TUTORIALS IN
OPERATIONS RESEARCH
INFORMS 2008

 | ISBN 978-1-877640-23-0
DOI 10.1287/educ.1080.0049

Using Operations Research to Reduce Delays for Healthcare

Linda V. Green
Columbia Business School, New York, New York 10027, lvg1@columbia.edu

Abstract The Institute of Medicine identified "timeliness" as one of six key "aims for improvement" in its most recent report on quality. Yet patient delays remain prevalent, resulting in dissatisfaction, adverse clinical consequences, and often, higher costs. This tutorial describes several areas in which patients routinely experience significant and potentially dangerous delays and presents operations research (OR) models that have been developed to help reduce these delays, often at little or no cost. I also describe the difficulties in developing and implementing models as well as the factors that increase the likelihood of success. Finally, I discuss the opportunities, large and small, for using OR methodologies to significantly impact practices and policies that will affect timely access to healthcare.

Keywords healthcare; delays; queuing

Healthcare is riddled with delays. The Institute of Medicine (IOM), the major advisory body to the government and the private sector on health-related issues, identified timeliness as one of the six key aims for improvement in its major report on quality of healthcare in 2001. This was in response to the growing evidence of significant delays across the healthcare industry and the impact of these delays on clinical outcomes.

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 wait some more until being seen. In hospitals, it is not unusual to find patients backlogged in the emergency department waiting room and on stretchers in hallways waiting for beds. Inpatient falls, which often result in broken bones and serious complications, are known to be more prevalent when nurses are delayed in answering call buttons. Although waiting for service is inconvenient and annoying in other situations, in healthcare, delays can be dangerous or even deadly.

This tutorial will describe some of the major types and sources of delay for healthcare, why they exist, and how operations research (OR) approaches can help provide insights and guidance for reducing delays, often with little or no increase in costs. In presenting specific examples, this tutorial will focus more on the contextual issues that are important to consider in developing, implementing, and interpreting the results of models than on the technical details of the models themselves. I will also discuss the regulatory, institutional, and systemic obstacles to using these methodologies to improve patient access to healthcare and look at efforts to reduce or eliminate these obstacles in the future. Finally, drawing from experiences across a range of applications, I will discuss the factors that appear to be most important in the adoption of OR approaches for reducing healthcare delays and how OR can play a role in influencing healthcare policy.

1. Three Sources of Dangerous Healthcare Delays

1.1. Emergency Department Delays

Arguably, the most critical delays for healthcare are the ones associated with healthcare emergencies. Unfortunately, emergency department (ED) overcrowding is a continuing and

growing problem. In a 2007 survey (American Hospital Association [2]), nearly half of all U.S. hospitals and 65% of urban hospitals reported being at or over capacity in their emergency rooms, resulting in long waits before being seen by a physician and delays of many hours or even days in getting a hospital bed. Each of these sources of delay can be life threatening.

1.1.1. Delays for ED Physicians. Delays for emergency care generally begin with the wait to be seen by a physician. Many patients who arrive to an ED are "nonurgent" and would not be harmed by significant delays in seeing a physician. However, many, if not most, are either "emergent" (requiring "immediate" care) or "urgent" (requiring care within a "short" period of time). Yet recent studies have revealed that waits to see an ED physician are long and getting worse. The overall median wait to see an ED physician increased from 22 minutes in 1997 to 30 minutes by 2004. Perhaps even more alarmingly, the median wait for patients diagnosed with acute myocardial infarction (AMI) (heart attacks) increased from 8 minutes in 1997 to 14 minutes in 2004 (Wilper et al. [35]). These figures are considerably higher in urban areas and for teaching hospitals, where most emergency care is delivered. Given the urgency of rapid treatment for victims of heart attack (as well as for sepsis, stroke, pneumonia, and trauma, among others) these figures, as well as anecdotal reports of patient deaths in EDs, indicate that these delays are a serious threat to the health and well-being of us all.

In addition, previous studies have established a strong link between long emergency department delays and the fraction of patients who leave without being seen (LWBS) (Green et al. [16], Fernandes et al. [8]). The proportion of patients who LWBS is itself an important measure of emergency department performance and quality of care. Several studies have concluded that patients who LWBS are sick and do require emergency care. One study has shown that up to 11% of patients who leave without being seen are hospitalized within a week and 46% of patients were judged to require immediate medical attention (Baker et al. [5]).

1.1.2. Delays for Inpatient Beds. Although insufficient ED physician levels are responsible for significant and serious delays in emergency care, the single biggest cause of ED overcrowding is a lack of inpatient beds (American Hospital Association [2]). About 25% of ED patients are admitted to the hospital and approximately 40% of inpatients come through the ED. For hospitals reporting ED capacity problems, the average time to move a patient from the ED to an inpatient bed once the decision to admit has been made is over 4.5 hours, and bed delays of over 24 hours are not uncommon (McCaig and Burt [21]). These delays lead to large numbers of patients in the ED who strain the ability of the physicians and nurses to care for them. This ED overcrowding also compromises the ability of ED physicians to treat new arrivals to the ED. When this kind of situation becomes untenable, a hospital will usually go on ambulance diversion—suspending new ambulance arrivals to the ED. In a 2007 survey, over half of all urban and teaching hospitals reported being on diversion in the past 12 months (American Hospital Association [2]). When hospitals are on diversion, seriously ill patients have to travel further to get to the nearest hospital, increasing their risk for adverse outcomes. In addition, due to the increased travel time, waits for ambulances increase, further endangering patients. A study based on data from New York City showed that when there are significant levels of ambulance diversion in a borough, the number of deaths due to AMI increases by 44% (Yankovic et al. [37]).

Bed delays can occur even when inpatient beds are available. This is because patients and beds are not all identical. Hospitals are divided into nursing units, which typically consist of between 20 and 50 beds. Each nursing unit is used for one or more clinical services (i.e., medical, surgical, pediatric, obstetrics, cardiology, neurology). So a patient with heart disease who is waiting in the ED for a bed may experience a significant delay if all the cardiology beds are occupied, even though beds may be available in another clinical unit. In addition, some patients need telemetry beds, which, in many hospitals, are not available in

all nursing units. Finally, most hospital rooms have two or more beds that can be used only by patients of the same gender. So it is possible for, e.g., a female patient to be delayed in getting an inpatient bed if the only beds that are available are in rooms occupied by male patients.

The most common delays for inpatient beds are for critical care beds (American Hospital Association [2]). Intensive care units (ICUs) are usually the most expensive units in the hospital due to both the technology utilized and the high level of staffing needed. The full per-day cost in an ICU is about three to five times as much as in a regular inpatient unit (Groeger et al. [17]), and therefore, hospital administrators believe that these beds should be highly utilized in order to be "cost efficient." Consequently, delays for these beds, which are used for the most seriously ill and injured patients, are often the longest.

1.2. Delays for Medical Appointments

Difficulty in getting a timely appointment to see a physician is a very common problem. In one study, 33% of patients cited inability to get an appointment soon as a significant obstacle to care (Strunk and Cunningham [32]). For most patients, their primary care physician is their major access point into the healthcare system. Yet primary care practices often have long waits for appointments. The average wait for a primary care appointment in the United States in 2001 was over three weeks (Murray and Berwick [25]). A more recent survey in Massachusetts found that the average delay was over six weeks (Mishra [24]). Long appointment backlogs result in difficulty in accommodating patients who have potentially urgent problems. As a result, patients experience delays in treatment that may result in adverse clinical consequences and patient dissatisfaction.

Two-thirds of all primary care physicians work in group practices, so patients who cannot get an appointment with their own physician may instead be seen by another physician in the practice. However, continuity of care has been identified as an important factor in accurate diagnosis and appropriate treatment (Smoller [30]). Another adverse consequence of long waits for appointments is the use of the ED for primary care, contributing to the overcrowding mentioned above.

From the physician practice perspective, large backlogs may require additional staff and resources to deal with patients trying to get appointments for the same day, and are often correlated with a high rate of cancellations, resulting in a loss of revenue for the practice.

1.3. Delays for Nursing Care

In a landmark report entitled "To Err is Human" (Institute of Medicine of the National Academies of Science [19]), the IOM reported that medical errors in hospitals are responsible for between 44,000 and 98,000 deaths each year and for more than one million injuries. A more recent report from Healthgrades [18] estimates the number of deaths to be 248,000 per year. Even at the lowest level, this implies that more people are killed by medical errors than die from traffic accidents or most forms of cancer. Although the causes of these preventable deaths are numerous, recent studies have indicated that insufficient nurse staffing levels are a major factor in the prevalence of medical errors (Aiken et al. [1], Needleman et al. [28]). This is not surprising because many nursing tasks are time sensitive. These include administration of medications, responding to patient call buttons, and dealing with emergency admissions. Although nursing workloads have increased in recent years due to a higher percentage of elderly and sicker inpatients and shorter hospital lengths of stay, staffing levels in most hospitals have either remained static or, in many cases, decreased.

Nursing costs comprise a very substantial fraction of hospital budgets, and therefore, cost-effective nurse staffing is very important. In most hospitals, the number of nurses assigned to a unit is determined by a specified ratio of patients to nurses. The historical norm for most types of clinical units has been 8:1, whereas for intensive care units it could be as little

as 1:1. Although most hospitals subscribe to these standards, cost pressures and a national nursing shortage have resulted in these ratios being exceeded in many cases. Sometimes, however, this is the result of a failure to adequately plan for the daily, weekly, and sometimes seasonal variations in hospital census that are common in most clinical units of virtually every hospital.

Insufficient inpatient nursing levels are also a major factor in delays for beds in the ED. For a patient to be transferred to an inpatient bed from the ED, a nurse needs to be available to handle the admission to the unit. So if the staffing level in the unit is such that the nurses are highly utilized, many patients waiting for a bed in that unit will continue to wait even when a bed becomes available.

Nursing levels also play a role in delays for care within the ED. As in other parts of the hospital, nurses are the primary managers and caretakers of patients. If the number of patients in the ED gets too high for the nurses to handle, either new arrivals will not be taken into the treatment area until some of the patients are discharged or moved to inpatient beds, or, if the ED is very congested, the hospital may go on ambulance diversion.

Delays due to lack of nursing personnel are arguably more problematic in the ED for several reasons. First, the need for speedy treatment is more urgent. Second, the variability in the volume and pattern of patient arrivals, as well as their diversity, make it more difficult to predict nursing needs. And finally, the stress of the work environment in the ED makes it more difficult to attract and retain nurses, and the levels of absenteeism are generally higher than in the rest of the hospital.

2. Why Are Healthcare Delays So Bad?

Given the enormous impact on society's well-being, it may be surprising that patients' delays in receiving healthcare, as described above, are so bad and, in some cases, may be getting worse. There are several major reasons for this.

First, until recently, these delays and their impact haven't been very visible either to the public or, in many instances, even to the healthcare community. This is because delays in healthcare have not, in general, been measured or reported. The quality movement came late to healthcare and the focus has traditionally been on clinical innovation, not timely delivery of service. In national rankings, hospitals have been primarily evaluated on clinical excellence, which is based on the reputation of their staff physicians. Because hospital prices were tightly regulated and compensation was based on the fee for service until the early 1990s, hospitals were largely insulated from competitive pressures and, consequently, paid little or no attention to patient satisfaction. Similarly, until the last several years, the major accreditation agency for hospitals, The Joint Commission on the Accreditation of Healthcare Organizations, focused attention on clinical processes rather than service or quality factors, which were rarely reported.

Second, with the advent of price deregulation and increasing pressure from government, employers, and managed care organizations, hospitals have been forced to focus on cutting costs. Due largely to decreased levels of compensation from Medicare and Medicaid, which account for over 50% of hospital revenues, about one-third of all hospitals have negative operating margins and the vast majority of them struggle to break even (American Hospital Association [3]). Until the last few years, the number of hospitals had been steadily declining for decades. Many more have downsized, particularly since the mid-1990s. Many of these closures and downsizings have been directly due to the use of target occupancy levels that have been used since the mid-1970s at both the federal and state government level to evaluate and control the supply of hospital beds (McClure [22]). As recently as December 2006, a commission set up by Governor Pataki of New York State mandated the closure of almost 20 hospitals and the downsizing of dozens more using their finding of an average occupancy rate of 77% compared with the "ideal rate of 85%" as their primary evidence of

"excess capacity" (Commission on Healthcare Facilities in the 21st Century [7]). This was done without any analyses of the impact on delays for emergency or inpatient care. At the institutional level, 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.

Third, healthcare needs are highly concentrated within a small percentage of the population. About two-thirds of all healthcare expenditures are attributable to about 10% of the population. This population consists primarily of people with chronic diseases such as cardiovascular disease, cancer, pulmonary illnesses, and diabetes. The percentage of people who have these illnesses has been steadily increasing, partially as a result of an aging population. This increase in illness is putting an increasing strain on all aspects of the healthcare delivery system, particularly emergency rooms and physician practices.

Finally, many delays for healthcare are due to inadequate levels of healthcare professionals, particularly doctors and nurses. Unfortunately, these are not just local problems, but often a result of national shortages. The nursing shortage has been well documented; see Figure 1 and U.S. Department of Health and Human Services [33]. This nursing shortage has seriously impeded hospital efforts to increase nurse staffing levels, particularly as opportunities for nursing work outside of hospitals have increased.

Although it is harder to quantify physician shortages, hospitals often have difficulty in staffing emergency rooms, particularly for evenings and weekends. Residency programs in emergency medicine are not as prevalent as for many other specialties, and the compensation is not as high, making it even more difficult to increase the supply. Emergency departments also have difficulty getting certain specialists, for example, orthopedist and neurosurgeons, to be "on call" particularly for evenings and weekends, for the same reasons of inconvenience and relatively poor compensation (American Hospital Association [2]).

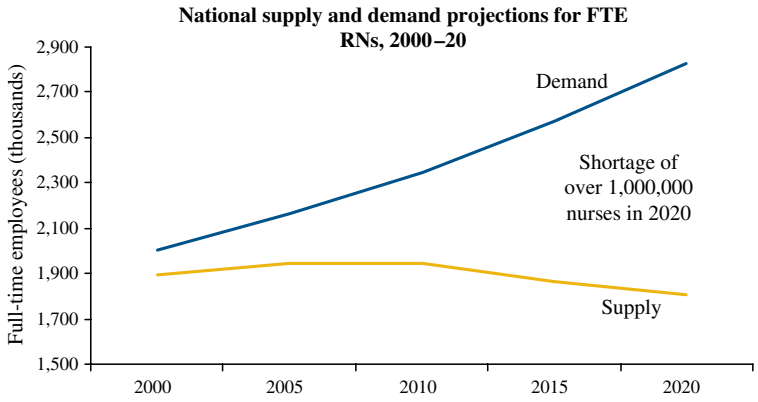
Growing backlogs for medical appointments for primary care are likely due, in part, to a lack of primary care physicians. The United States has fewer primary care physicians per capita than almost any other industrialized nation and there is a particular lack in rural areas (American Medical Association [4]).

3. How Can OR Help?

3.1. Using OR to Reduce Delays for Emergency Care

In this section, I describe some examples of OR-based studies related to understanding and reducing the types of healthcare delays described above.

FIGURE 1. The nursing shortage.



Source. Adapted from National Center for Health Workforce Analysis, Bureau of Health Professions, Health Resources and Services Administration, 2004. *What Is Behind HRSA's Projected Supply, Demand, and Shortage of Registered Nurses*? Link: ftp://ftp.hrsa.gov/bhpr/workforce/behindshortage.pdf.

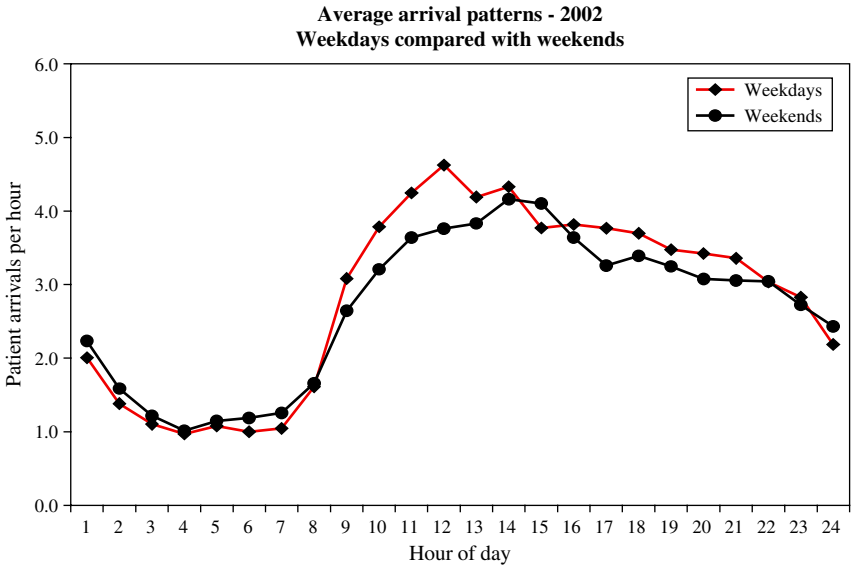
3.1.1. Delays for ED Physicians. As mentioned above, visits to EDs have been steadily increasing. Constrained provider capacity relative to demand volume is exacerbated by the extreme variability in demand during each 24-hour period experienced by a typical emergency department. This time-of-day pattern, as reported in the National Hospital Ambulatory Medical Care Survey for 2002 (McCaig and Burt [21]), is distinguished by a relatively low level of demand during the night followed by a precipitous increase starting at about 8 or 9 A.M., a peak at about noon, and persistently high levels until late evening. In addition, though the general pattern of demand is similar across the days of the week, individual days are likely to experience different overall volumes as well as slight differences in the exact timing of peaks and valleys. In particular, emergency departments are likely to have fewer visits on weekends than on weekdays.

Though physician staffing levels are generally varied over the day to meet these changing levels, this is usually done without the aid of operations research models. In a study conducted in the ED of a midsize urban hospital in New York City (Green et al. [14]), the overall volume varied from a low of 63 patients per day on Saturdays to a high of 72 per day on Mondays. 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 queuing analyses to develop two schedules: weekday and weekend. This required aggregating ED arrival data into these two groups. For each, demand data were collected for each hour of the day using the hospital's admissions database to understand the degree of variation over the day (see Figure 2) to explore the impact of different shift starting times and/or staffing levels on delays.

Estimating the average provider service time per patient was more difficult. This time includes several activities such as 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 were estimated indirectly from direct observation and historical productivity data.

Traditionally, in a service system with time-varying arrivals, the desired staffing levels would be determined by the *stationary independent period by period* (SIPP) approach, which

FIGURE 2. Average arrival patterns for the Allen Pavilion.



begins by dividing the workday into planning periods, such as shifts, hours, half-hours, or quarter-hours. Then a series of stationary queuing models, most often $M/M/s$-type models, are constructed, one for each planning 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. In Green et al. [12], the SIPP approach was shown in many cases to seriously underestimate the number of servers needed to meet a given delay performance target. This is particularly true when the mean service times are high (e.g., 30 minutes or more) and planning periods are long (two hours or more). In these situations, it was demonstrated that a simple variant of SIPP, called Lag SIPP, performs far better than the simple SIPP approach. The major reason is that in cyclical demand systems, there is a time lag between the peak in the arrival rate and the peak in system congestion. This lag is significant when the mean service time is long. Lag SIPP corrects for this factor.

In our ED physician staffing study Green et al. [12], the *Lag SIPP* approach was applied by first advancing the arrival rate curve by our estimate of the average physician time per patient, 30 minutes. We then constructed a series of $M/M/s$ models for each two-hour staffing interval, using the average arrival rate for each based on the time-advanced curve and the estimated average 30-minute service time. The delay standard we choose was that no more than 20% of patients wait more than one hour before being seen by a provider. The use of one hour 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 [21]). The 20% criterion reflected the approximate percentage of nonurgent arrivals at the study institution.

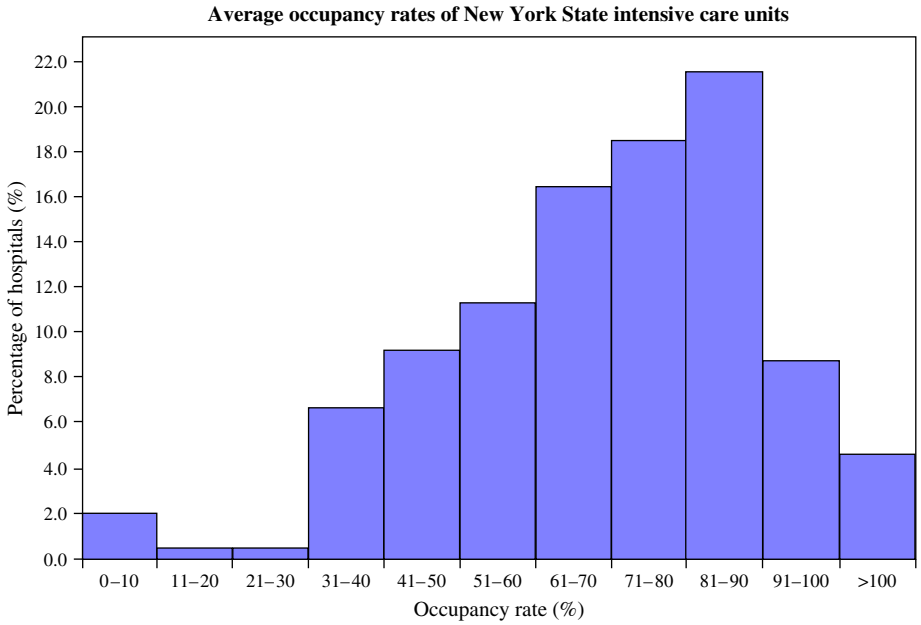
The modeling results gave the number of ED physicians needed in each of the two-hour staffing intervals to meet the delay standard. In total, 58 physician hours were needed on weekdays to achieve the desired service standard, which represented an increase of three hours over the existing staffing level of 55 hours. Model runs for the weekend indicated that the target performance standard could be achieved with a total of 53 provider hours. In both of these cases, the queuing 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 the staffing level to handle the morning surge in demand needed to occur earlier than in the original schedule. Although resource limitations and physician availability prevented the staffing suggested by the queuing 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, four hours were moved from the weekends and added to the Monday and Tuesday afternoon shifts (because these were the two busiest days of the week), and a shift that previously started at noon was moved to 10 A.M..

At the time of the study, the study hospital did not collect data on delays to be seen by a physician. So to evaluate the impact of the new staffing levels, we focused primarily on the effect on the percentage of patients who LWBS. Despite an increase in patient arrival volume of 6.3%, an increase in provider hours of only 3.1% resulted in a decrease in the proportion of patients who left without being seen by 22.9%. Restricting attention to a four-day subset of the week during which there was no increase in total provider hours, a reallocation of providers based on the queuing analysis resulted in a decrease in the fraction of patients who left without being seen of 21.7%, although the arrival volume increased by 5.5%.

3.1.2. Delays for Inpatient Beds. As mentioned above, the most frequently cited reason for ED overcrowding is the lack of critical care beds. Yet, these units are usually very small and intentionally operated at high levels of utilization so as to be "cost efficient."

In a study of the impact of target occupancy levels on delays for time-sensitive conditions (Green [10]), we used the basic $M/M/s$ model to estimate delays for ICUs beds in New York State. The assumption of a Poisson arrival process is reasonable in this situation because

FIGURE 3. Average occupancy levels of NY State ICUs.



many arrivals to ICUs come through the ED and these arrivals have been shown to be well approximated by this process (see, e.g., Green et al. [14]).

Figure 3 shows the distribution of average occupancy levels for the ICUs in New York State in 1997. The average size of the units in this sample was only 15 beds and the mode was 10 beds and the data show an average occupancy of 75%. Using the traditional criterion of hospital occupancy levels, many of these units would be considered to have "excess" beds.

Because patients needing an ICU bed are emergent and any delay can be dangerous, probability of delay was used as the our metric of performance. This measure was also chosen because it depends only on size and server utilization, which were available from New York State Institutional Cost Reports. Thus, the number of beds needed to meet any delay target could be computed easily.

Adopting a standard of a maximum probability of delay of 10%, 112 of the 194 ICUs, or about 58%, were overutilized. If that target is reduced to 5%, 143 or 74% of the units were too small to handle their experienced workloads; for a 1% target, 175 or over 90% were of inadequate size. These estimates are likely to be conservative for several reasons. First, in an analysis of intensive care units at Beth Israel Deaconess in Boston, the coefficient of variation of the length of stay ranged from 1.1 to 1.6 (Green and Nguyen [15]), suggesting that the $M/M/s$ assumption of exponential service times leads to underestimates of actual congestion. Second, several of the hospitals had multiple divisions or locations, and the reported units are sometimes the sum of two or more smaller units. Third, in many larger hospitals, there are several types of ICUs (e.g., medical, surgical, neurological, and cardiac). Therefore, some of the unit sizes reported in this data likely represents the combined size of smaller, specialized units. It is also important to note that the average reported length of stay in these units was almost 18 days, so that when delays occur, they are likely to be long. This is consistent with observations reported by ED administrators.

It is also important to note that in performing these types of simplistic analyses, the data used are not really reflective of the true demands for the units being studied. For example, because many ICU units are often fully occupied, the hospitals in which they are housed will likely be on "critical adult" diversion on occasion, so their occupancy levels are based on some unmet demands. Another reason this may occur is that when patients are waiting

in the ED for an ICU bed, less seriously ill patients in the ICU unit may be transferred to a "step-down" unit to make room for the more needy new arrivals, hence, reducing the length of stay.

3.2. Using OR to Reduce Delays for Medical Appointments

Many primary care practices have been pressured by managed care companies and lower compensation levels to take on more patients. As a result, many physicians have been wrestling with chronically long backlogs for appointments, leading to patient dissatisfaction, increasing levels of late cancellations, and greater administrative difficulties.

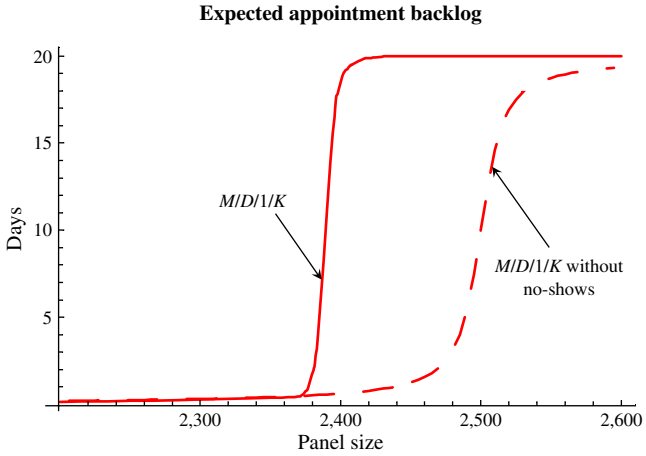
The ability to offer patients timely appointments requires some minimum physician capacity relative to patient demand. In discussions with practitioners, I found that one of the central questions in trying to improve access for patients is, what is a "manageable" panel size? Primary care practices and many specialty care practices, such as cardiology, have a "patient panel"—a set of patients who receive their care from the practice on some regular basis. So, in these practices, the primary lever to bring demand and supply into a relationship that is compatible with being able to offer short appointment dates is patient panel size.

To identify a panel size that will result in short waits for appointments with high probability, it is necessary to explicitly consider the impact of cancellations. Although some patients cancel their appointments far enough in advance of their scheduled time to allow for a new appointment request to be substituted, many practices experience a high level of patients who cancel too late for this to happen or who simply do not show up at the scheduled time. For simplicity, I will refer to both of these patient types as "no-shows." Empirical evidence shows that the rate of no-shows increases with increasing appointment backlogs (Galucci et al. [9]), resulting in more unused appointment slots and wasted physician time. Therefore, instead of increased patient demand rate resulting in higher physician utilization, it may actually result in lower utilization levels. In addition, although some patients fail to appear at the appointed time because the original reason for the visit no longer exists, other no-shows are due to personal or work-related problems, or the patient's decision to seek treatment elsewhere rather than wait. In the latter situations, many no-shows schedule a new appointment with their original physician. This is even true when they have sought treatment elsewhere because it is common practice for clinics and emergency rooms to advise the patient to see their own physician as well.

In Green and Savin [11], we modeled a physician practice appointment system as a single-server queuing system with deterministic service times in which customers who are about to enter service have a state-dependent probability of not being served and, with some specified probability, rejoin the queue. The model's estimates of backlogs for various panel sizes and cancellation behavior were compared with those produced from a more realistic simulation model and the results were shown to be very reliable.

Three important observations resulted from our analyses. First, the impact of cancellations on appointment backlogs can be very significant. This is illustrated in Figure 4, which, using data based on an appointment system for a diagnostic imaging facility, compares the estimates of expected backlog for various patient panels using the queuing model with cancellations as well as the standard $M/D/c$ model. Second, the panel sizes predicted by the model to be consistent with short delays for appointments are smaller than the 2,500-patient level often suggested by managed care companies and in the literature (Murray and Tantau [26]). This figure also highlights another important feature of appointment dynamics in the presence of no-shows: as the panel size increases, the transition to unmanageable backlogs occurs more rapidly, and over a narrower interval of panel size values than in the no-cancellation case. This phenomenon is due to the presence of a positive feedback effect that no-shows exert on the appointment backlog.

FIGURE 4. The impact of no-shows on appointment backlogs.



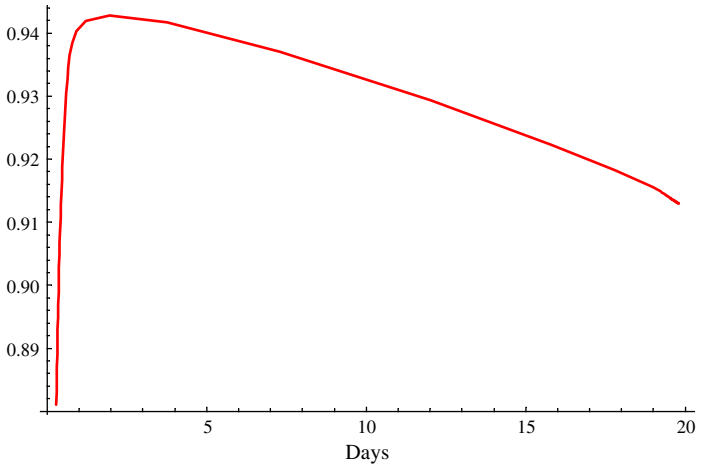
Third, it is possible to identify an optimal patient panel size for a given medical practice. As illustrated in Figure 5, physician utilization initially increases with increasing backlog, as would be expected in a conventional queuing situation. However, because of the cancellation and rescheduling dynamics, physician utilization ultimately decreases as the expected backlog increases. And this occurs at relatively low levels of delay, showing that the right panel size results in both good service for patients as well as high levels of utilization and, hence, revenue for physicians.

In response to a very high level of interest expressed by physicians who have learned about our work in this area, we are now developing an easy-to-use tool that will be accessible on the Web.

3.3. Reducing Delays for Nursing Care

There has been a great deal of discussion and debate in the professional nursing world as well as in state legislatures about how many nurses are really needed in hospitals. Although nursing ratios are still the most common methodology for determining nursing levels, the old guidelines of 8:1 have been challenged with the result that, in 2004, California implemented the first state law mandating minimum nurse-to-patient staffing levels of one to six in general medical-surgical units (California Department of Health Services [6]). Other states are

FIGURE 5. Physician utilization as a function of expected backlog.



contemplating similar legislation. However, there have been many arguments made against the use of mandated staffing ratios (Lang et al. [20]). Opponents of mandated ratios observe that none of the studies of staffing and quality have identified an optimal ratio, and that ratios are too inflexible to account for variation in nursing skills and the severity of patients' illnesses.

From an analytical perspective, a clinical unit can be viewed as a finite source queuing system because demands are generated from the inpatients in that unit. However, due to admissions, discharges, and transfers, the number of inpatients varies over a shift. Furthermore, each of these changes in inpatient census triggers a demand for nursing care. So, to identify appropriate nursing levels, it is necessary to model a single hospital clinical unit as a queuing system with two sets of servers: nurses and beds. Although patients are usually assigned to a specific nurse for each shift, it is common practice for any available nurse to attend to a patient if the designated nurse is busy with other patients. So, it is reasonable to assume that both the nursing system and the bed system are multiserver.

In Yankovic and Green [36], we developed an easy-to-use two-dimensional queuing model for evaluating the impact of any given nursing level on performance measures such as the expected waiting time for a nurse, the probability of being delayed, and other performance indicators. This will enable hospital administrators and nurse managers to determine nurse staffing levels for various units and shifts that best fit their target service standards.

The model that we developed is part of a major project being undertaken in conjunction with the Hospital for Special Surgery in New York City to determine the feasibility and usefulness of using a queuing model to help guide nurse staffing levels in hospitals. In addition to the development of the model, the project focuses on the identification of nursing tasks and the ability to obtain the required input data from both electronic and nonelectronic sources. A specific goal of the project is to inform the development of hospital information technology (IT) systems with regard to operational data that should be routinely collected to inform nurse staffing decisions and evaluate the impact of nurse staffing levels. This will be the first attempt to tie nurse staffing levels to measures of responsiveness to patient needs such as medication orders and call button requests for help.

From a policy perspective, the model can be used to demonstrate the potential problems with using the type of mandated nurse to patient ratios legislated by California. Specifically, our analyses showed that whereas a 1:6 nurse to patient ratio may insure a small probability of delay, i.e., less than 5%, for nursing demands in a busy clinical unit of 60 beds, an otherwise similar unit with only 30 beds would require a minimum of 1:5 ratio to achieve the same delay performance. This is not surprising for anyone who understands that queuing systems exhibit economies of scale. Less obviously, our results showed that the average length of stay can also have an effect of effective nursing levels. In particular, since the admissions and discharge processes require a significant amount of nursing time, units with shorter lengths of stay will require higher staffing levels than those with longer lengths of stay.

Because the model tracks the number of patients in the unit and assumes, as in reality, that a new patient cannot be admitted to the unit unless a nurse is available, the model can be used to estimate the impact of nursing levels on delays for patients waiting in the ED. This is particularly important because nurse availability is often overlooked as a factor affecting ED delays.

4. Obstacles to Using OR to Reduce Healthcare Delays

4.1. Lack of Operational and Performance Data

As mentioned above, hospitals and other healthcare providers have not, in general, collected data on operational performance including delays. So, for example, in trying to evaluate the use of the queuing-based physician staffing levels that were implemented at the Allen Pavilion at NewYork-Presbyterian Hospital, as described in the previous section, we couldn't

compare the effect on arguably the most important performance metric—delay in being treated by a physician. Similarly, although some hospitals do collect the delays for inpatient beds in the ED, they do usually do not collect these by clinical unit or service type. Therefore, there is no easy way to determine which units/services are chronically short on bed capacity.

Although certain types of operational data are routinely kept by hospitals and/or physicians, e.g., arrivals to a facility, lengths of stay in a hospital, lengths of appointment slots, etc., there is a great deal of data that are needed for a model that are not usually not documented. For example, in our ED physician study (Green et al. [14] the time that ED physicians spend with patients was not recorded. In speaking with physician practices we found that information on actual requests for appointments by physician and no-show rates are usually not kept. And one of the biggest challenges to using a queuing model to support nurse staffing is identifying IT systems and processes for collecting data on both requests and durations for nursing care.

4.2. Physician Resistance

Although the resources of a hospital are largely controlled by the physicians who work there, in most hospitals, most physicians are not employees of the hospital. This makes it very difficult to get physicians to adopt new practices. Even if the physicians are employees, hospital administrators are often loath to put pressure on them to change their behavior. This is because the reputation of a hospital is primarily based on the reputation of the physicians who work there and competition for good physicians can be intense, particularly in certain specialties. So, it is important for the hospital to keep physicians happy, and this is reflected in the culture of the hospital. Nonphysicians generally defer to physicians on nonclinical as well as clinical issues.

This physician independence can clearly interfere with efforts to change the allocation or management of resources. An example of this occurred after the implementation of the new ED physician staffing plan at the Allen Pavilion. Under the new plan, the number of ED physicians working the Sunday night shift was reduced from two to one to increase the level of staffing in the late evening on Mondays and Tuesdays. However, within a year of implementing the new plan, staffing reverted to the old pattern because the physician working solo on the Sunday shift was unhappy with that arrangement. Similarly, we have observed on several occasions how physician requests for preferential treatment for specific patients take precedence over any existing schedule for, e.g., imaging, even if there is no clinical urgency associated with the request.

Physician influence is particularly important in issues of inpatient bed use. Physicians can be very territorial about the number and use of beds in their services, particularly in academic medical centers where the teaching and research functions often lead to clinical units or subunits being restricted to relatively small categories of patients.

4.3. Lack of Cost/Revenue Information

Athough the mission of hospitals and other health delivery facilities is to provide care for patients in an effective and timely manner, these facilities are under enormous cost pressures from managed care and government payers. Therefore, it is not surprising that efforts are focused on cutting costs and increasing revenues. This often leads to myopic behavior, which may endanger patients and even be suboptimal for the facility. For example, many hospitals view their emergency departments as cost centers because they are very expensive to operate and the compensation provided by Medicare and insurance companies does not cover the costs. This may be compounded by a high level of uninsured patients using the ED for whom the hospital may never receive any payment. Therefore, it is not surprising that many hospitals are not very motivated to shorten delays for ED care. However, many hospitals who have analyzed the financial implications of high levels of patients who LWBS

and ambulance diversions due to long ED delays have come to the conclusion that these can lead to significant loses of revenue for two reasons: (1) 25% to 50% of patients arriving to the ED are admitted to the hospital and the revenues generated from their stays can easily compensate for the costs of ED care, and (2) patients who experience long ED delays may not return to the hospital when they have elective surgery.

4.4. "Toxic" Payment Systems

Hospital and physician compensation are largely based on Medicare reimbursement systems, which have historically favored procedures over medical care. So, even at present, services like cardiac and orthopedic surgery are far more profitable than care for, e.g., pneumonia or asthma. For this reason, hospitals try to admit as many surgical patients as possible and often engage in marketing efforts to attract these patients. Because many hospitals have general medical/surgical units, the greater the fraction of surgical patients, who are generally elective, the fewer beds are available for patients admitted from the ED. Recent cuts in Medicare payments to hospitals have resulted in an even greater incentive for hospitals to focus on elective surgical patients, for whom they have a positive operating margin, as opposed to medical patients who often lead to operating losses. This is one of the major underlying reasons that ED overcrowding has become a growing problem.

A similar situation exists for physician practices. Most physician payment systems reimburse based on the volume of patient visits. This encourages physicians to generate patient visits for issues that could be handled by phone or e-mail exchanges, which are generally not compensated and, thus, can lead to longer delays for appointments.

4.5. Estimating Demands

One of the biggest obstacles to using OR methodologies to reduce delays for healthcare is in identifying the "true" demand for healthcare services. It is well known in the healthcare world that the demand for healthcare is distorted by several factors. First, because much of healthcare is paid for by third parties rather than the consumers themselves, demand is very much influenced by the size of deductions and copayments. A famous study by Rand Health [29] showed that people who paid a larger share of their healthcare used fewer services, but in general, had the same health outcomes. Demand has also been shown to be highly dependent on supply. For example, cities with more hospital beds per capita have higher rates of hospitalization. Physician practice patterns also significantly influence the use of healthcare resources (Wennberg et al. [34]). These studies illustrate the difficulty of estimating and predicting demands for healthcare services and, hence, the capacities needed to serve a given population.

5. Summary and Conclusions

5.1. Opportunities

Healthcare managers and physicians are increasingly aware of the need to use their resources as efficiently as possible to continue to assure that their institutions survive and prosper. This is particularly true in light of the growing demands for healthcare due to an increase in older patients and the prevalence of chronic diseases, and a continuing threat of sudden and severe demand surges due to outbreaks of epidemics such as SARS and avian flu, or terrorist incidents. Effective resource 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 healthcare managers to control the cost and improve the quality of healthcare 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.

The growing capabilities and installation of healthcare IT systems in hospitals and healthcare practices are making it increasingly possible to collect and organize the data needed to create and implement OR methodologies. However, OR professionals are needed to help guide the development of these systems so that the appropriate data will be identified and processes developed to collect and use these data to fuel the methodologies. Sometimes, the process of conducting a study in a healthcare facility can lead to the needed IT capabilities. This happened subsequent to our ED physician staffing project (Green et al. [14]) where, as a result of our study, the hospital implemented a new electronic data collection system for capturing physician delays.

5.2. Some Caveats

As described above, it is essential to be aware of the organizational structure, culture, and constraints when trying to develop and implement OR models in healthcare delivery systems. This requires a general understanding of the healthcare system and the financial, human resource and regulatory environment in which it operates. It also requires close working relationships with the physicians, nurses, and other professionals who will be affected by the proposed methodology to fully appreciate all of the factors relevant to structuring the model, estimating the parameters, and developing performance objectives.

This does not mean that an OR model need be complex to be useful. To the contrary, sometimes a simple model can lead to important insights. An example of this is the use of a standard Markovian queuing model to gain insights about the bed capacity needs of intensive care units. Similarly, even though the noncontinuous nature of physician interactions with patients in the ED did not fit the assumption of exponential service times that was used in the study of ED staffing, and the staffing changes suggested by the analyses could not be implemented exactly because of shift constraints, the insights about the relative capacity needs of different days of the week and different times of days led to new staffing patterns that improved service.

In fact, there is reason to believe that a very simple-to-use OR-based tool has the greatest potential for widespread adoption. An article describing a spreadsheet-based simple binomial model for estimating patient panel sizes for medical practices was published in a widely read medical journal (Green et al. [13]) and subsequently featured on the *Business Week* website (Miller [23]) and the newsletter of the American Medical Association (Stevens [31]). The result was hundreds of e-mails to the authors requesting the spreadsheet from various types of practices, clinics, and hospitals across the country. This reveals that there is a significant level of interest in using models to improve patient service, particularly if they are easily accessible.

5.3. The Bigger Picture

Healthcare delivery systems are increasingly the subject of political discussions and debates as the United States and other countries struggle to reduce costs and improve quality and access. Much of these discussions center around whether there are too many or too few healthcare resources such as inpatient beds, physicians, nurses, and imaging facilities. Unfortunately, the politicians and policy professionals are often guided, or misguided, by simplistic frameworks and guidelines that totally inappropriate and potentially dangerous. This was dramatically demonstrated in 1987–1988 when New York City experienced a severe and protracted citywide shortage of inpatient hospital beds. During this period, ambulances were routinely turned away from full hospitals and urgently sick patients experienced delays

of days waiting for an open bed. This hospital crisis followed a two-year period in which capacity declined by 9% because of new state regulations linking Medicaid reimbursement to achieving occupancy levels of at least 85% to reduce the number of "excess" beds in the city [36].

Operations research can and should play a role in these critical discussions and decisions regarding capacity needs and uses. To achieve this, operations researchers need to develop and/or use models that can illustrate the potential adverse consequences of existing or proposed policies, such as hospital closings or mandated nursing ratios, and publish and publicize their findings in medical journals, newsletters, and newspapers. The potential impact can literally make a difference of life or death for many people.

References

[1] L. A. Aiken, S. P. Clarke, D. M. Sloane, J. A. Sochalski, and J. H. Silber. Hospital nurse staffing and patient mortality, nurse burnout and patient satisfaction. *Journal of the American Medical Association* 288:1987–1993, 2002.

[2] American Hospital Association. Survey of hospital leaders. American Hospital Association, Washington, D.C., 2007.

[3] American Hospital Association. Trendwatch Chartbook. Trends affecting hospitals and health systems. American Hospital Association, Washington, D.C., 2008.

[4] American Medical Association. Physician characteristics and distribution in the U.S. American Medical Association, Chicago, 2005.

[5] D. W. Baker, C. D. Stevens, and R. H. Brook. Patients who leave a public hospital emergency department without being seen by a physician. Causes and consequences. *Journal of the American Medical Association* 266:1085–90, 1991.

[6] California Department of Health Services. Final statement on reasons hospital nurse staff ratios and quality of care. California Department of Health Services, Sacramento, CA, 2003.

[7] Commission on Healthcare Facilities in the 21st Century. A plan to stabilize and strengthen New York's healthcare system. Commission on Healthcare Facilities in the 21st Century, New York, 2006.

[8] C. M. Fernandes, A. Price, and J. M. Christenson. Does reduced length of stay decrease the number of emergency department patients who leave without seeing a physician? *Journal of Emergency Medicine* 15:397–399, 1997.

[9] G. Galucci, W. Swartz, and F. Hackerman. Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services* 56:344–346, 2005.

[10] L. Green. How many hospital beds? *Inquiry* (Winter), 400–412, 2003.

[11] L. Green and S. Savin. Reducing delays for medical appointments: A queueing model approach. *Operations Research*. Forthcoming.

[12] L. Green, P. J. Kolesar, and J. Soares. Improving the SIPP approach for staffing service systems with cyclic demand. *Operations Research* (July–August), 549–464, 2001.

[13] L. Green, S. Savin, and M. Murray. Providing timely delivery of care: What is the right panel size? *Joint Commission Journal on Quality and Patient Safety* (April), 211–218,2007.

[14] L. Green, J. Soares, J. Giulio, and R. Green. Using queueing theory to increase the effectiveness of physician staffing in the emergency department. *Academic Emergency Medicine* (January), 61–68, 2006.

[15] L. V. Green and V. Nguyen. Strategies for cutting hospital beds: The impact on patient service. *Health Services Research* 36:421–442, 2001.

[16] R. A. Green, P. C. Wyer, and J. Giglio. ED walkout rate correlated with ED length of stay but not with ED volume or hospital census (abstract). *Academic Emergency Medicine* 9:514, 2002.

[17] J. S. Groeger, K. K. Guntupalli, M. Strosberg, N. Halpern, R. C. Raphaely, F. Cerra, and W. Kaye. Descriptive analysis of critical care units in the United States. *Critical Care Medicine* 21:279–291, 1992.

[18] Healthgrades. Third Annual Patient Safety in American Hospitals Study. Healthgrades, Golden, CO, 2006.

[19] Institute of Medicine of the National Academies of Science. To err is human. Report, Institute of Medicine of the National Academies of Science, Washington, D.C., 1999.

[20] T. A. Lang, M. Hodge, V. Olson, P. S. Romano, and R. L. Kravitz. Nurse-patient ratios: A systematic review on the effects of nurse staffing on patient, nurse employee, and hospital outcomes. *Journal of Nursing Administration* 34:326–337, 2004.

[21] L. F. McCaig and C. W. Burt. National Hospital Ambulatory Medical Care Survey: 2002 Emergency Department Summary. Advance Data from Vital and Health Statistics—CDC: National Center for Health Statistics, Hyattsville, MD, 340:1–36, 2004.

[22] W. McClure. Reducing excess hospital capacity. Bureau of Health Planning, Rockville, MD, 1976.

[23] K. Miller. An open-access doctor's office. *Business Week Online* (February 12), http://www.businessweek.com, 2007.

[24] R. Mishra. State's patients endure long wait. *The Boston Globe* (June 7), 2005.

[25] M. Murray and D. Berwick. Advanced access: Reducing waiting and delays in primary care. *Journal of the American Medical Association* 289:1035–1040, 2003.

[26] M. Murray and C. Tantau. Same-day appointments: Exploding the access paradigm. *Family Practice Management* 7(8):45–50, 2000.

[27] L. P. Myers, K. S. Fox, and B. C. Vladeck. Health services research in a quick and dirty world: The New York city hospital occupancy crisis. *Health Services Research* 25:739–755, 1990.

[28] J. Needleman, P. Buerhaus, S. Mattke, M. Stewart, and K. Zelevinsky. Nurse-staffing levels and the quality of care in hospitals. *New England Journal of Medicine* 346:1715–1722, 2002.

[29] Rand Health. Research Highlights, The health insurance experiment. http://www.rand.org/pubs/research_briefs/RB9174/index1.html, 2006.

[30] M. Smoller. Telephone calls and appointment requests: Predictability in an unpredictable world. *HMO Practice* 6(2):25–29.

[31] L. Stevens. Calculate your ideal patient load: How to strike the correct balance. *AMNEWS (American Medical Association News* (May 12), 2008.

[32] B. C. Strunk and P. J. Cunningham. Treading water: Americans' access to needed medical care, 1997–2001. Center for Studying Health System Change, Washington, D.C., 2002.

[33] U.S. Department of Health and Human Services. The registered nurse population: Finding from the 2000 National Sample Survey. National Center for Health Workforce Analysis, Bureau of Health Professions, Health Research and Services Administration, U.S. Department of Health and Human Services, Washington, D.C., http://bhpr.hrsa.gov/healthworkforce/reports/rnsurvey/default.htm, 2004.

[34] J. E. Wennberg, E. S. Fisher, and J. S. Skinner. Geography and the debate over Medicare reform. *Health Affairs Web Exclusive*, http://content.healthaffairs.org/cgi/content/full/hlthaff.w2.96v1/DC1, 2002.

[35] A. P. Wilper, S. Woolhandler, K. E. Lasser, D. McCormick, S. L. Cutrona, D. H. Bor, and D. U. Himmelstein. Waits to see an emergency department physician: U.S. trends and predictors, 1997–2004. *Health Affairs Web Exclusive* 27:w84–w95, 2008.

[36] N. Yankovic and L. Green. A queuing model for nurse staffing. Working paper, Columbia Business School, New York, 2008.

[37] N. Yankovic, S. Glied, M. Grams, and L. V. Green. Ambulance diversion and myocardial infarction mortality. Working paper, Columbia Business School, New York, 2008.