Improving Access to Healthcare: Models of Adaptive Behavior

Carri W. Chan and Linda V. Green

Abstract : Patient access to healthcare is a major problem area due to inadequate supplies and misallocation of resources including physicians, nurses, and hospital beds. Increasing patient demands due to an aging and more chronically ill population will exacerbate this situation, leading to longer delays for care, hurried treatment times, and adverse clinical outcomes. Though there is a significant operations literature focused on methods to mitigate these effects, suggested remedies may be ineffective due to adaptive behavior by both physicians and patients. This chapter will focus on the quantification and impact of such adaptive behavior on the ability to provide timely patient access to limited health services.

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

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

1 Introduction

Demand for healthcare is increasing due to a growing and aging population, making access to care more difficult. Beyond anecdotal evidence, there is increasing empirical evidence of access problems, most notably through overcrowding in Emergency Departments (EDs) [6, 12]. While demand is increasing, the supply of hospital beds, physicians, nurses, and other health resources remains relatively stagnant or, worse, is potentially decreasing. It is already the case that the supply of nurses is insufficient to meet demands [8] and there are predictions of severe physician shortages in the coming years [13, 34, 41].

As a consequence of high demand and insufficient supply, many patients experience delays in receiving treatment. The overall median wait to see an ED physician increased from 22 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 [44]. In one study of patients and their primary care physicians, 33% of patients cited inability to get an appointment soon as a significant obstacle to care [42]. The average wait for a primary care appointment in the U.S. in 2001 was over three weeks [36]. 60% of physicians reported being dissatisfied with delays [37].

Delays can result in adverse patient outcomes such as increased mortality rates and an overall reduction in quality of outcome [39]. For emergent patients, such as those suffering acute myocardial infarction, timely access to care is imperative as even delays on the order of minutes can increase mortality [32, 11, 5, 23]. Delays can also result in increased length-of-stay (LOS), resulting in patients consuming more resources and further intensifying the problem. For example, delays in transfers from the ED to the Intensive Care Unit (ICU) have been shown to increase ICU and hospital LOS [9, 38, 40].

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

There are a number of behavioral factors in the healthcare setting which exacerbate access problems. One such factor is planned variability in capacity due to physician preferences. For example, surgeons often have significant ability to influence their own operating schedules. Most surgeons prefer operating in the morning so they can see new patients in the afternoon. This often results in surgeries being scheduled within a tight time window without adequate attention to the variability of their durations. Not surprisingly, many surgeries get delayed and recovery rooms get congested causing cancelations of subsequent surgeries. Since inpatient beds are often reserved for surgical patients, these surgical delays can translate into ED congestion due to the inability to move ED patients into inpatient beds. In one noted hospital study, the level of ambulance diversions (ambulances turned away from the ED) was better correlated with the variability in the *scheduled* surgical load than with emergency admissions [33]. While some variability in the surgical schedule is certainly unavoidable, there is potential to utilize better scheduling of elective admissions to smooth load variability [30]. For instance, using stochastic linear programming, Denton et al. consider how to assign surgeries to various specialties and how to determine the number of operating rooms (ORs) to open given unavoidable uncertainty in the duration of various surgeries [14]. Golden et al. use integer programming methods to improve scheduling the OR and reduce boarding of patients in the post-anesthesia recovery room due to ICU congestion [21]. In fact, there has been considerable operations literature dealing with surgical scheduling, see [7] and related references.

In this chapter, we will focus on a distinctive and prevalent characteristic of healthcare delivery systems–adaptive behavior. There has been growing evidence that patients and providers dynamically alter their behavior based on congestion and backlogs. These adaptive behaviors have been observed in both outpatient and inpatient settings. For instance, if patients have to wait a long time for an appointment with a physician, they may cancel at the last minute or just not show up [20]. When delays in the ED are long, patients are more likely to leave without being seen, even though they require care [19]. Hospital EDs sometimes adapt to increasing backlogs by diverting ambulances away from the ED, effectively reducing patient arrivals and ED load [29]. Though some of these behaviors may reduce the system workload, some adaptive behavior may actually worsen the situation. In one study of a hospital ED, nurses were found to be more likely to not show up for work when the anticipated patient load was higher, creating an even larger imbalance between supply (nurses) and demand (patients) [24].

In the inpatient environment, providers are often faced with the difficult task of caring for more patients than their resources allow and, hence, adopt practices to attempt to mitigate these high stress periods. For instance, physicians may discharge patients early from an ICU when it is full and space is needed for new patients [26]. If there is no room in a hospital stroke unit at the time of a stroke patient's arrival, the patient may be placed in a less specialized unit which could result in a longer LOS and a poorer clinical outcome [46]. Indeed, patients are often assigned to less appropriate clinical units due to congestion in the desired unit.

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

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

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

2 Quantifying the Impact of Adaptive Behavior

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

2.1 Empirical Evidence: Adaptive Behavior of Patients

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

There are a growing number of healthcare practices and outpatient facilities that operate on an appointment basis. One of the difficulties faced by these facilities are patients who make last-minute cancelations or fail to arrive to their scheduled appointments. These patients are classified as 'no-shows'. No-shows often waste already limited physician availability since it is usually impossible to fill a last minute cancelation with another patient. This can result in significant monetary losses (up to 14% of annual revenues) for the clinic [35]. **[Note to editor: Reference chapter on no-shows]**

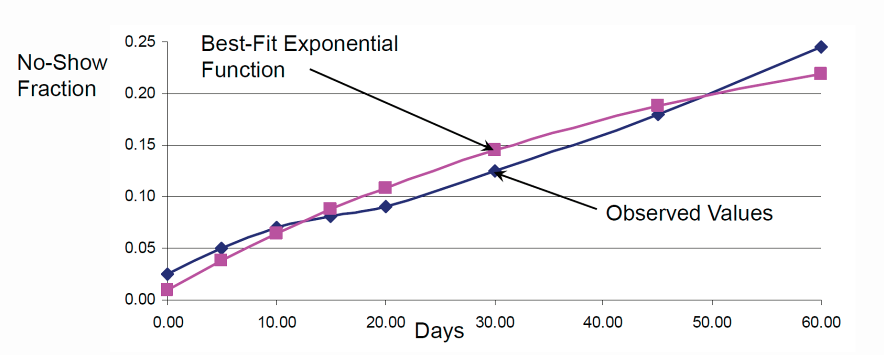

Fig. 1 Observed no-show fraction values and the best-fit exponential functions for Columbia MRI data as reported in [22] as reported in [20]. Reprinted by permission, L. V. and S. Savin, Reducing Delays for Medical Appointments: A Queueing Approach, Operations Research, volume 56, issue 6, (November/December, 2008). Copyright 2008, the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 300, Hanover, MD 21076 USA.

Empirical evidence has shown that the rate of no-shows increases with the increase in appointment backlog, i.e. the longer a patient has to wait until their appointment, the more likely he is to fail to show up. This phenomenon was observed at a mental health clinic, an MRI facility, and a family practice clinic [20, 22, 31]. In [22], data on the connection between the appointment backlog and the likelihood of a patient no-show from both a mental health clinic and an imaging facility were fit to an exponential function as depicted in Figure 1. The percentage of no-shows is monotonically increasing in backlog in these two independent data sets, though the rates are quite different, as would be expected with such different patient characteristics across the two facilities.

In another setting, there has been growing evidence that increased crowding in the ED has resulted in an increase in patients who leave the ED without being seen [15, 25, 19]. This often means that patients who require care do not have the access they need [1].

2.2 Empirical Evidence: Adaptive Behavior of Physicians

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

In [27, 26], Kc and Terweisch examine the relationship between occupancy levels and patient care. Using data from an ICU unit for cardiothoracic surgery patients, they demonstrate that, after controlling for patient severity, a patient's LOS decreases as the unit occupancy level increases. This is illustrated in Figure 2. This supports anecdotal reports that patients are sometimes discharged prematurely in order to accommodate new, more critical patients. We refer to such a discharge as a *demand-driven* discharge.

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

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

In this section, we will focus on the first two questions; the third will be addressed in Section 4. In addition to the effect of high occupancy levels on patient LOS it may also affect patient readmission likelihood due to the early discharge of patients.

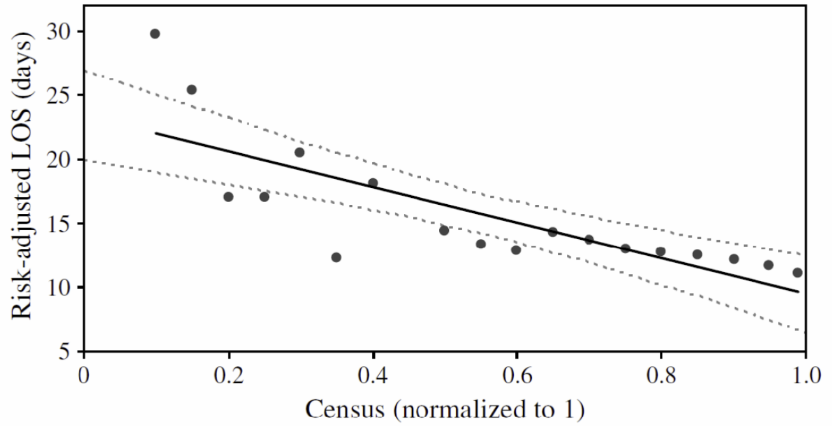

Fig. 2 Length of Stay as a Function of Census. Note: Census is defined as the number of patients in the cardiac unit at the time a patient is admitted. Length of stay (LOS) is the total number of days a patient spends at the hospital. Dashed lines represent 95% confidence intervals. Reprinted by permission, D. Kc and C. Terweisch, Impact of workload on service time and patient safety: An econometric analysis of hospital operations, Management Science, volume 55, issue 9, (July, 2009). Copyright (2009), the Institute for Operations Research and the Management Sciences (INFORMS), 7240 Parkway Drive, Suite 300, Hanover, MD 21076 USA

However, there is an inherent endogenity bias, since more severe patients are likely to have longer length-of-stay in the ICU and have higher readmission risks, which could lead to a positive bias in estimating the effect of LOS on readmissions. The exogenous factors affecting LOS–variables that affect the time spent in the ICU, but otherwise do not directly affect patient outcomes–constitute potential instrumental variables (IVs) to mitigate the endogeneity bias. In particular, an indicator variable which specifies whether or not the ICU is busy (i.e. at high occupancy levels) upon discharge becomes a valid IV (see [45] for details on this methodology). Because operational factors are unlikely to be correlated with patient medical factors, such as severity, which may affect patient outcomes, they can often be used as instrumental variables to generate unbiased estimates of these outcomes.

In a study of cardiac surgical patients in a single hospital, a 10% increase in occupancy level corresponded to a 20% decrease in ICU LOS–a reduction of nearly 2.5 days [26]. This shortened length-of-stay corresponded to increases in the likelihood of readmission. Specifically, being discharged one day earlier than one's expected LOS translated to an increase of 60% in the odds of being readmitted to the ICU [27]. This modification of patient LOS due to congestion may initially free capacity in the ICU, but it can also negatively impact patient outcomes in the long run.

2.3 Quantifying Effects via Modeling

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

Using the ICU described as an example, the first step is to build a stochastic model of an ICU which incorporates the fact that patients may be demand-driven discharged. Such a model can be used to consider how changes in arrival patterns, ICU capacity, and surgical schedules can affect the likelihood of being discharged early.

In [17], it is assumed that patients are either scheduled or unscheduled. The state of the system is given by the remaining length-of-stay of the patients who occupy the ICU. If a new patient arrives and there is no space available, a current patient is demand-driven discharged to accommodate the new patient. Using an aggregation-disaggregation technique to reduce computational complexity, Dobson et al. calculate the desired performance metrics such as the probability of being *bumped* and the expected number of days remaining when a patient is bumped.

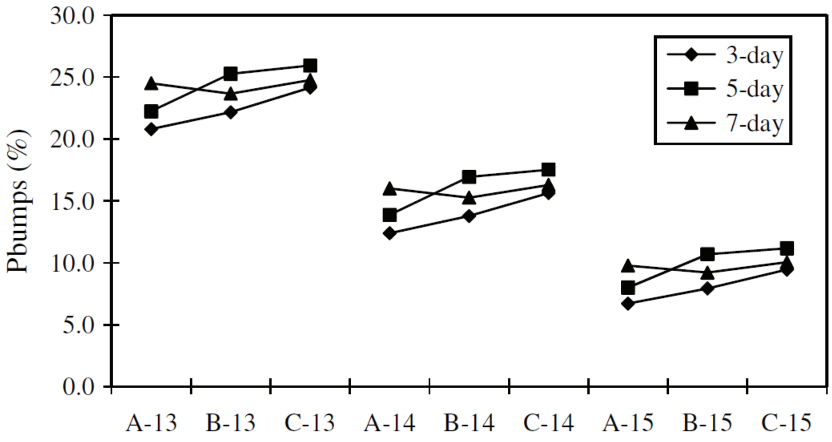

Fig. 3 Average probability of being demand-driven discharged for A) 70% B) 50% and C) 30% scheduled patients and an ICU of size 13, 14, and 15 beds as reported in [17]. Reprinted by permission, G. Dobson, H.-H. Lee, E. Pinker. A model of ICU bumping. Operations Research, volume 58, issue 6, (November/December, 2010). Copyright (2010), the Institute for Operations Research and the Management Sciences (INFORMS), 7240 Parkway Drive, Suite 300, Hanover, MD 21076 USA.

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

3 Incorporating Adaptive Behavior into Decision-Making

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

3.1 Accounting for no-shows when determining patient panel size

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

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

In [22], Green et al. model a single-physician practice via a modified M/D/1/K queue where patients arrive according to a Poisson process, service times are deterministic, and there is a finite appointment backlog limit K such that any patients who arrive when the queue length is K are ‘lost’ in the sense that they are not given an appointment and so potentially seek treatment elsewhere. For more information on the standard M/D/1/K queue, we refer the reader to examine a book on queueing, such as the one by Kleinrock [28]. The modified M/D/1/K model approximates the no-show process by assuming that a customer who is scheduled to begin service has a state dependent probability of being a no-show, resulting in an idle period for the server and, with a fixed probability, the customer rejoining the queue. The likelihood of no-show is non-decreasing in the number of patients who are still in the backlog upon the appointment (i.e. service) time of the patient in question. Such an approximation is able to capture wasted capacity by patient no-shows as well as the increased likelihood of such events when the system is more congested. Additionally, it allows for analytical tractability of the steady-state behavior of the system.

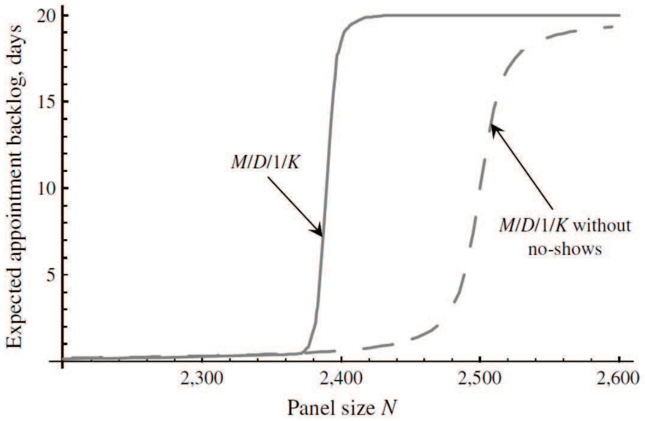

Fig. 4 Expected appointment backlog as a function of the patient panel size for the M/D/1/K model with and without no-shows (using a no-show model based on an MRI facility data, assuming 20 slots per day, K = 400 appointment slots and a probability of rescheduling equal to 1) as reported in [22]. Reprinted by permission, L. V. and S. Savin, Reducing Delays for Medical Appointments: A Queueing Approach, Operations Research, volume 56, issue 6, (November/December, 2008). Copyright 2008, the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 300, Hanover, MD 21076 USA.

Figure 4 compares the expected appointment backlog of the M/D/1/K queue with and without no-shows. Using a no-show model calibrated from data of a MRI facility, one can see that the impact of patient no-shows is very significant [22]. Since no-shows result in wasted appointment slots and rescheduled appointments, they result in longer appointment backlogs and hence more no-shows. Thus there is a adverse feedback cycle and the backlog grows much more rapidly than in a model without no-shows. Ignoring no-shows in a model of a clinic may result in a physician electing to maintain a panel size that is too large to provide timely access to care for his/her patients.

An increasingly important performance metric for access in this setting is the probability of being able to get a same-day appointment. In fact, 33% of patients reported that the 'inability to get an appointment soon' inhibited access to care [42]. Figure 5 compares the probability of getting a same-day appointment with and without no-shows. If no-shows are not considered, the figure suggests that a panel size of 2,400 would provide timely access since patients will be able to get a same-day appointment 80% of the time. (This performance level would be consistent with data that suggests about 20% of appointments are for follow-up care and are scheduled weeks in advance.) However, in actuality, this probability is likely to be close to

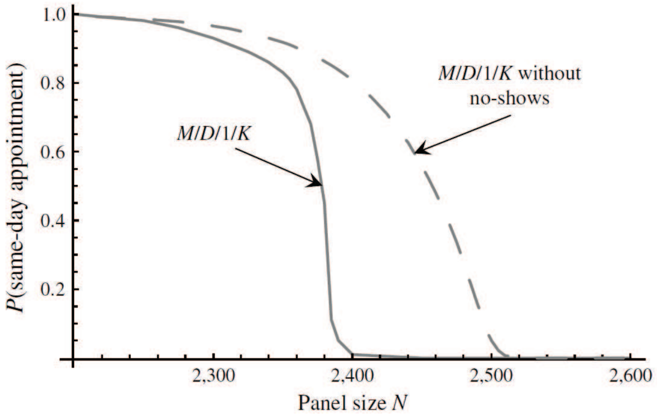

Fig. 5 Probability of getting a same-day appointment as a function of the patient panel size for the M/D/1/K model with and without no-shows (using a no-show model based on an MRI facility data, assuming 20 slots per day, K = 400 appointment slots and a probability of rescheduling equal to 1) as reported in [22]. Reprinted by permission, L. V. and S. Savin, Reducing Delays for Medical Appointments: A Queueing Approach, Operations Research, volume 56, issue 6, (November/December, 2008). Copyright 2008, the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 300, Hanover, MD 21076 USA.

0, due to the no-show phenomenon. Such an analysis highlights the importance of accounting for the adaptive behavior of patients when making operational decisions.

4 Dynamic Policies which Account for Adaptive Behavior

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

4.1 *Accounting for no-shows when scheduling patient appointments*

An important aspect of outpatient clinic management is scheduling patients as they call for appointments. As with panel size planning, patient no-shows are an important factor in crafting schedules so that the number of idle appointment slots is minimized. In [31], Liu et al. analyze a dynamic scheduling model which captures this no-show phenomenon. Each day, the appointment scheduler must determine which day to assign to each patient who calls for an appointment. The longer a patient waits for an appointment, the more likely she is to cancel or be a no-show. On any given day, a scheduled patient can show up or not show up for her appointment that day, or cancel an appointment which may be on a future date.

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

Two simple heuristics are to optimize the scheduling policy assuming appointments depending on the how quickly a patient must be seen [31]. The first heuristic, referred to as **Open Access**, requires appointments to be provided on the current day. Hence, patients are guaranteed same-day appointments, even if this requires the physician to spend significant overtime to treat all patients beyond the initial allocated daily capacity. Another heuristic, the **Two-day policy**, requires an appointment to be provided on the current or following day. In doing so, it tries to reduce overtime costs at the expense of immediate access.

These heuristics can serve as the basis for additional heuristics using policy improvement. The policy improvement heuristic selects the best scheduling decision in the current state under the assumption the suboptimal base policy is used for all subsequent decisions. Hence, it is a one-step policy improvement over the base heuristic (see [3] for more details on this methodology). Finally, these heuristics are compared to the following benchmarks: a **Threshold heuristic** where patients are scheduled on the earliest day with fewer than M patients scheduled; a **Load balancing heuristic** where patients are scheduled on the day with the fewest appointments; and a **Random heuristic** where patients are scheduled on a random day.

Using data calibrated from empirical data of a family medicine clinic, [31] compares the performance of the proposed algorithms via simulation. Table 4.1 summarizes the relative costs of the various heuristic policies in comparison to the Open Access scheduling policy for various daily capacities (M) and cost of scheduling one patient (h). Note that the number of patients scheduled in a day, z, can be greater or less than M. If there are more patients than appointment slots ($z > M$) all of these patients will be treated and overtime cost is incurred. This is in contrast to the K in the M/D/1/K model of Section 3 as in that case, overtime was not allowed. Smaller M suggests the clinic is more overloaded. The simulations suggest that the threshold heuristic, two-day policy, and policy improvement heuristics based on Open Access and the two-day policy generate more revenue compared to Open Access. Interestingly, even the Random heuristic sometimes outperforms Open Access. In an under loaded system, Open Access would be optimal. Under Open Access, all patients are scheduled on the current day, which minimizes their likelihood of being a no-show. However, as the practice becomes more heavily loaded, this will result in frequent overload, requiring physicians to work overtime, thus incurring high costs. Hence, Open Access is not the best scheduling policy to use in general.

		PI 2-day	2-day	PI Open Access	Threshold	Load Balancing	Random
$M = 55$	$h = 0$	2.11	.78	2.18	2.11	-6.30	-3.28
	$h = .2$	4.10	3.23	3.08	3.25	-5.48	-1.53
	$h = .5$	12.74	12.14	3.72	4.39	-5.48	2.68
$M = 50$	$h = 0$	6.77	2.75	5.42	6.45	-2.22	-1.20
	$h = .2$	8.28	5.48	6.96	8.21	-1.09	0.50
	$h = .5$	18.56	15.29	9.25	12.11	0.72	5.31
$M = 45$	$h = 0$	10.63	6.23	9.25	5.24	4.11	1.81
	$h = .2$	13.35	9.13	11.53	6.28	4.91	3.28
	$h = .5$	25.01	20.32	21.78	10.40	8.10	9.16
$M = 40$	$h = 0$	9.84	9.12	10.21	2.79	2.99	4.23
	$h = .2$	13.03	12.48	13.69	3.57	3.83	6.05
	$h = .5$	27.41	26.82	28.13	6.79	7.32	13.58

Table 1 Simulation Study Results: percentage improvement of total reward (number of patients served less daily fixed cost and scheduling cost) compared to Open Access scheduling for daily capacity M and cost h for scheduling a patient as reported in [31]. Adapted with permission from N. Liu, S. Ziya, V. G. Kulkarni, . Dynamic Scheduling of Outpatient Appointments Under Patient No-Shows and Cancellations, Manufacturing & Service Operations Research volume 12, issue 2, (October, 2010). Copyright (2010), the Institute for Operations Research and the Management Sciences, 7240 Parkway Drive, Suite 300, Hanover, Maryland 21076 USA.

4.2 *Improving demand-driven discharge decisions*

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

Various factors can affect how often demand-driven discharges must occur [17]. A natural question is how one should determine which patient to discharge. Section 2.3 assumed that patients were discharged in order of shortest remaining LOS. However, given the natural variability in patient stays, this quantity is not always known. Additionally, it ignores the potential impact of readmissions on ICU congestion. In [10], a dynamic optimization model is developed to help guide such decisions.

The model in [10] assumes that whenever a new patient arrives and there are no available beds, a physician must decide which patient to discharge in order to accommodate the new, higher acuity patient. Each patient is identified by type, which specifies the expected initial ICU length of stay, the likelihood of readmission upon a demand-driven discharge, and the expected ICU length-of-stay upon readmission. Any cost function which accounts for a patient's disservice due to a demand-driven discharge can be incorporated. A cost function which is estimable from currently available data is given by the *readmission load*, i.e. expected ICU treatment time required by the demand-driven discharged patient following the initial discharge.

In principle the optimal policy can be computed numerically via dynamic programming. Unfortunately, the size of the state space makes it practically infeasible to solve. Utilizing properties of the optimal value function, the authors show that the performance of a greedy heuristic, which discharges the patient with the lowest readmission load is a $(\hat{\rho}+1)$-approximation for the optimal policy, where $\hat{\rho}$ is a measure of utility [10]. Such a bound is useful to quantify the worse-case performance of such a greedy policy.

Using patient data from 7 different hospitals in a single hospital network, Chan et al. simulate the performance of the greedy discharge policy relative to several relevant benchmarks. In the medical community, the decision of which patient to discharge is made by assessing which patient is the 'least critical' (see, for instance, [43]) which can be somewhat subjective and is generally not based upon quantitative measures. Each of the discharge policies studied below can be interpreted as a measure of criticality:

- **Probability of readmission index:** Discharge the patient with the smallest probability of readmission. Readmitted patients tend to be more critical (see [18]), so that the rationale here is that a lower likelihood of readmission translates to lower patient criticality.
- **Length-of-stay (LOS) index:** Discharge the patient with the smallest remaining service time. This policy thus equates criticality with the nominal length-of-stay of a patient. This policy is analyzed in [17] albeit for a model that is agnostic to readmission loads.

- **The Greedy index:** This is the proposed heuristic from [10] which prioritizes patients in increasing order of readmission load.

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

Figure 6 compares the readmission load of the greedy heuristic compared to the other benchmark policies. The savings relative to the next best policy corresponds to 23.7 hours over one week at a net patient arrival rate of $\lambda = 0.021$ (or 1 ICU bed out of 10 for 1 day per week). Figure 7 shows the number of deaths per week for the same discharge policies. One can see that the number of deaths is practically identical for all policies, while the readmission load is very different. Hence, without sacrificing patient quality, in terms of mortality, the greedy heuristic which incorporates readmission risks can significantly reduce the patient load on the ICU, and subsequently increase the number of patients who receive critical care.

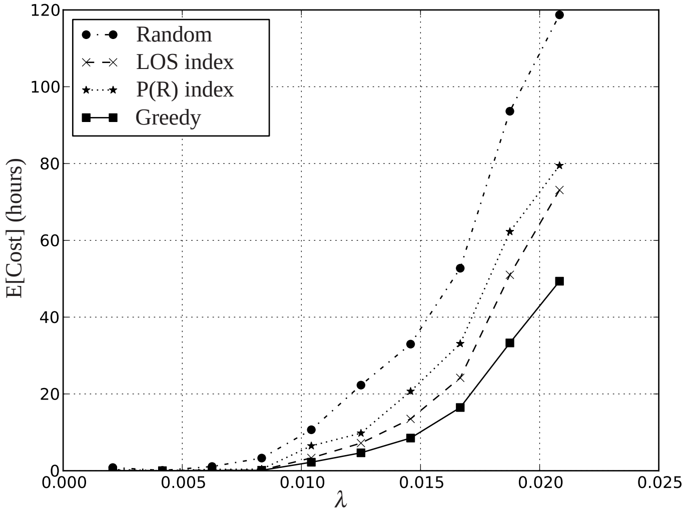

Fig. 6 Performance of greedy policy compared to benchmarks for various arrival rates and distribution across patient types according to the proportions seen in the empirical data as reported in [10].

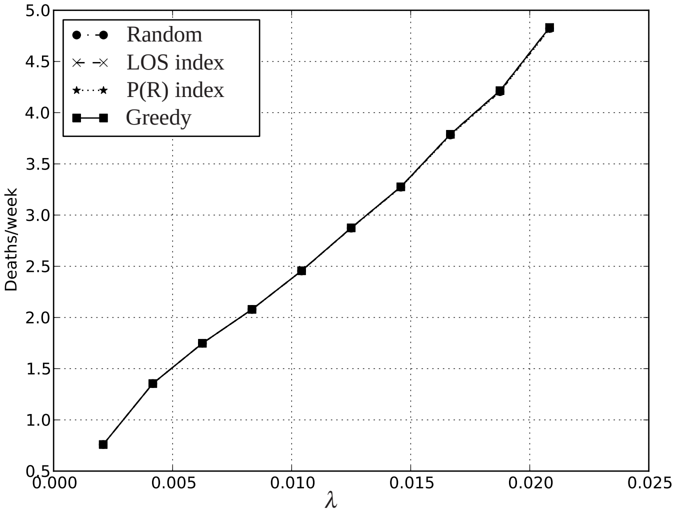

Fig. 7 Number of deaths for greedy policy compared to benchmarks for various arrival rates and distribution across patient types according to the proportions seen in the empirical data as reported in [10].

5 Conclusions and Future Research

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

One consequence of adaptive behavior is that the true service requirements and arrival rates of patients may be censored. These potentially misleading measurements of the system load make it difficult to assess the actual required capacity requirements. Due to the chronic mismatch of supply and demand, much of the observed behavior of healthcare systems do not accurately reflect the true dynamics. Hence, there is a need to analyze the impact of adaptive behaviors on patient demands and treatment times when estimating required capacity for many healthcare resources (e.g., ICU beds, obstetrics beds, surgical suites, primary care physicians, nurses). For instance, ignoring endogenous nurse absenteeism can result in understaffing [24].

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

There is significant heterogeneity in patient types. In Section 3, all patients were assumed to have identical characteristics. However, patients are likely to have different service requirements and/or no show rates. Similarly, one could consider appointment scheduling for two patient types: routine versus urgent patients as in Dobson et al. [16]. This work does not account for the no-show phenomenon. An interesting research direction would be to consider how to combine heterogenous patients with the no-show phenomenon.

There has been a growing interest in the operations research community to understand how human behavior impacts operations. Accounting for “behavioral operations” when designing system can improve performance [4]. A recent survey of this area emphasizes the need for experimental data to identify the impact of human behavior [2]. While randomized controlled experiments are often not possible in healthcare settings, adaptive behavior can be measured from retrospective data as described in this chapter.

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

References

1. 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. *JAMA*, 266:1085–1090, 1991. 2.1
2. E. Bendoly, K. Donohue, and K. L. Schultz. Behavior in operations management: Assessing recent findings and revisiting old assumptions. *Journal of Operations Management*, 24(6):737–752, 2006. 5
3. D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2005. 4.1
4. J. Boudreau, W. Hopp, J. O. McClain, and L. J. Thomas. On the interface between operations and human resource management. *MSOM*, 5:179–202, 2003. 5
5. M.D. Buist, G.E. Moore, S.A. Bernard, B.P. Waxman, J.N. Anderson, and T.V. Nguyen. Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: preliminary study. *British Medical Journal*, 324:387–390, 2002. 1
6. C.W. Burt and S.M. Schappert. Ambulatory care visits to physician offices, hospital outpatient departments, and emergency departments: United States, 1999-2000. *Vital Health Stat.*, 13(157):1–70, 2004. 1
7. B Cardoen, E Demeulemeester, and J Belien. Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, 201:921–932, 2009. 1
8. S. Chagaturu and S. Vallabhaneni. Aiding and abetting - nursing crises at home and abroad. *New England Journal of Medicine*, 353(17):1761–1763, 2005. 1
9. D. B. Chalfin, S. Trzeciak, A. Likourezos, B. M. Baumann, and R. P. Dellinger. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine*, 35:1477–1483, 2007. 1
10. C. W. Chan, V. F. Farias, N. Bambos, and G. Escobar. Maximizing Throughput of Hospital Intensive Care Units with Patient Readmissions. *Working Paper, Columbia Business School*, 2011. 4.2, 6, 7
11. P.S. Chan, H.M. Krumholz, G. Nichol, and B.K. Nallamothu. Delayed time to defibrillation after in-hospital cardiac arrest. *New England Journal of Medicine*, 358(1):9–17, 2008. 1
12. Committee on the Future of Emergency Care in the United States. *Emergency medical services at the crossroads*. Washington, DC: The National Academies Press, 2007. 1
13. R. Cooper, T. Getzen, H. McKee, and P. Laud. Economic and demographic trends signal an impending physician shortage. *Health Affairs*, 21(1):140–154, 2002. 1
14. B.T. Denton, A.J. Miller, H.J. Balasubramanian, and T.R. Huschka. Optimal Allocation of Surgery Blocks to Operating Rooms Under Uncertainty. *Operations Research*, 58(4):802–816, 2010. 1
15. R. Derlet and J. Richards. Overcrowding in the nations emergency departments: complex causes and disturbing effects. *Ann Emerg Med*, 35:63–68, 2000. 2.1
16. G. Dobson, S. Hasija, and E. J. Pinker. Reserving Capacity for Urgent Patients in Primary Care. *Production and Operations Management*, 20(3):456–473, 2011. 5
17. G. Dobson, H.-H. Lee, and E. Pinker. A Model of ICU Bumping. *Operations Research*, 58:1564–1576, 2010. 2.3, 3, 4.2
18. C.G. Durbin and R.F. Kopel. A Case-Control Study of Patients Readmitted to the Intensive Care Unit. *Critical Care Medicine*, 21:1547–1553, 1993. 4.2
19. 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? *J Emerg Med*, 15:397–399, 1997. 1, 2.1
20. 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. 1, 1, 2.1
21. B. Golden, M. Harrington, R. Konewko, E. Wasil, and W. Herring. Reducing Boarding in a Post-Anesthesia Care Unit. *Production and Operations Management (to appear)*, 2011. 1
22. L. V. Green and S. Savin. Reducing Delays for Medical Appointments: A Queueing Approach. *Operations Research*, 56:1526–1538, 2008. 1, 2.1, 3.1, 4, 3.1, 5

23. L.V. Green and S. Glied. The impact of ambulance diversion on myocardial infarction mortality. *Inquiry (to appear)*, 2011. 1
24. L.V. Green, S. Savin, and N. Savva. 'Nursevendor Problem': Personnel Staffing in the Presence of Endogenous Absenteeism. *Working paper, Columbia Business School*, 2011. 1, 5
25. 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]. *Acad Emerg Med*, 9:514, 2002. 2.1
26. D. Kc and C. Terwiesch. Impact of Workload on Service Time and Patient Safety: An Econometric Analysis of Hospital Operations. *Management Science*, 55:1486–1498, 2009. 1, 2.2, 2.2
27. D. Kc and C. Terwiesch. An Econometric Analysis of Patient Flows in the Cardiac ICU. *MSOM (to appear)*, 2011. 2.2, 2.2
28. Leonard Kleinrock. *Queueing Systems*, volume I: Theory. Wiley Interscience, 1975. 3.1
29. A. Kolker. Process modeling of emergency department patient flow: Effect of patient length of stay on ED diversion. *Journal of Medical Systems*, 32:389–401, 2008. 1
30. E. Litvak, P.I. Buerhaus, F. Davidoff, M.C. Long, M.L. McManus, and D.M. Berwick. Managing Unnecessary Variability in Patient Demand to Reduce Nursing Stress and Improve Patient Safety. *Joint Commission Journal on Quality and Patient Safety*, 31:330–338, 2005. 1
31. N Liu, S Ziya, and V Kulkarni. Dynamic Scheduling of Outpatient Appointments Under Patient No-Shows and Cancellations. *MSOM*, 12:347–364, 2010. 2.1, 4.1, 1
32. G. De Luca, H. Suryapranata, J.P. Ottervanger, and E.M. Antman. Time delay to treatment and mortality in primary angioplasty for acute myocardial infarction: every minute of delay counts. *Circulation*, 109(10):1223–1225, 2004. 1
33. M.L. McManus, M.C. Long, A. Cooper, J. Mandell, D.M. Berwick, M. Pagano, and E. Litvak. Variability in Surgical Caseload and Access to Intensive Care Services. *Anesthesiology*, 98:1491–1496, 2003. 1
34. J. Merritt, J. Hawkins, and P.B. Miller. Will the Last Physician In America Please Turn Off the Lights? A Look at Americas Looming Doctor Shortage. Technical report, Irving, TX: Practice Support Resources, Inc., 2004. 1
35. C. G. Moore, P. Wilson-Witherspoon, and J. C. Probst. Time and money: Effects of no-shows at a family practice residency clinic. *Family Medicine*, 33:522–527, 2001. 2.1
36. M. Murray and D.M. Berwick. Advance access: Reducing waiting and delays in primary care. *J. Amer. Medical Association*, 289:10351039, 2003. 1
37. E.G. Poon, T.K. Gandhi, T.D. Sequist, H.J. Murff, A.S. Karson, and D.W. Bates. 'I Wish I Had Seen This Test Result Earlier!': Dissatisfaction With Test Result Management Systems in Primary Care. *Arch Intern Med*, 164:2223–228, 2004. 1
38. B. Renaud, A. Santin, E. Coma, N. Camus, D. Van Pelt, J. Hayon, M. Gurgui, E. Roupie, J. Hervé, M.J. Fine, C. Brun-Buisson, and J. Labarère. Association between timing of intensive care unit admission and outcomes for emergency department patients with community-acquired pneumonia. *Critical Care Medicine*, 37(11):2867–2874, 2009. 1
39. F. Rincon, S.A. Mayer, J. Rivolta, J. Stillman, B. Boden-Albala, M.S V. Elkind, R. Marshall, and J.Y. Chong. Impact of Delayed Transfer of Critically Ill Stroke Patients from the Emergency Department to the Neuro-ICU. *Neurocritical Care*, 13:75–81, 2010. 1
40. E. Rivers, B. Nguyen, S. Havstad, J. Ressler, A. Muzzin, B. Knoblich, E. Peterson, and M. Tomlanovich. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *New England Journal of Medicine*, 345(19):1368–1377, 2001. 1
41. E.S. Salsberg and G.J. Forte. Trends in the physician workforce, 1980-2000. *Health Affairs*, 21:165–173, 2002. 1
42. B. C. Strunk and P. J. Cunningham. Treading water: Americans access to needed medical care, Report, 1997-2001. *Center for Studying Health System Change, Washington, D.C.*, 2002. 1, 3.1
43. M.D. Swenson. Scarcity in the Intensive Care Unit: Principles of Justice for Rationing ICU Beds. *American Journal of Medicine*, 92:552–555, 1992. 4.2
44. 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*, 27:w84–95, 2008. 1

45. J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge,MA, 2002. 2.2
46. N. Yankovic. *Models for Assessing the Impact of Resource Allocation in Hospitals*. PhD thesis, Columbia Business School, 2009. 1