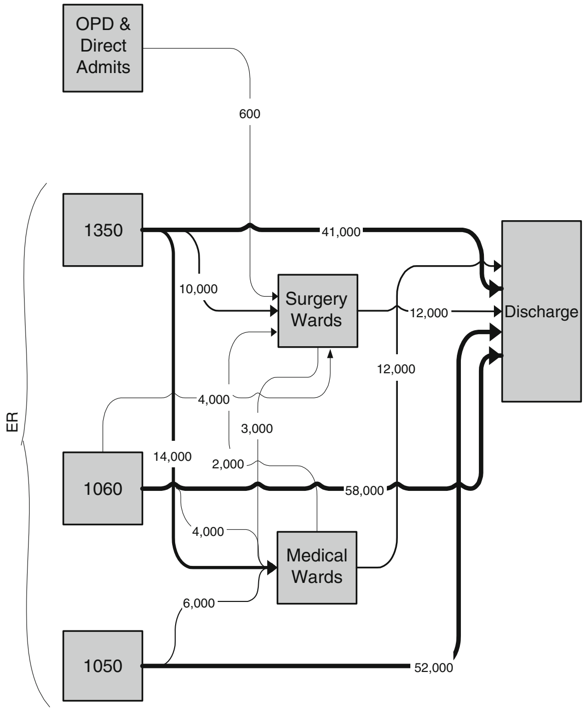
The statistical data gathering at the hospital was problematic because the Hospital Information System (HIS) was not well suited to providing the quantitative patient flow data needed. The HIS was primarily used to track individual transactions for individual patients and summary statistics could only be provided with special programming. In the end, the project team wrote its own programs to analyze the HIS data externally through spreadsheets and graphs.

The patient flow information we desired was to quantify the numbers of patients moving from one area to another. This was done at a high level (Fig. 21.4). In follow-on projects, this helped demonstrate the relative volume of patients moving between areas.

Another portion of the project was gathering input from individual hospital staff. We found that a focus group discussion was particularly productive. This involved meetings with about 20 groups of people (5–15) who were involved directly in patient care (nurses, technicians, etc.). These group meetings were conducted on a confidential basis without anyone present from hospital administration or other management. The participants were given a formal document assuring their confidentiality and explaining the purpose of the meeting. The meetings were catered to create an informal and positive atmosphere. Resulting from the meetings was a wide variety of detailed suggestions to improve patient flow. Some of the time was spent on complaints about the environment, but in each meeting there were a few individuals who had many useful ideas. These people proved to be good resources for future patient flow improvement work.

The last phase of the project involved developing recommendations, which grew from the analysis of the statistical data. The flowcharts pointed to areas with the most severe bottlenecks and the data quantified the size of the problems.

Fig. 21.4 Patient flow between hospital areas



Brainstorming by the project team also produced many recommendations once the team had the background of the flowcharting, data analysis and focus group discussions. These recommendations were then reported back to the client, which constituted the people in the initial project kick-off meeting, as well as other interested individuals.

Findings of the project were recommendations that addressed the most serious delays and it identified where redesign of processes could be accomplished with little expense. As is the case with many local government owned hospitals, budget limitations were severe, so that only recommendations that did not add to staffing or require expensive new technology were practical. The recommendations included the following:

1. Implement improved radiology department procedures to reduce queues. Radiology proved to be a bottleneck, which contributed to an increased length of stay and total inpatient costs.
2. Expand the discharge waiting room to accelerate patient bed availability. By moving inpatients promptly from a hospital bed to a waiting area made it feasible to reduce queuing in the emergency department. This is an example of the interconnected nature of patient flow problems.

3. Formalize document control. A lack of document management led to duplicate and unnecessary paperwork.
4. Optimize patient transportation staffing. Delays here also contributed to the length of stay as well as delays in areas such as radiology and surgery.
5. Revise patient ID scheme. It turned out that patient status identification contributed to patient discharge process delays and also impacted the length of stay.
6. Bed management system. Due to the high bed utilization and the need to coordinate patient discharge and transportation a more powerful control system was needed.
7. Centralize appointment system. The hospital's outpatient clinics were numerous. Decentralized appointment processes resulted in underutilization of clinics and resulted in increased ER visits by patients who were unable to gain access to outpatient services.

Lessons learned and project documentation is also an important step in all projects. An organization should not have to relearn the experience. We prepared a formal report that documented this experience. Although the hospital was more interested in the immediate recommendations the project report will help future patient flow improvement efforts.

4.1 Case Study: Implementing the Findings of a Project

The aforementioned project identified bottlenecks in patient flow, but real improvements require real implementation of changes with significant measurable results.

The initial project identified several hospital ancillary service areas where significant delays occurred. These were a particularly serious problem because they caused a costly increase in the patient's length of stay at the hospital. We had observed that patient wait times in radiology were long and unpleasant for many patients.

Radiology consisted of a variety of functions—too many to change all at once. In conjunction with managers of the radiology department we decided to focus on the CT (Computerized Tomography) function. It was important because many patients require the CT scans and the delays they experienced were long. Even with an appointment, outpatients regularly waited for 4 h in the hospital to complete their scan. We therefore organized a project to address patient flow in the CT area in the hospital.

This project was intended to implement changes based on findings identified in the previous project. Patient flow through radiology could be quantified and the net effect of changes resulting from the project measured. The physical implementation of recommendations was to be the next step.

The first part of this project involved direct observation of the CT function by an engineering member of our team. CT is basically a simple process flow with patients queuing after arrival and being processed (scanned) by one of three identical machines. Our engineering oriented observation was done daily over 1 month period in order to get a clear understanding of the work procedures and to accurately measure the time taken for each step. A flowchart was prepared (Fig. 21.5) and patient records were checked to compare to overall statistical reports generated by the department.

The results of this subproject were as follows:

- We developed a computer simulation of CT patient flow in order to test ideas and potential changes to the flow. The computer simulation allowed us to determine that total patient throughput could be doubled, increased by 100 %. Of course, management was very interested in such a finding and had a clear incentive to act on our recommendations.
- We recommended specific operational changes to significantly increase the number of patients served daily without increasing costs. These included the following:
  - When the technologist is nearing completion of a scan, he should contact the coordinator who should bring up the next patient so that there is no delay between patients.
  - Patients should be seen in the order of their appointments, rather than first come first served. While this will not change the total throughput it will reduce the average patient waiting time.
  - Staffing shortages must be reduced. Frequently CT machines were not in use due to too few technicians available.
  - Accurate throughput data must be created and verified. Previously several systems were in use, which produced conflicting estimates of the number of patients served each day.
  - A CT leadership position must be created. Someone must be identified and assigned to assure that throughput is maximized. Previously no one person held that responsibility.
  - The CT technician duties could be relieved of certain paperwork preparation and filing. My spending more of their time on the scan process, more scans will be done because the availability of technicians was a bottleneck.
  - A CT function report card report should be established. This would include daily throughput measurement and comparisons to industry benchmarks. This report should be provided to management and the results compared to expectations.

4.2 Implementation Steps

These results were presented to hospital administration and a detailed report prepared. The management of radiology also reviewed the recommended changes and agreed with nearly all the changes suggested. Just circulating these

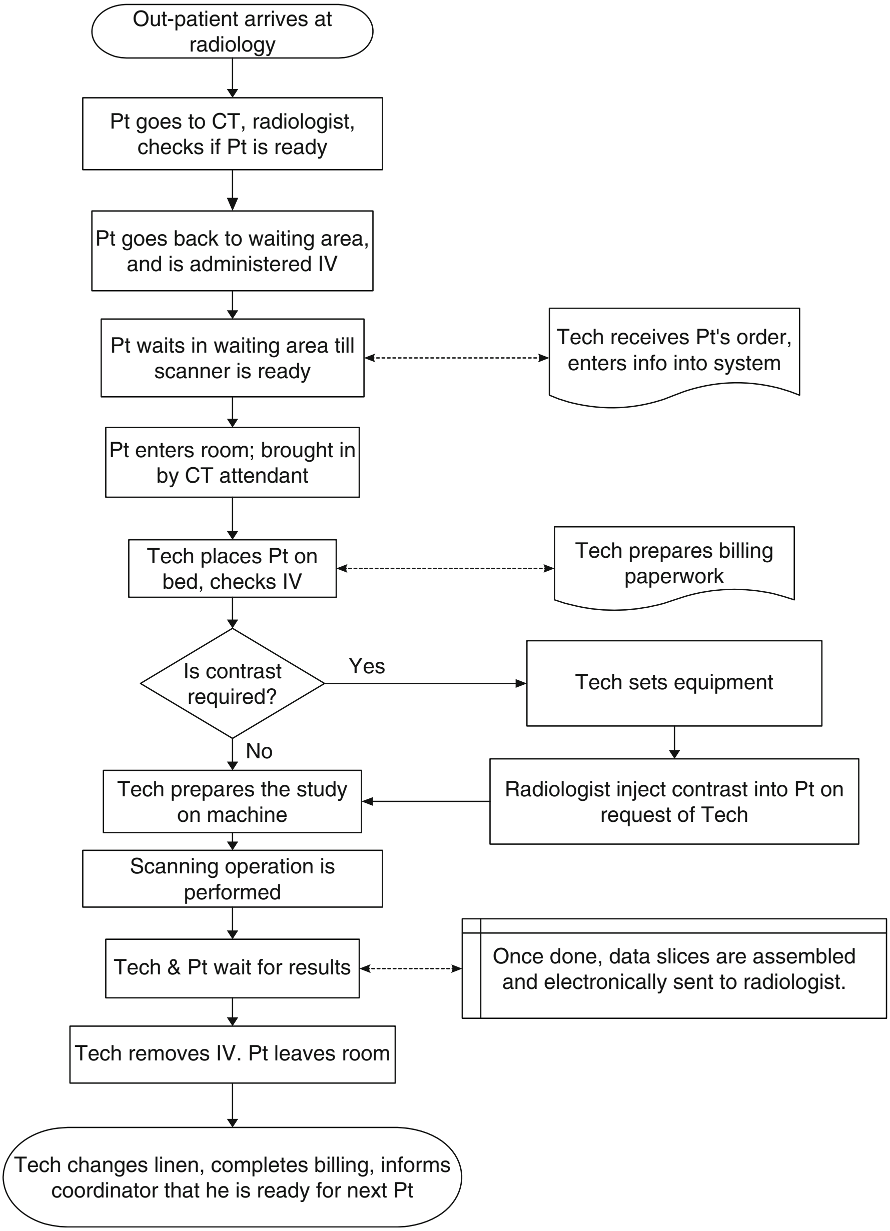

Out-patient arrives at radiology
Pt goes to CT, radiologist, checks if Pt is ready
Pt goes back to waiting area, and is administered IV
Pt waits in waiting area till scanner is ready
Tech receives Pt's order, enters info into system
Pt enters room; brought in by CT attendant
Tech places Pt on bed, checks IV
Tech prepares billing paperwork
Is contrast required?
Yes
Tech sets equipment
No
Tech prepares the study on machine
Radiologist inject contrast into Pt on request of Tech
Scanning operation is performed
Tech & Pt wait for results
Once done, data slices are assembled and electronically sent to radiologist.
Tech removes IV. Pt leaves room
Tech changes linen, completes billing, informs coordinator that he is ready for next Pt

Fig. 21.5 Radiology flowchart

recommendations seemed to have a beneficial effect. The hospital reported "dramatic" improvements resulting from awareness in the CT area of what must be done. Initially a reporting scheme was implemented which produced a daily report on the number of patients served. The calling up of patients by the coordinator seemed to accelerate and daily attendance was more closely monitored. This resulted in an increase in the number of patients served by about 30 % within a few weeks of the publication of the study.

The hospital organized an implementation committee of radiology personnel, nurses and technicians, to proceed with all the changes. By involving the staff in planning the changes, it is more likely that the changes will be fully accepted by the direct care providers themselves.

The committee meets weekly to plan, design and implement changes. The first step was for the committee to develop its own flowchart of the process. This clarified their understanding of the patient flow. They also developed additional changes to improve patient flow. The committee decided that a separate waiting area was needed for CT patients. The hospital administration agreed to the construction budget as a sign of support for the change team. Thus, changes not only included the recommendations from the patient flow project but also additional changes that were developed during the implementation process. The net effect will probably result in an even greater improvement than the 100 % forecasted by the initial study.

4.3 Report Cards to Measure Implementation

The measurement of results is an important ingredient in the implementation process. A report card provides a picture of the results of a patient flow project as well as a comparison to the target for which the project was intended. Quantitative measures provided tell the team whether the changes that they are making are actually leading to hospital improvements. It also tells management whether the project team is achieving the results they desire. The design of a "balanced scorecard" has been well established as a device to move an organization towards desired goals (Kaplan and Norton 1992). Balanced is achieved by considering several objectives, such as those of the patient as well as the hospital.

Many formats are possible for a report card as well as various possible individual measurements. Table 21.1 shows a report card for the CT patient flow improvement project. Both the patient's interest (wait time) and the hospital's objective (patients scanned per day) are included, along with modest targets. Initially, this report was done on a daily basis and senior management insisted on seeing the results each day. The reports were used as a basis for meeting with the CT department staff and management. All concerned agreed that the report card was how they should be evaluated.

In the case of this scorecard, the report was first used to inform the team designing and implementing the changes what was the current patient flow in

Table 21.1 CT Report Card

Measure	Target value	Actual average	Most recent measure
Days until next available appointment	23 days	36 days	35 days
Average outpatient wait at CT	0.5 h	2.25 h	2.3 h
Average inpatient wait for CT	6 h	28 h	25 h
Throughput, number of scans per day	150 scans	120 scans	135 scans

comparison to targeted values. Targets were developed from both industry standards and from what the clinical people in the department felt were reasonable. Many people were surprised initially at the low level of performance reported on the scorecard but this served as useful lever in motivating change.

Balanced scorecards can be used in nearly every type of project from improving an emergency room (Huang et al. 2004; Cleverley and Cleverley 2005) to hospital information systems (Gordon and Geiger 1999).

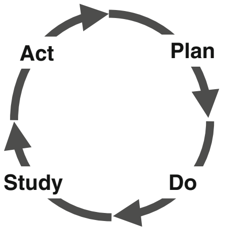
5 Creating Change

Since projects are very often about creating change, it is worthwhile to look at how we can assure that change successfully takes place in a health care organization. Many organizations have found that change is best done on an iterative basis with the initial steps being a relatively minor change. This may not seem desirable from a project management viewpoint intended to complete a project promptly. However, if a change is to be sustained, it must be implemented carefully and be supported by those affected.

A popular approach to conducting a patient flow improvement project is the so-called Lean approach. This is sometimes referred to as the Toyota approach, referencing where it was developed. “Lean” focuses on eliminating unnecessary work, as well as utilizing successful productivity improvement concepts developed in Japan in recent decades.

PDSA (Plan, Do, Study, Act), as a way to create change and organize a project, comes from the Lean approach. In order to assure a successful change, an organized and feasible approach must be used. Quality improvement changes, such as those that result from research findings, seem to succeed best when repeated small steps are used. PDSA cycles consist of planning the change (Plan), carrying out the change (Do), observing, checking and analyzing the results of the change (Study) and then deciding what additional changes should be made (Act). This is a cycle, done repeatedly until the results meet the objectives. It is best to start out with a small change or test before full implementation.

PDSA is a cycle and often shown in a graphic such as the one below:



6 Summary and Conclusions

Project management's methods provide resources to increase the likelihood of success in efforts to improve patient flow. Delays in health care services are a frequent concern for owners, managers, providers and users of health care. In an effort to reduce delays many projects are launched. However, only with proper project management methods do these projects have much chance for success in making significant improvements. Project management knowledge is available, but those involved in reducing patient delays must familiarize themselves with the fundamentals before their projects begin.

Certain project management approaches seem particularly important in projects involving delays in patient flow. One suggestion is to keep projects as simple and focused as possible. Projects that involve all the many aspects of patient flow risk the danger of being bogged down in trying to change too many things at once. Another important idea is to make sure that each project is clearly defined and well thought out at the beginning. Objectives and scope must be clear and agreed to by management and the project team. By creating a written plan at the start gives the best possible opportunity for success. Such planning may uncover flaws or gaps in the undertaking that might not otherwise be known. There is always an opportunity for reducing delays, which can always be reduced by a well-managed project.

References

Bovend'Eerdt, T. J. H., Botell, R. E., & Wade, D. T. (2009). Writing SMART rehabilitation goals and achieving goal attainment scaling: A practical guide. *Clinical Rehabilitation, 23*(4), 352–361.

Cleverley, W., & Cleverley, J. (2005). Scorecards and dashboards: Using financial metrics to improve performance. *Healthcare Financial Management, 59*(7), 64–70.

Gent, L., Parry, A. E., & Parry, M. E. (1998). The high-cooperation hospital project team. *Team Performance Management, 4*(6), 253–268.

Gordon, D., & Geiger, G. (1999). Strategic management of an electronic record project using the balanced scorecard. *Journal of Healthcare Information Management, 13*, 113–123.

Gray, C. G., & Larson, E. (2011). *Project management: The managerial process*. New York, NY: Irwin.

Huang, S. H., Chen, P. L., Yang, M. C., Chang, W. Y., & Lee, H. J. (2004). Using a balanced scorecard to improve the performance of an emergency department. *Nursing Economics, 22*(3), 140–147.

Jones, K. R., & Redman, R. W. (2000). Organizational culture and work redesign: Experiences in three organizations. *Journal of Nursing Administration, 30*(12), 604–610.

Kaplan, R., & Norton, D. (1992). The balanced scorecard – Measures that drive performance. *Harvard Business Review, 70*, 71–79.

Kerzner, H. (2004). *Advanced project management*. Hoboken, NJ: John Wiley & Sons.

MacGregor, S. P. (2005). Achieving project management success. *Journal of Product Innovation Management, 22*(3), 293.

Project Management Institute. (2013). *A guide to the project management body of knowledge (PMBOK)* (5th ed.). Newtown Square, PA: Project Management Institute.

Verzuh, E. (2011). *The fast forward MBA in project management*. New York, NY: John Wiley & Sons, Inc.

Wideman, M. (2002). Wideman Comparative Glossary of Common Project Management Terms v3.1, http://maxwideman.com/pmglossary/index.htm

Biographies

Douglas Andrusiek is a Senior Lecturer and Discipline Lead in Paramedical Sciences at Edith Cowan University in Perth, Western Australia, where he teaches a unit on trauma for the Bachelor of Paramedical Sciences degree. Formally, Dug was the Director of Research and Evaluation for British Columbia Emergency Health Services, the governing body responsible for the British Columbia Ambulance Service, Trauma Services BC, and the Patient Transfer Network for British Columbia. Andrusiek completed an M.Sc. in Epidemiology, Biostatistics, and Health Services Research at the University of British Columbia, School of Population and Public Health in 2005, during which time he validated a tool for evaluating paramedic compliance with pre-hospital trauma treatment protocols. He is a Ph.D. candidate in the School of Population and Public Health, studying Emergency Medical Services system response to cardiac arrest. Additionally, Andrusiek is an investigator with the Resuscitation Outcomes Consortium, British Columbia site, a North American-wide multicenter trauma and cardiac arrest resuscitation research network.

David Belson is a Senior Research Associate and Lecturer in the Epstein Department of Industrial and Systems Engineering, at the University of Southern California (USC). He teaches a class at the USC entitled *Improving Health Care Operations*. He previously was in management consulting with Tefen International and Ernst & Young. He served in executive positions with IBM and Universal Studios. He was the project manager on projects aimed at improving patient flow in the Los Angeles County Department of Health Services. He has worked with several hospitals in California regarding productivity improvement and information systems. Dr. Belson received his Ph.D. in Industrial Engineering from the USC.

Emilio Cerdá is a Professor of Economic Analysis. He holds a Ph.D. in Mathematics. Areas of research include dynamic optimization, stochastic programming, multiple criteria decision making, environmental and resource economics, and health care management. He is currently Head of the Department of Foundations of Economic Analysis I at the Complutense University, Madrid, Spain.

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3, © Springer Science+Business Media New York 2013

Peter Congdon is a quantitative geographer with particular interests in spatial epidemiology, statistical modeling, and health services research. Since 2001 he has been a Research Professor in Geography at QMUL and is also affiliated to the QMUL Centre for Statistics. He is the author of a range of articles and books, including the recent Wiley publications “Bayesian Models for Categorical Data” and “Applied Bayesian Modelling.” Dr. Congdon has been involved in several ESRC-funded projects, including studies on the relationship between adolescent health and neighborhood context, influence of geographic setting on social differences in health, and locality-level mortality and socioeconomic change. He has also been involved in a Department of Health Study of the public health workforce and analysis of psychiatric need indicators for the New York Office of Mental Health. His professional activities include associate editorship of Biometrics and project review work for a number of research agencies (e.g., NHS Service Delivery and Organisation R&D Programme, Alberta Heritage Foundation for Medical Research, Arts and Humanities Research Board).

Maged M. Dessouky is a Professor in the Daniel J. Epstein Department of Industrial and Systems Engineering at the USC. Dr. Dessouky’s research interests are in production and operations management, supply chain management, transportation, scheduling, simulation, and applied operations research. He is an area editor of Planning and Scheduling for *Computers & Industrial Engineering* and area editor of Transportation Simulation & Methodology for *ACM Transactions on Modeling and Computer Simulation*. Dr. Dessouky received his B.S. and M.S. degrees in Industrial Engineering from Purdue University in 1984 and 1987, respectively. He received a Ph.D. in Industrial Engineering and Operations Research from the University of California, Berkeley, in 1992.

David Evans is a Clinical Assistant Professor of Surgery, Department of Surgery, University of British Columbia, where his practice focuses on trauma and acute general surgery at the Vancouver General Hospital. He is Medical Director of Trauma for Vancouver Coastal Health and Service Director for the Vancouver General Hospital Trauma Service. He served as clinical faculty in general surgery and critical care medicine in the Department of Surgery at the McGill University Health Centre, McGill University, from 1997 to 2005 where he was director of both the Adult Trauma Program at the McGill University Health Centre (2000–2005) and the Montreal General Hospital Trauma and Acute Care General Surgery Service (1997–2005). Dr. Evans was a principal founder of the Canadian Trauma Trials Collaboration, forerunner of the current Trauma Association of Canada (TAC) Research Committee, and has served on the TAC Executive Board for over a decade. He has headed the Trauma/Critical Care Committee of the Canadian Association of General Surgeons and is a Fellow of both the Royal College of Physicians and Surgeons of Canada and the American College of Surgeons. As principal or co-investigator, Dr. Evans has participated in numerous clinical trials and currently holds a Canadian Institutes of Health Research Partnerships in Health System Improvement Grant to study policy-relevant reporting in trauma systems.

Linda V. Green is the Armand G. Erpf Professor at Columbia Business School. She earned her doctorate in Operations Research from Yale University. Her research, which has focused on the development and application of mathematical models of service systems, has resulted in numerous publications in the major technical journals including *Operations Research, Management Science,* and *The Journal of Applied Probability* as well as prominent health care journals such as *Health Services Research, Inquiry* and *Academic Emergency Medicine*. Her current research is on identifying operational policies to improve the delivery of health care. Specific projects focus on improving emergency responsiveness, developing strategies for the efficient and effective use of major diagnostic equipment such as MRIs, and developing a new nurse staffing methodology. She is a co-founder and co-director of the Columbia Alliance for Healthcare Management, a unique partnership of the College of Physicians and Surgeons, the Mailman School of Public Health, and the Business School of Columbia University dedicated to promoting interdisciplinary research and education in health care management. In recognition of her professional achievements, Professor Green was elected a Fellow of INFORMS.

Randolph Hall is Vice President for Research and a Professor in the Epstein Department of Industrial and Systems Engineering, at the University of Southern California. He previously served as the Principal Investigator for the Center for Risk and Economic Analysis of Terrorism Events (CREATE) and the Director for the National Center for Metropolitan Transportation Research (METRANS) as well as the Chair of the Epstein Department and Senior Associate Dean for Research in Engineering. He is the author of *Queueing Methods for Services and Manufacturing* and Editor for the *Handbook of Transportation Science* and *Handbook of Healthcare System Scheduling* and is the Principal Investigator on a pair of projects aimed at improving patient flow in the Los Angeles County Department of Health Services. Dr. Hall received his Ph.D. in Transportation Engineering from the University of California at Berkeley.

Shane N. Hall is a Captain in the United States Air Force and a Ph.D. candidate in the Simulation and Optimization Laboratory, Department of Mechanical and Industrial Engineering at the University of Illinois at Urbana-Champaign. He has a B.S. in Mathematics from Brigham Young University and an M.S. in Operations Research from Air Force Institute of Technology. His dissertation research includes the formulation and analysis of discrete optimization models meant to address problems in pediatric vaccination.

Katherine Harding is a researcher in the Allied Health Clinical Research Office at Eastern Health, Melbourne. She has a clinical background as an occupational therapist with experience in a range of health care settings and also holds a Masters of Public Health (James Cook University) and a Ph.D. (La Trobe University). Her research interest is in the area of access and triage systems in allied health services and their relationship to patient flow and waiting times.

Sheldon H. Jacobson is a Professor, Willett Faculty Scholar, and Director of the Simulation and Optimization Laboratory in the Department of Mechanical and Industrial Engineering at the University of Illinois. He has a B.Sc. and M.Sc. (both in Mathematics) from McGill University and an M.S. and Ph.D. (both in Operations Research and Industrial Engineering) from Cornell University. His research interests include the analysis and design of heuristics for discrete optimization problems, aviation security, and health care delivery systems. In 1998, he received the Application Award from the Institute of Industrial Engineers Operations Research Division. In 2002, he was named an Associate in the Center for Advanced Study at the University of Illinois and was awarded the Aviation Security Research Award by Aviation Security International, the International Air Transport Association, and the Airports Council International. In 2003, he received the Best Paper Award in the HE Transactions Focused Issue on Operations Engineering and was named a Guggenheim Fellow by the John Simon Guggenheim Memorial Foundation. His research has been published in a wide spectrum of journals, including *Operations Research*, *Mathematical Programming*, *INFORMS Journal on Computing*, *Operations Research Letters*, *Naval Research Logistics*, *HE Transactions*, *and the Journal of the Operational Research Society*. He has received research funding from the National Science Foundation, the Air Force Office of Scientific Research, and the Federal Aviation Administration.

Kirk Jensen, MD, MBA, FACEP, has spent over 20 years in emergency medicine management and clinical care. Board-certified in emergency medicine, Dr. Jensen is Chief Medical Officer for BestPractices, Inc, and Executive Vice President for EmCare. He has worked extensively on emergency department and hospital-wide patient flow both within his own practices and in partnership with the Institute for Healthcare Improvement (IHI). He has coached over 300 emergency departments through the process of improving operations and clinical services. He chaired the IHI communities Operational and Clinical Improvement in the Emergency Department and Improving Flow Through the Acute Care Setting and currently leads the innovative seminars Cracking the Code to Hospital-wide Patient Flow and Perfecting Emergency Department Operations. He was on the expert panel and site examination team for Urgent Matters, a Robert Wood Johnson Foundation Initiative, and is a Medical Director for the Studer Group. Dr. Jensen is coauthor of the 2008 ACHE Hamilton Award winning book Leadership for Smooth Patient Flow. He is also coauthor of Hardwiring Flow and The Hospital Executive's Guide to Emergency Department Management. Dr. Jensen teaches at the American College of Emergency Physicians (ACEP) Directors Academy, leading ED directors through process and operational improvements.

Hongzhong Z. Jia received his dual B. Eng degree in Mechanical Engineering and Industrial Management from Shanghai Jiao Tong University, China, and then completed his M. Eng study at the Department of Mechanical Engineering, National University of Singapore (NUS), in 2001. He is currently a Ph.D. candidate majored in Operations Research at the USC. His research interests include vehicle routing, manufacturing scheduling, agent-based systems, and product development. He is a member of INFORMS.

Alexander Kolker, Ph.D., has more than 10 years of practical experience in quantitative health care operations management and simulation modeling for hospital capacity planning, system-wide patient flow, optimized staffing, forecasting, and business analytics. He is the author of *Healthcare Management Engineering: What Does This Fancy Term Really Mean*? as well as the lead editor and author of *Management Engineering for Effective Healthcare Delivery: Principles and Applications*. He has published 6 book chapters and 8 peer-reviewed journal papers and delivered 18 national and international conference presentations and webinars in the area of simulation modeling and operations management in health care settings. As an adjunct faculty of the University of Wisconsin-Milwaukee, Lubar School of Business, he developed and taught the graduate course "Healthcare Administration & Delivery Systems." Dr. Kolker currently works at API Healthcare in Hartford, Wisconsin.

Linda Kosnik is the Chief Nursing Officer at Overlook Hospital, Atlantic Health System in Summit, NJ. Ms. Kosnik received her undergraduate degree from Columbia University and her MSN from Seton Hall University. She is currently a doctoral candidate in health science leadership also at Seton Hall University. Ms. Kosnik has lectured nationally and has published extensively on collaboration, quality improvement, and process and performance management as applied to health care. She is a nationally recognized expert on reducing waits and delays and in the overarching domain of improving patient flow in hospitals, especially as it relates to demand and capacity management. Through the IHI and other forums, Linda has successfully mentored hundreds of teams through improvement initiatives. Using a collaborative approach to leadership, she has established national recognition for Overlook Hospital, including receiving the New Jersey Governor's Award for Quality Improvement that is based on Malcolm Baldrige Criteria.

Lisa Kuramoto is a Statistical Analyst in the Centre for Clinical Epidemiology and Evaluation, a part of the Vancouver Coastal Health Research Institute in Vancouver. She holds an M.Sc. in Statistics from the University of British Columbia. Her research interests are primarily in investigating surgical waiting times and their impact on patient outcomes.

Adrian R. Levy is Associate Professor in the Department of Health Care and Epidemiology at the University of British Columbia and is an investigator at St Paul's Hospital in the Centre for Health Evaluation and Outcome Sciences. He received his Ph.D. from McGill University in epidemiologist, and his academic interests are in health services research, specifically in the areas of access to care, pharmacoepidemiology, and economic evaluation. He has considerable experience using administrative databases and population registries for carrying out population-based health services research.

Mark Lindsay, MD, MMM, is an Assistant Professor of Medicine at Mayo Clinic School of Medicine. He presently serves as Medical Director for Allevant Solutions LLC, a joint venture of Mayo Clinic and Select Medical that is dedicated to optimizing post-acute care pathways. Dr. Lindsay served as Quality Officer for Mayo Clinic Health System, providing leadership to clinical outcomes, patient safety, and service excellence for 19 hospitals and 70 clinics for 5 years. He previously served as Department Chair Pulmonary and Critical Care at Mayo Clinic Health System, Eau Claire. He completed his Medical School, and postgraduate training, at the USC School of Medicine where he stayed on staff serving as Assistant Professor of Medicine at the USC. He completed his Masters in Medical Management at the USC School of Business in 2004. His project was expanding ventilator and transitional care programs across Mayo.

Megan McHugh is a Research Assistant Professor in the Center for Healthcare Studies and Department of Emergency Medicine, Northwestern University, Feinberg School of Medicine. She is also director of the Center's Program in Health Policy and Implementation. Her research focuses on hospital quality improvement, emergency department operations, value-based purchasing, and health policy making. Previously, Dr. McHugh was research director at the Health Research & Educational Trust of the American Hospital Association and Senior Program Officer at the Institute of Medicine. Dr. McHugh received her Ph.D. in public policy from The George Washington University.

Pavan Murali is a doctoral student in Industrial Engineering at the USC. He holds a Bachelor's degree in Mechanical Engineering from the Indian Institute of Technology, Madras, and a Master's degree in Operations Research from the USC. His research interests mainly lie in the design and implementation of algorithms for combinatorial and nonlinear optimization. He is currently working as a research assistant on a project involving simulation and mathematical modeling to improve health care operations in the radiology, surgery, and bed management departments at the Los Angeles County Hospital.

Fernando Ordóñez is a professor of industrial engineering at Universidad de Chile. His research focuses on convex optimization, robust optimization, complexity of algorithms, sensitivity analysis, condition number theory, and applications of optimization to engineering and management science. He received his BS and Mathematical Engineering degree, from the University of Chile in 1996 and 1997, respectively, and his Ph.D. in Operations Research from MIT in 2002.

Laura de Pablos is Head of the Department of Applied Economics VI at the Complutense University, Madrid, Spain. She has published several papers and books about Public Economics. Areas of research include budget control, evaluation and control of public spending policies, analysis of the impact of public spending, education economics, and health economics.

Roger Resar, MD, is an Assistant Professor of Medicine at the Mayo Clinic School of Medicine and Change Agent for Luther Midelfort and Mayo Foundation. He shares his time between consulting within the Mayo System and the IHI where he is a Senior Fellow and functions as faculty and a founding member of the IHI Innovation Team. Dr. Resar has contributed to the development and dissemination of key safety and improvement strategies, including Medication Reconciliation (which has been deemed a Joint Commission standard for 2006), a trigger tool methodology that is currently being used in hundreds of hospitals to measure adverse events, "bundle science," reliability concepts for design improvement work, and basic work in hospital flow. Dr. Resar is boarded in Pulmonary and Critical Care Medicine and graduated from the University of Wisconsin Medical School in 1972.

María V. Rodríguez-Uría is a well-known scientific expert in the MCDM field with numerous articles in the concerned literature. She is a Professor and Head of the Department of Quantitative Economics at the University of Oviedo in Spain. She has also supervised several Ph.D. theses, and she is responsible for the Multicriteria Decision Making Group of her University.

Sergei Savin is an Associate Professor of Decision, Risk, and Operations at the Graduate School of Business, Columbia University. Professor Savin holds Ph.D. degrees in Physics from the University of Pennsylvania and in Operations and Information Management from the Wharton School. His research encompasses the areas of health care management, revenue management, marketing–operations coordination, e-business, and new product development. Professor Savin's works have been published in leading academic journals *Operations Research* and *Management Science* and have been supported by the grants from Columbia Health Alliance and Columbia Center for Excellence in E-Business. He serves as an Associate Editor for *Operations Research*, as a Senior Editor for *Production and Operations Management*, and as an Editorial Board member for *Manufacturing and Service Operations Management*.

Zhihong Shen is currently a Ph.D. student majored in Operations Research in the Industrial and Systems Engineering department at the USC. She received her BS in Electronics and Information Science and Technology and Bachelor of Economics from Peking University, China, in 2000 and then completed her MS in High Performance Computing at Singapore-MIT Alliance in 2001. Her current research interests include vehicle routing, production and operations management, and applications of optimization.

Boris Sobolev is a Professor at the School of Population and Public Health at the University of British Columbia. As a Canada Research Chair in Statistics and Modeling for Health Care he contributed extensively to the area of health services research. His research interests include understanding the link between processes and outcomes of care, with a specific focus on access to care, patient safety, adverse drug events, surgical outcomes, and comparative effectiveness of interventions. His research at UBC culminated in the publication of two books: Analysis of Waiting-Time Data in Health Services Research, which was met with a warm reception from scholars and professionals around the world, and more recently, his second book, Health Care Evaluation Using Computer Simulation, considered an authoritative reference in this domain. He is Editor-in-Chief for the Handbook of Health Services Research, a six-volume major reference work commissioned by Springer.

James R. Swisher is Director of Performance Improvement for Mary Washington Hospital (MWH). Mr. Swisher currently oversees MWH's quality, service, and industrial engineering programs. Upon his arrival at MWH in 1999, he established and has grown the industrial engineering function at MWH. He now works to leverage industrial engineering and operations research approaches to improving health care processes at MWH, particularly in the fields of clinical quality, patient satisfaction, and patient flow. Prior to joining MWH, he worked as Director of Analysis and Design for Biological & Popular Culture, Inc where he was responsible for management engineering. He holds a B.S. and an M.S. in Industrial Systems Engineering (Operations Research) from Virginia Tech and an M.B.A. from Mary Washington College. He is interested in the application of operations research techniques to complex business problems, particularly in the field of health care. Mr. Swisher is a member of INFORMS, HE, and Alpha Pi Mu.

Nicholas Taylor, Ph.D., is a Professor of Allied Health at Eastern Health and La Trobe University in Melbourne. He leads the Allied Health Clinical Research Office at Eastern Health. He is an active researcher on improving allied health services and has published widely on the effects of physical activity and exercise.

Jan M.H. Vissers is a Professor in Health Operations Management at the Institute for Health Policy and Management at Erasmus University Medical Centre Rotterdam at Rotterdam, NL. He is also affiliated with the Department of Technology Management of Eindhoven University of Technology, Eindhoven. He is furthermore a senior management consultant at Prismant—Institute for Health Care Management Development in Utrecht. Dr. Vissers Received his M.Sc. Industrial Engineering and Management Science from EUT in 1975 and his Ph.D. from EUT in 1994. He acts as current chairman of the European Working Group on Operational Research Applied to Health Services and is a member of the editorial board of Health Care Management Science. He received the 1995 Baxter Award for his thesis "Patient Flow based Allocation of Hospital Resources" for its contribution to Health Care Management. Together with Roger Beech, he edited the book *Health Operations Management. Patient Flow Logistics in Health Care*, published in 2005 by Routledge. His research focuses on the analysis, design, and control of operational health care processes and systems.

Michael Warner is Chairman and Chief Science Officer of AtStaff Incorporated, designing staff management systems for health care. Previously, he co-founded and was president of Atwork, Inc., a software company which he sold in 1995. Previous to Atwork, Warner was a faculty member at Duke University's Fuqua School of Business, Duke's Medical School, and the University of Michigan's School of Public Health. He is author of two textbooks in Health Administration and numerous articles on the application of Operations Research to Health Administration.

Dr. Warner received a BS in Math and a Masters in Health Administration from Duke University and a Ph.D. in Operations Research from Tulane University.

Michael Williams is the President of The Abaris Group, a firm that specializes in emergency department and inpatient program patient flow and capacity building strategies. He has personally conducted greater than 350 hospital studies on improving performance, productivity, and market share. He is a recognized expert on health care performance, benchmarking, and financing having conducted 15 presentations during the past 24 months for organizations such as ACEP, California Hospital Association, Connecticut Hospital Association, American Trauma Society, Urgent Matters, The Zoll Corporation, The Coding Institute, and Beta. He is a frequent contributor to the Healthcare Advisory Board on a number of hospital and ED subjects. He is also a member of CAL/ACEP's Reimbursement Committee. He is also on the editorial board for *ED Management*. He is a senior faculty member for The Robert Wood Johnson Foundation's Urgent Matters Project (urgentmatters.org), which is a ten-hospital collaborative throughout the country and an instructor for Harvard University on their course on "Designing EDs for the Future."

Index

R. Hall (ed.), *Patient Flow: Reducing Delay in Healthcare Delivery*,
International Series in Operations Research & Management Science 206,
DOI 10.1007/978-1-4614-9512-3, © Springer Science+Business Media New York 2013