PRODUCTION AND OPERATIONS MANAGEMENT

POMS

Vol. 20, No. 3, May–June 2011, pp. 359–374
ISSN 1059-1478 | EISSN 1937-5956 | 11 | 2003 | 0359

DOI 10.1111/j.1937-5956.2011.01231.x

Design and Analysis of Hospital Admission Control for Operational Effectiveness

Jonathan E. Helm, Shervin AhmadBeygi, Mark P. Van Oyen
Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109-2117, USA
jhelm@umich.edu, shervin@umich.edu, vanoyen@umich.edu

Variability in hospital occupancy negatively impacts the cost and quality of patient care delivery through increased emergency department (ED) congestion, emergency blockages and diversions, elective cancelations, backlogs in ancillary services, overstaffing, and understaffing. Controlling inpatient admissions can effectively reduce variability in hospital occupancy to mitigate these problems. Currently there are two major gateways for admission to a hospital: the ED and scheduled elective admission. Unfortunately, in highly utilized hospitals, excessive wait times make the scheduled gateway undesirable or infeasible for a subset of patients and doctors. As a result, this group often uses the ED gateway as a means to gain admission to the hospital. To better serve these patients and improve overall hospital functioning, we propose creating a third gateway: an expedited patient care queue. We first characterize an optimal admission threshold policy using controls on the scheduled and expedited gateways for a new Markov decision process model. We then present a practical policy based on insight from the analytical model that yields reduced emergency blockages, cancelations, and off-unit census via simulation based on historical hospital data.

Key words: health care operations; patient flow modeling; Markov decision processes; admission control
History: Received: December 2008; Accepted: September 2010, after 2 revisions.

1. Introduction

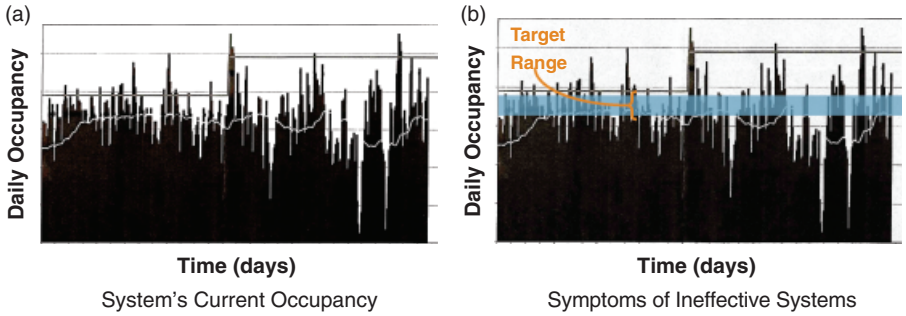
In the United States, health care lags significantly behind the manufacturing sector in process improvement practice. One consequence of this is that hospital care services are subject to significant, unnecessary, and detrimental fluctuations in patient census and associated workload. This variability in patient census has been linked specifically to congestion and chaos in the emergency department (ED), excessive radiology backlogs, strains on nurse and ancillary staff, and overcrowding in the post-acute care unit to name a few. This system-wide congestion results in compromised quality of care, emergency patient blockage for lack of beds, excessive patient length of stay (LOS), and significant understaffing and overstaffing costs (see Forster et al. 2003, Han et al. 2007, Haraden and Resar 2004, Machlin and Carper 2007, Price et al. 2011, and Sprivulis et al. 2006).

This paper introduces and evaluates innovative mechanisms for managing the variability in hospital workload at a key source: inpatient admissions. Currently, most hospitals use only one mechanism for daily admission control; that is, they reactively cancel elective surgeries only when there are no more inpatient beds available. In such systems, no control exists to increase the short-term utilization of hospital resources. This occurs because hospitals often categorize admissions as either scheduled elective or emergency patients, foregoing the potential to redesign the ED and admissions to accommodate patients who need to be seen within a few days but are not true emergency cases. As a result, hospitals have some control over the arrival rate of scheduled patients, but very little control over the arrival rate of emergency patients.

It can be counterproductive to lump the entire range of unscheduled patient types into one category: emergency. As has been noted (see, for example, Griffith et al. 1976, Lim et al. 1975, and Munson and Hancock 1972) one can identify a third category of patient that we refer to here as *expedited patients*. For patients in this category, the acuity of their medical condition is less than most ED patients who are admitted, and their admission to the hospital can be delayed 1–3 days, for example, without compromising their treatment. Without an efficient expedited admission process, these patients are often admitted through the ED due to excessive waiting times if they seek admission as an elective patient.

The novelty of this paper lies in developing a dynamic control structure that effectively uses a call-in mechanism for servicing this third class of patient and a controlled cancelation mechanism of elective patients. By properly servicing expedited patients, the excess load they placed on the ED during periods of peak congestion is reduced. This also reduces the arrival rate of uncontrollable random admissions. Thus

Figure 1 Controlling Census Variability in Hospitals



the expedited call-in queue can be used to smooth hospital occupancy levels over time to increase utilization of a hospital's expensive resources, such as staff and beds. The effect of creating both a call-in and a proactive cancelation mechanism is to squeeze the hospital occupancy variation (see Figure 1), while leaving a sufficient capacity buffer to accommodate potential future emergency arrivals. This, in turn, increases the quality of care, facilitates patients' access to health care, and decreases the overall hospital costs resulting from occupancy variability.

This paper develops a stochastic model for dynamic inpatient admission control that uses detailed information on the number of occupied beds by bed unit (ward) to show that:

1. Using both model-based proactive cancelation and call-in control mechanisms has advantages to using only reactive cancelation, as is the prevailing practice.
2. An easily implementable multi-dimensional double threshold policy for controlling both these mechanisms can effectively balance the opportunity cost of unfilled beds against the potential for cancelation of electives and bed block for emergency and elective patients.

It is important to make a distinction between our proposed dynamic *control* approach vs. *scheduling* or *planning* models. Admission scheduling models are typically concerned with decisions to generate efficient schedules (see Beliën and Demeulemeester 2007, Helm and Van Oyen 2010a, May et al. 2011) or capacity plans (see Zhang et al. 2008) for operating rooms, diagnostic labs (see Patrick et al. 2008) or outpatient clinics (see Gupta and Denton 2008). The schedules are created by assigning patients to time slots, sequencing the operations or dynamically managing the capacity. In contrast, we propose a closed-loop (i.e., feedback driven) control model linking system-wide behavior to operational-level decisions to stabilize hospital occupancy. An optimized scheduling system is not a substitute for our proposed control approach, but rather a preamble for it. In practice, our control mechanism would take the schedule as an input and provide a planned response to the realization of the schedule and exogenous random events; we simulate such an approach in section 4.

This paper contributes to the literature a clearer understanding of the value of feedback control for inpatient admissions. First, it develops a stylized model to generate insight into the structure of a typical admission system. Second, it argues the optimality of a double threshold admission policy for this stylized model and proves several properties for two new queueing operators. Third, the insight from the analytical model is used to develop a practical hospital admission policy with significant potential to improve health care delivery, which is demonstrated via simulation on historical hospital data.

The structure of this paper is as follows. In section 2, we discuss the motivation and background of hospital admission control systems and review the related literature. In section 3, we develop and analyze two stochastic models for understanding the essential dynamics of hospital admission control systems. In section 4, we propose a practical admission policy based on the insight from section 3 and use a simulation framework to demonstrate the benefits for real-world applications.

2. Models for Hospital Admission Control

Based on our discussions with several major hospitals in the state of Michigan, 40–60% of inpatients are admitted through the ED. The Emergency Medical Treatment and Active Labor Act requires US hospitals, by law, to medically stabilize emergency patients. Though ED services are significantly impacted by inpatient admissions from other hospital services, hospitals often lack an effective system for coordinat-

ing of hospital admissions, bed units, and the ED. Olshaker and Rathlev (2006) argue that a primary cause of ED overcrowding and ambulance diversion is a lack of empty inpatient beds for emergency patients. ED overcrowding negatively impacts the quality of patient care, leading to increased LOS and/or worsening patient disposition (see Forster et al. 2003) and in extreme cases, increased mortality rate (see Sprivulis et al. 2006).

In the United States, ED overcrowding is also exacerbated because many non-emergency patients use the ED as a convenient means to get admission into a highly utilized hospital. While it is not currently common practice, we seek to show that ED congestion can be reduced by using a patient flow management system that effectively uses an expedited call-in queue and uses the hospital census levels in admission decisions. We review the literature on patient flow models that link census to admission decision making (section 2.1) and the use of expedited call-in queues in hospitals (section 2.2).

2.1. Patient Flow Modeling and Linking Admission to Census

Dunn (1967), Connors (1970), and Shonick and Jackson (1973) represent early efforts in the modeling of hospital occupancy and admissions. Harrison et al. (2005) used a multi-stage stochastic approach to establish that variability in daily hospital occupancy in combination with high occupancy levels can increase the risk of hospital overflows.

Using simulation, Hancock and Walter (1979) developed (1) an inpatient admission scheduling and control system to achieve high average occupancy subject to constraints on the number of cancelations and emergency diversions, and (2) prediction models for the maximum average occupancy attainable using their control system. Jun et al. (1999) and Jacobson et al. (2006) used surveys to establish that an effective patient flow model can enable high patient throughput, low patient wait times, short LOS, and low clinic overtime. Hall et al. (2006) provided a systematic approach to health care engineering based on four different levels of granularity: *macro*, *regional*, *center*, and *department*. Our model targets the "center" (hospital level) and can be used at the departmental level. Proudlove et al. (2007) used forecasting and simulation to predict and manage inpatient flows into hospital beds. This paper builds upon the above work by using feedback driven control to optimize the use of reactive admission mechanisms with respect to system-level metrics. Recent advances in patient flow modeling can also be found in Bretthauer et al. (2011) and White et al. (2011).

2.2. An Expedited Call-in Queue for Quick Response

Long wait times for elective (scheduled) admission—not only for the hospital but for primary care providers as well—can force patients to use the ED as an expedient means for hospital admission. A lack of bed availability sometimes causes doctors to work around the system by declaring an emergency—even if a true emergency does not exist—to get their patients admitted more quickly. This "work-around" behavior strains the ED and subverts proper admission control. Our goal is to extend the concept of a call-in queue to meet the needs of patients requiring "expedited" inpatient care and allow them to be seen within a few days (a faster turnaround than for a typical elective admission). The literature below demonstrates the efficiency of patient call-in queues in a variety of settings.

A call-in mechanism gives patients who need expedited care a new pathway into the hospital and thus reduces the load on the ED. Because a majority of hospital costs are fixed, an empty hospital bed carries a significant *opportunity cost* (see Kroneman and Siegers 2004 and Machlin and Carper 2007). The expedited call-in queue can be used to increase the hospital occupancy during low occupancy periods to avoid this opportunity cost. Streamlined management of an expedited call-in queue for inpatients has been partially implemented before (see Griffith et al. 1976, Lim et al. 1975, and Munson and Hancock 1972). Though call-in queues are not commonly used in the United States, this practice is more prevalent in Europe. Worthington (1991) analyzes the practice of using inpatient call-in queues in the United Kingdom. Patrick and Puterman (2007) discuss how establishing a pool of on-call outpatients can improve resource utilization in an undercapacitated diagnostic center. On-call patient queues are also widely used for outpatient clinics (see Klassen and Rohleder 2004) and elective surgeries (see Bowers and Mould 2002).

3. A Markov Decision Process Model for Hospital Admission Control

Our dynamic admission control approach uses two mechanisms—elective admission cancelation and the call-in of expedite patients from a waiting list—to strike a balance between bed utilization and hospital congestion. The purpose of this section is to develop intuition into the structure of an optimal admission policy that is used to build the practical admission control mechanism presented in section 4. For the sake of intuition we present a stylized Markov decision process (MDP) model that focuses only on key dynamics of an effective admission control system.

Of relevance to our methodology, Gerchak et al. (1996) presented a dynamic programming model to optimize elective surgery schedules considering uncertainty in operating room capacity due to emergency arrivals. Green et al. (2006) used a dynamic programming approach to control admission of inpatients, outpatients, and emergency patients into a diagnostic medical facility. Patrick et al. (2008) developed an approximate dynamic programming approach for scheduling multi-priority patients to a public diagnostic facility. Dobson et al. (2011) analyzed the effect of reserving slots for urgent patients in the context of primary health care practice using MDP models. See Gupta (2007) and Gupta and Denton (2008) for a detailed survey on the use of Markov decision models for appointment scheduling and controlling admissions in a variety of health care delivery environments.

3.1. The Model

Our proposed MDP models seek to balance the opportunity cost of unutilized resources with the penalties associated with heavy congestion in those resources. In particular we consider balancing the opportunity cost of unfilled hospital beds with the penalties of (1) canceling elective admissions and (2) emergency arrivals that are blocked from entering a hospital bed. By analyzing the structure of this model, we determine the form of an optimal admission control policy to support the development of specific, practical policies for application.

For the general model consider a queueing system with two queues—a call-in queue and the hospital itself—as shown in Figure 2. Scheduled patients and emergency patients are assigned hospital beds according to a Poisson process with rate λ_s' and λ_e' while the call-in patients are placed on the call-in queue according to a Poisson process with rate λ_q' and are then assigned a hospital bed (admitted) via the admission controller's call-in action. The controller also has the option of canceling an arriving scheduled patient. Patients receive service according to an exponential distribution with rate μ'.

Let $\mathbf{X}^\pi(t) = (X_1^\pi(t), X_2^\pi(t))$ denote the number of patients in the system at time t under policy π, where $X_1^\pi(t) \in \bullet^+$ is the number of patients in the hospital and $X_2^\pi(t) \in \bullet^+$ is the number of patients on the call-in queue. Let B be the number of inpatient hospital beds. Let h_1' be the opportunity cost of an *empty* bed—quantified as $(B - X_1^\pi(t))^+$. In contrast, h_2' represents the per unit time (inventory) holding cost associated with holding a patient in the call-in queue.

Let τ' be the cost associated with having too many patients in the hospital—quantified as $(X_1^\pi(t) - B)^+$. When an emergency patient arrives at the door, it is forbidden by law to turn them away even if the hospital is full. This means that when a hospital reaches peak occupancy, emergency patients start to back up in the ED and/or are placed in the halls until a bed opens up. Similarly, scheduled surgeries remain in the operating room, occupying critical OR resources if there is no bed available for them post-operation. These are adverse situations for the hospital and the patient and thus a penalty of τ is assessed for each patient that is forced to wait in the OR, the ED, or in the hallway for a bed to open up.

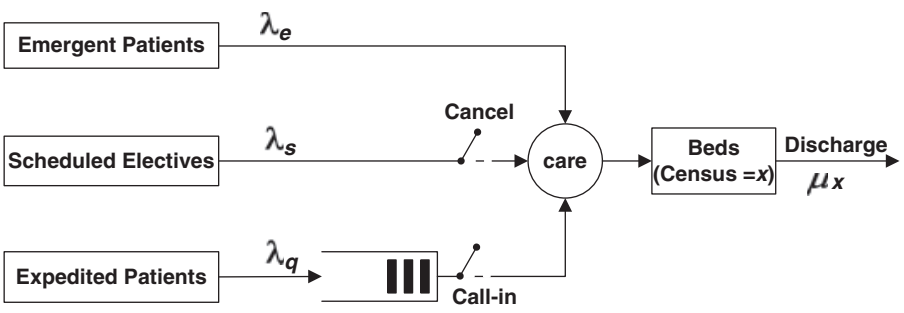
Finally, let c' be the cost of canceling a scheduled patient. If we let $N^\pi(t)$ be the counting process for the number of cancelations by time t under policy π, we can formulate the average cost per unit time objective function for policy π as

$$Z^\pi = \limsup_{T\to\infty} E\frac{1}{T}\left[\int_0^T (h_1'(B - X_1^\pi(z))^+ + \tau'(X_1^\pi(z) - B)^+ + h_2' X_2^\pi(z))dz + c'N^\pi(T)\right]. \tag{1}$$

A policy is then optimal if it achieves the optimal cost among all admissible policies Π as follows:

$$Z^* = \inf_{\pi\in\Pi} Z^\pi. \tag{2}$$

Figure 2 Two-Dimensional Admission Control System



Our goal is to use the structure of the optimal policy for this model to gain insight into hospital management practice. The following theorem, proved in Helm and Van Oyen (2010b), shows this goal can be achieved by analyzing the finite-horizon discounted version of the problem.

THEOREM 3.1. *For the average-cost optimality equation, if* $\lambda_e' + \lambda_q' < B\mu'$ *then the following hold.*

(i) *There exists an average-cost optimal stationary policy.*
(ii) *The optimal average cost can be computed by:* $Z^* = \inf_{\pi\in\Pi} Z^\pi = \lim_{\beta\to 1^-} \lim_{n\to\infty(1-\beta)} V_{n,\beta}(x)$, *where* $V_{n,\beta}(x)$ *is the n stage discounted value function.*
(iii) *Let* $\pi_{n,\beta}$, *denote an optimal policy for the n-period (discounted cost) problem. Then any limit point* π_β *of the sequence* $\{\pi_{\beta,n}\}_{n\geq 1}$ *as* $n\to\infty$ *is optimal for the infinite-horizon discounted cost. Moreover, any limit point of the sequence* $\{\pi_\beta\}_{\beta\in(0,1)}$ *(as* $\beta\to 1^-$*) is average-cost optimal.*

Theorem 3.1 allows us to (1) guarantee the existence of an optimal stationary policy and (2) establish that the finite-horizon problem converges to the infinite-horizon average cost problem in both cost and policy. This means that the insights from analyzing the finite-horizon problem are conferred to the original infinite-horizon average-cost optimal problem in Equation (2). The condition $\lambda_e' + \lambda_q' < B\mu'$ is sufficient to guarantee that the birth death model of the Markov chain induced by the certain policies is stable and thus the objective function cannot be infinite.

3.2. MDP Formulation

The final step before beginning our analysis of the n-period discounted model is to apply uniformization (Lippman 1975) to formulate the discrete time equivalent of the discounted problem. To do so, let the uniformization factor $\psi = \lambda_e' + \lambda_s' + \lambda_q' + B\cdot\mu'$. Let $\lambda_e = \lambda_e'/\psi$, $\lambda_s = \lambda_s'/\psi$, $\lambda_q = \lambda_q'/\psi$, $\mu = \mu'/\psi$ denote the discrete time parameters after uniformization corresponding to the transition probabilities in the embedded discrete time Markov chain (DTMC). Let α be the continuous time discount factor in the original problem and ξ be an exponential random variable with rate ψ (the length of time for one transition in the discrete chain). The equivalent discrete time discount factor for the DTMC then becomes

$$\beta = E[e^{-\alpha\xi}] = \int_0^\infty (e^{-\alpha t})(\psi e^{-\psi t})dt = \frac{\psi}{\psi+\alpha}. \qquad (3)$$

This implies that the discrete time costs can be defined in terms of the continuous time costs as $h_1 = h_1'/(\psi+\alpha)$, $h_2 = h_2'/(\psi+\alpha)$, and $\tau = \tau'/(\psi+\alpha)$. Because c is assessed per event and not assessed per unit time set $c = c'$. Thus, the discrete time instantaneous one stage cost can be written as

$$\begin{aligned} C(\mathbf{X}) &= E\left[\int_0^\xi (h_1'(B-X_1)^+ + \tau'(X_1-B)^+ + h_2'X_2)e^{-\alpha t}dt\right] \\ &= \frac{1-\beta}{\alpha}(h_1'(B-X_1)^+ + \tau'(X_1-B)^+ + h_2'X_2) \\ &= \frac{h_1'}{\psi+\alpha}(B-X_1)^+ + \frac{\tau'}{\psi+\alpha}(X_1-B)^+ + \frac{h_2'}{\psi+\alpha}X_2 \\ &= h_1(B-X_1)^+ + \tau(X_1-B)^+ + h_2X_2. \end{aligned} \qquad (4)$$

Now we can formulate the finite-horizon optimal expected discounted cost recursive optimality equation as

$$\begin{aligned} V_{n+1,\beta}(\mathbf{x}) = C(\mathbf{x}) &+ \beta\{\lambda_e\cdot V_{n,\beta}(\mathbf{x}+e_1) \\ &+ \lambda_q\cdot\min\{V_{n,\beta}(\mathbf{x}+e_1), V_{n,\beta}(\mathbf{x}+e_2))\} \\ &+ \min(x_1,B)\cdot\mu\cdot[\mathbb{1}_{\{x_2>0\}} \\ &\times\min\{V_{n,\beta}(\mathbf{x}-e_1), V_{n,\beta}(\mathbf{x}-e_2)\} \\ &+ \mathbb{1}_{\{x_2=0\}}V_{n,\beta}((\mathbf{x}-e_1)^+)] \\ &+ \lambda_s\min\{V_{n,\beta}(\mathbf{x}+e_1), c/\beta + V_{n,\beta}(\mathbf{x})\} \\ &+ (B-x_1)^+\mu\cdot[\mathbb{1}_{\{x_1=0,x_2>0\}} \\ &\times\min\{V_{n,\beta}(\mathbf{x}), V_{n,\beta}(\mathbf{x}+e_1-e_2)\} \\ &+ (1-\mathbb{1}_{\{x_1=0,x_2>0\}})V_{n,\beta}(\mathbf{x})]\}, \end{aligned} \qquad (5)$$

where $V_{n,\beta}(\mathbf{x})$ represents the optimal cost of the n-period β-discounted problem starting in state $\mathbf{x} = (x_1, x_2)$. $\mathbb{1}\{\cdot\}$ is the indicator function, and e_i represents the ith unit vector. In other words, e_i is the vector that contains all zeros except for a 1 in the ith position. The initial condition, $V_{0,\beta}\equiv 0$ is assumed for mathematical convenience and has no effect on the results for the infinite-horizon problem.

The first term of the value function is the one period instantaneous cost described in Equation (4). The second term represents an emergency arrival, which is always admitted. The third term represents the arrival of an expedited patient. When this event occurs, the controller either calls them into the hospital and assigns a bed upon arrival (term 1 of the minimization) or places them on the call-in queue (term 2). The fourth term represents the discharge of a patient from the hospital. When a patient is discharged, a bed is freed up, so at this point the hospital can decide to admit an expedited patient and backfill the empty bed (term 2 of the minimization) or to free the bed for scheduled/emergency patients that may arrive in future periods (term 1). Note that, if there are no patients in the call-in queue, no call-in action is available (the second indicator of the discharge term). The fifth term represents a scheduled arrival and the decision to begin treatment and assign a bed to the

patient (term 1 of the minimization) or to cancel the patient (term 2). The final term combines the uniformization term (at the end) with a call-in decision when $x_1 = 0, x_2 > 0$. That is, an empty system may call in patients at a rate of $B\mu$. While this state has essentially zero probability for a realistic parametrization it is important in the case of $B = 1$ (see section 3.5).

For convenience of analysis, we reformulate this n-period problem using *event-based dynamic programming* operators (see Koole 1998). The cost, arrival, admission control and routing operators are defined as $T_{\text{cost}}f(\mathbf{x}) = C(\mathbf{x}) + f(\mathbf{x})$, $T_{\text{A}(1)}f(\mathbf{x}) = f(\mathbf{x}+e_1)$, $T_{\text{AC}}f(\mathbf{x}) = \min\{c+\beta\cdot f(\mathbf{x}), \beta\cdot f(\mathbf{x}+e_1)\}$, and $T_{\text{R}(\{1,2\})}f(\mathbf{x}) = \min\{f(\mathbf{x}+e_1), f(\mathbf{x}+e_2)\}$. Let $T_{\text{unif}}[f_1, f_2, f_3, f_4](\mathbf{x}) = \lambda_e\beta f_1(\mathbf{x}) + \lambda_q\beta f_2(\mathbf{x}) + \mu\cdot\beta f_3(\mathbf{x}) + \lambda_s f_4(\mathbf{x})$ be the uniformization operator. In addition we define a multi-server backfill operator that encompasses the dynamics of multiple servers as well as the dynamics of backfilling a discharge with a patient from the call-in queue

$$T_{\text{MB}}f(x) = \begin{cases} (x_1 \wedge B)\min\{f(\mathbf{x}-e_1), f(\mathbf{x}-e_2)\} + (B - x_1 \wedge B)f(\mathbf{x}) & \text{if } x_1, x_2 > 0 \\ B\ \min\{f(\mathbf{x}), f(\mathbf{x}+e_1-e_2)\} & \text{if } x_1 = 0, x_2 > 0 \\ (x_1 \wedge B)f((\mathbf{x}-e_1)^+) + (B - x_1 \wedge B)f(\mathbf{x}) & \text{if } x_2 = 0. \end{cases} \tag{6}$$

Using these operators we can rewrite the value function of Equation (5) as

$$V_{n+1,\beta}(\mathbf{x}) = T_{\text{cost}}\cdot T_{\text{unif}} \times [T_{\text{A}(1)}V_{n,\beta}, T_{\text{R}(\{1,2\})}V_{n,\beta}, T_{\text{MB}}V_{n,\beta}, T_{\text{AC}}V_{n,\beta}](\mathbf{x}). \tag{7}$$

The following sections develop intuition into the structure of an optimal policy for the above system using the n-period discounted model described in Equation (5). After discussing a few modeling assumptions, we begin our analysis with a simplified case of the general model that disregards a detailed model of expedited call-in patients to isolate the effect of hospital occupancy alone on decision making. This model provides a sense of "what's best" for the hospital. Next we analyze the general formulation to gain insight into a policy that balances the hospital's needs for operational efficiency with the needs of the patients on the call-in queue.

3.3. MDP Modeling Assumptions

To reflect practice, we focus on the equilibrium properties of the steady-state MDP model, rather than finite-horizon effects. To model arrivals, we follow Harrison et al. (2005), who showed that a Poisson process is a good model both for elective and emergency arrivals to hospital beds. The Poisson assumption for call-in arrivals follows directly from Bernoulli splitting of the Poisson process for emergencies. As in Harrison et al. (2005), we are not addressing elective admission *scheduling* optimization and so we assume that the hospital has no control over the elective admission schedule. This is in part to isolate the value of a call-in queue and in part because this is a reality in many hospitals in the United States due to decentralized control that allows surgery services to schedule independently of the rest of the hospital. The surgery schedule is constantly changing even up until the last minute, random case times also lead to a nearly Poisson process of elective patients (surgical and medical) requesting a bed. It is important that we are modeling only the flow of elective patients into beds and do not explicitly model the operating room or prior activities.

Generally the arrival rates vary in a non-stationary manner by time of day and day of week. However, because the purpose of the model is to gain insight into the structure of a policy that balances the trade-offs between utilization and congestion and not to solve the MDP for specific values, making the stationary Poisson assumption for the aggregate flow of patients does not detract from our purpose. We allow for non-stationary arrivals that match historical arrival patterns in a real hospital in section 4, where we analyze a practical policy that harnesses the insight presented in this section.

The LOS in the hospital is assumed to be exponential for the purposes of tractability. Note that the policies developed in section 4 based on the insight gained from this queueing model and using distributions fitted from historical hospital data still show significant benefits to the hospital despite the modeling simplifications made for the MDP analysis.

3.4. Isolating Hospital System Efficiency with Occupancy-Based Decision Making

To isolate the effects of occupancy on hospital decision making and analyze in isolation, the value of census-based cancelation and call-in admission control, we simplify the modeling of call-in queue patients with the following model changes: (1) there is always a patient for the hospital to call in and (2) there is no waiting penalty for call in patients. While these assumptions lack the realism of a true hospital, they enable us to analyze the hospital's "best case" control options with a simple, insightful model which is later shown to provide reliable insights for complex models.

The value function in Equation (8) represents the uniformized version of this parsimonious model. We proceed by analyzing the n-period discounted problem for this special case of the general model of section 3.1

$$V_{n+1,\beta}(x) = C(x) + \beta\{\lambda_e V_{n,\beta}(x+1) + \lambda_s \min\{V_{n,\beta}(x+1), c/\beta + V_{n,\beta}(x)\} + \min(x,B)\cdot\mu\min\{V_{n,\beta}(x-1), V_{n,\beta}(x)\} + (B-x)^+\mu\cdot V_{n,\beta}(\mathbf{x})\}. \tag{8}$$

Note that the state $x \in \bullet^+$ now represents only patients in the hospital, as we are not modeling the length of the call-in queue. $C(x)$ and the event operators are modified accordingly. To simplify the analysis, we do not allow call-ins when the system is empty. This is a good approximation because we are interested in large bed size B, so this state has negligible probability of occurrence.

We show that the value function is convex by demonstrating the closure of the event operators from section 3.2 under convexity, because $V_{n,\beta} = T_1 \cdots T_k V_{0,\beta}$ is just a composition of a combination of the above operators with the initial convex value function V_0.

THEOREM 3.2. *The value function as defined in Equation (8) is convex.*

PROOF. $V_{0,\beta} \equiv 0$ is trivially convex. The closure under convexity of the operators T_{cost}, $T_{\text{A(1)}}$, T_{AC}, $T_{\text{R}(\{1,2\})}$, and T_{unif} is shown in Koole (1998) under the operators of the same name. All that remains is the multiple server backfill operator T_{MB}. To show closure of T_{MB} under convexity we consider four combinations of minimizers a_1 and a_2 of the left-hand side (LHS) of

$$(x-1)\min\left\{\underbrace{f(x-2)}_{a_1}, \underbrace{f(x-1)}_{a_2}\right\} + (B-(x-1))f(x-1) + (x+1)\min\left\{\underbrace{f(x)}_{a_1}, \underbrace{f(x+1)}_{a_2}\right\} + (B-(x+1))f(x+1) \geq 2x\cdot\min\{f(x-1), f(x)\} + 2\cdot(B-x)f(x),$$

where a_1 represents the case where the discharge action minimizes the function and a_2 represents the case where the call-in action minimizes the function. We ignore the discount factor, β, because it can be divided out on both sides. To show closure under convexity, it suffices to show closure for every combination a_i, a_j on the LHS. Case a_1, a_1 reduces to the operator T_{MD} shown to be closed under convexity in Koole (1998). Case a_2, a_2 follows directly from the convexity f and from the properties of minimization. Case a_1, a_2 is easy, because for convex f, $f(x-2) \leq f(x-1) \Rightarrow f(x) \leq f(x+1)$, therefore $\min\{f(x-2), f(x-1)\} = f(x-2) \Rightarrow \min\{f(x), f(x+1)\} = f(x)$. The final case, a_2, a_1, follows from

$$\begin{aligned}
&(x-1)f(x-1) + (B-(x-1))f(x-1) + (x+1)f(x) + (B-(x+1))f(x+1) \geq \\
&(B-(x+1))2f(x) + 2f(x) + 2f(x-1) + (x-1)[f(x-1) + f(x)] = \\
&(B-x)2f(x) + (x+1)f(x-1) + (x-1)f(x) \geq \\
&(B-x)2f(x) + (x+1)\min\{f(x), f(x-1)\} + (x-1)\min\{f(x), f(x-1)\} = \\
&2(B-x)f(x) + 2x\ \min\{f(x), f(x-1)\}.
\end{aligned}$$

The first inequality follows by combining terms 2 and 4 and using convexity. Next, the equality involves rearranging terms. The following inequality follows from properties of minimization, and the final equality again involves rearranging terms and using properties of minimization. If $x > B$ then T_{MB} reduces to $B\ \min\{f(x-1), f(x)\}$, which preserves convexity because it is equivalent to T_{AC} where $c = 0$. The case $x = B$ can be shown for cases a_2, a_1 and a_2, a_2 directly using the properties of minimization and convexity. For the boundary, where $x = 1$ the result follows directly from convexity for the only two cases on the LHS: a_2, a_1 and a_2, a_2. Because V_0 is convex and all the operators that make up the t-stage value function, $T_{(\cdot)}$, are closed under convexity, the value function itself is convex for all t. □

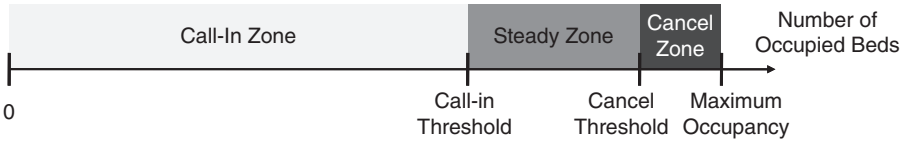
COROLLARY 3.1. *The optimal policy for the value function as defined in Equation (8) is (i) of double threshold type for both the cancelation action as well as the call-in action with a call-in threshold, θ^{S}, and a cancelation threshold, θ^{C}, and moreover (ii) $\theta^{\text{S}} \leq \theta^{\text{C}}+1$.*

PROOF. (i) The threshold structure of the optimal policy for both actions follows directly from convexity of the value function.

(ii) At θ^{S} the discharge action is cheaper, i.e., $f(\theta^{\text{S}}) - f(\theta^{\text{S}}-1) \geq 0$, and the call-in action is cheaper at $\theta^{\text{S}}-1$, i.e., $f(\theta^{\text{S}}-1) - f(\theta^{\text{S}}-2) \leq 0$. At θ^{C}, the cancelation action is cheaper, i.e., $f(\theta^{\text{C}}+1) - f(\theta^{\text{C}}) \geq c$, and the admit action is cheaper at $\theta^{\text{C}}-1$, i.e., $f(\theta^{\text{C}}) - f(\theta^{\text{C}}-1) \leq c$. From these equations, it is clear that at $\theta^{\text{C}}+1$, the discharge action will dominate the call-in action because $f(\theta^{\text{C}}+1) - f(\theta^{\text{C}}) \geq c \geq 0 \Rightarrow f(\theta^{\text{C}}+1) \geq f(\theta^{\text{C}})$. Because θ^{S} is the smallest state at which the discharge action will dominate the call-in action, clearly $\theta^{\text{S}} \leq \theta^{\text{C}}+1$. □

From Corollary 3.1(i) one can use thresholds θ^{S} and θ^{C} to decompose the hospital occupancy into zones based on the optimal admission control actions. Corollary 3.1(ii) suggests that the call-in threshold generally lies below the cancelation threshold (i.e.,

Figure 3 Zones for Zone-Based Admission Control vs. Number of Filled Beds



$\theta^S \leq \theta^C+1$). When $\theta^S<\theta^C$, the state space can be decomposed specifically into three zones as shown in Figure 3: a *call-in* zone, a *steady* zone, and a *cancel* zone. In the call-in zone, the controller admits expedited and scheduled patients. In the steady zone, the controller admits the scheduled patients but does not call in expedited patients. In the cancel zone, the controller cancels the scheduled patients and does not admit expedited patients. In practice, the solution is elegant and the hospital need only determine what zone the census falls into and take action accordingly.

3.5. Balancing Hospital Efficiency and Call-in Patient Service

In this section, we capture the additional effect of the length of the call-in queue to penalize the waiting time of expedited patients as in the general model formulated in Equation (5). Recall the state space is $\mathbf{x} = (x_1, x_2) \in \bullet^+ \times \bullet^+$, where x_1 is the number of patients in the hospital and x_2 is the number of patients in the call-in queue.

The key result is that we identify properties of the value function in general and again find a double threshold policy to be optimal, except that in this model we have a two-dimensional double threshold policy. This section proves the structure for a special (but non-trivial) case of one bed, and the next section numerically extends the insight for an arbitrary number of beds. That is, there is a call-in and cancelation threshold in both number of occupied beds *and* in number of patients on the call-in queue. To extend the insight about threshold policies from the previous parsimonious model, the following properties are sufficient for a threshold policy in two dimensions to be optimal with respect to the call-in and cancel actions: (1) supermodularity, (2) superconvexity, and (3) convexity in x_1 and x_2. For notational convenience, we denote the class of functions that has these properties as $\Im$.

A supermodular function has the property $f(x+e_1)+f(x+e_2) \leq f(x)+f(x+e_1+e_2)$. A superconvex function has two symmetric properties called (1) *superconvexity 1* defined as $f(x+e_1)+f(x+e_1+e_2) \leq f(x+e_2)+f(x+2e_1)$ and (2) *superconvexity 2* defined as $f(x+e_2)+f(x+e_1+e_2) \leq f(x+e_1)+f(x+2e_2)$ (see Koole 1998).

In general, the interaction between the *call-in action* dynamics and the *multiple server* dynamics disrupts the propagation of the properties of $\Im$. In fact, section 3.6 reveals counterexamples where the $V_{n,\beta} \notin \Im$ even though the threshold structure is still optimal; moreover, the two-dimensional threshold structure holds in all cases of a large random test suite.

Following the form of analysis from section 3.4 we first show that $V_0(\cdot), C(\cdot) \in \Im$. Then we show that if $f \in \Im$ then $T_i f \in \Im$ for all operators. Note that combining supermodularity and superconvexity gives convexity in both components immediately, so we omit the proof of convexity in x_1 and x_2.

LEMMA 3.1. $V_{0,g}C(\cdot) \in \Im$.

PROOF. $V_{0,\beta} \equiv 0$ is trivially supermodular and superconvex. The proof of these properties for $C(\cdot)$ is straightforward and can be easily checked by using the definition of $C(\cdot)$ both above, below, and at the boundary, B. Convexity in x_1 and x_2 follows directly from supermodularity and superconvexity, thus $V_0C(\cdot) \in \Im$. □

T_{cost}, T_{unif}, $T_{\text{A(1)}}$, $T_{\text{R}(\{1,2\})}$, T_{AC} are standard operators in Koole (1998), whose closure under the properties of the functions in $\Im$ was shown in that paper (as long as $C \in \Im$). It remains to show that the backfill operator, T_{MB}, is also closed under the properties of $\Im$. This is a new queueing operator, though it shares similarities with Ha (1997), whose closure under various properties has not yet been shown to the best of our knowledge. The closure of the backfill operator, T_{MB}, under supermodularity and superconvexity is shown first for $x_1>0$, $x_2>0$ and then for the boundary cases $x_1>0$, $x_2=0$, $x_1=0$, $x_2>0$, and $x_1=0$, $x_2=0$.

LEMMA 3.2. *T_{MB} as defined in Equation (6), is closed under supermodularity for $B=1$.*

PROOF. Suppose $f \in \Im$ and let $x_1>0$, $x_2>0$. For supermodularity we need to show the following:

$$T_{\text{MB}} \circ f(x) + T_{\text{MB}} \circ f(x+e_1+e_2) \geq T_{\text{MB}} \circ f(x+e_1) + T_{\text{MB}} \circ f(x+e_2). \tag{9}$$

In order to prove Equation (9) we need to consider four different cases based on the specific minimizing

actions of the LHS. We can re-write Equation (9) as

$$\min\left\{\underbrace{f(x-e_1)}_{a_1},\underbrace{f(x-e_2)}_{a_2}\right\}+\min\left\{\underbrace{f(x+e_2)}_{a_1},\underbrace{f(x+e_1)}_{a_2}\right\} \geq \min\{f(x),f(x+e_1-e_2)\}+\min\{f(x-e_1+e_2),f(x)\}, \tag{10}$$

where a_1 represents a discharge where the "no backfill" action minimizes the function. In a_2 the "backfill" action minimizes the function. Cases a_1,a_1 and a_2,a_2 follow directly from the supermodularity of $f\in\Im$. Cases a_1,a_2 and a_2,a_1 follow from convexity in x_1 and x_2, respectively.

Boundary conditions $x_1>0, x_2=0$: Here we want to show:

$$f(x)+\min\{f(x-e_1+e_2),f(x)\} \leq f(x-e_1)+\min\left\{\underbrace{f(x+e_2)}_{a_1},\underbrace{f(x+e_1)}_{a_2}\right\}.$$

For case a_1 the LHS $\leq f(x)+f(x-e_1+e_2)\leq f(x-e_1)+f(x+e_2)$ and using change of variable $y=x-e_1$, it is easily seen that this equation reduces to supermodularity of f. For case a_2, the LHS $\leq 2f(x)\leq f(x-e_1)+f(x+e_1)$ by convexity in x_1.

Boundary condition $x_1=0,\ x_2\geq 0$: The inequality holds by observing that at the boundary $T_{MB}f(x)=T_{MB}f(x+e_1)$ and $T_{MB}f(x+e_2)=T_{MB}f(x+e_1+e_2)$. □

The following lemma establishes closure of T_{MB} under superconvexity. This result holds for some intuitive conditions on the costs as shown.

LEMMA 3.3. *If* $h_1\leq\lambda_e\beta\tau$ *and* $h_2\geq 0$ *then* T_{MB} *is closed under superconvexity for* $B=1$.

PROOF. Suppose f is superconvex. First we need to show that the operator T_{MB} satisfies the first equation of superconvexity:

$$T_{MB}\circ f(x+e_2)+T_{MB}\circ f(x+2e_1)\geq T_{MB}\circ f(x+e_1)+T_{MB}\circ f(x+e_1+e_2). \tag{11}$$

We proceed as before, re-writing Equation (11) as

$$\min\left\{\underbrace{f(x-e_1+e_2)}_{a_1},\underbrace{f(x)}_{a_2}\right\}+\min\left\{\underbrace{f(x+e_1)}_{a_1},\underbrace{f(x+2e_1-e_2)}_{a_2}\right\} \geq \min\{f(x),f(x+e_1-e_2)\}+\min\{f(x+e_2),f(x+e_1)\}. \tag{12}$$

Similar to Theorem 3.2, we can show that case a_1, a_2 need not be considered. By the superconvexity of f we have $f(x)-f(x-e_1+e_2)\leq f(x+e_1)-f(x+e_2)\leq f(x+2e_1)-f(x+e_1+e_2)\leq f(x+2e_1-e_2)-f(x+e_1)$. The first two inequalities follow from superconvexity (1) and the second follows from superconvexity (2). Thus if $f(x-e_1+e_2)\leq f(x)$, as is implied by a_1,a_2, then $f(x+2e_1-e_2)\geq f(x+e_1)$ so a_1,a_2 can be eliminated. Cases a_1,a_1 and a_2,a_2 follow directly from the superconvexity of $f\in\Im$. Case a_2,a_1 follows directly from the properties of minimization.

The arguments for the proof for superconvexity (2) are completely symmetric to those used to prove superconvexity (1) so the proof is omitted.

Superconvexity (1)—boundary conditions: For the first equation of superconvexity, boundary case $x_1>0$, $x_2=0$ reduces to

$$f(x)+\min\{f(x+e_2),f(x+e_1)\} \leq \min\left\{\underbrace{f(x-e_1+e_2)}_{a_1},\underbrace{f(x)}_{a_2}\right\}+f(x+e_1).$$

For case a_1 the LHS $\leq f(x)+f(x+e_2)\leq f(x-e_1+e_2)+f(x+e_1)$. Using change of variable $y=x-e_1$, this equation reduces to superconvexity of f. Case a_2 follows directly from the properties of minimization.

Because $T_{MB}f(x+e_2)=T_{MB}f(x+e_1+e_2)$, the boundary case $x_1=0, x_2>0$ reduces to

$$\min\{f(x),f(x+e_1-e_2)\}\leq\min\{f(x+e_1),f(x+2e_1-e_2)\}.$$

It can easily be shown that if $h_1\leq\lambda_e\beta\tau$ and $h_2\geq 0$, then f is increasing in x_1 and therefore we obtain that $f(x+e_1)\geq f(x)\geq\min\{f(x),\ f(x+e_1-e_2)\}$ and likewise $f(x+2e_1-e_2)\geq f(x+e_1-e_2)\geq\min\{f(x), f(x+e_1-e_2)\}$.

Superconvexity (2)—boundary conditions: For the second equation of superconvexity, the boundary case $x_1>0,\ x_2=0$ reduces to

$$\min\{f(x-e_1+e_2),f(x)\}+\min\{f(x+e_2),f(x+e_1)\} \leq f(x)+\min\left\{\underbrace{f(x-e_1+2e_2)}_{a_1},\underbrace{f(x+e_2)}_{a_2}\right\},$$

where in case a_1 the LHS $\leq f(x-e_1+e_2)+f(x+e_2)\leq f(x)+f(x-e_1+2e_2)$. Using change of variable $y=x-e_1$, this equation reduces to superconvexity of f. Again case a_2 follows directly from the properties of minimization. The boundary case $x_1=0, x_2>0$ reduces to

$$2\min\{f(x+e_2),f(x+e_1)\}\leq\min\left\{\underbrace{f(x)}_{a_1},\underbrace{f(x+e_1-e_2)}_{a_2}\right\} +\min\left\{\underbrace{f(x+2e_2)}_{a_1},\underbrace{f(x+e_1+e_2)}_{a_2}\right\}.$$

Case a_1,a_2 follows directly from supermodularity, cases a_1,a_1 and a_2,a_2 follow directly from convexity in x_2. Case a_2,a_1 need not be considered because $f(x)-f(x+e_1-e_2)\leq f(x+2e_2)-f(x+e_1+e_2)$ by using superconvexity (2) twice. Thus if $f(x+e_1-e_2)\leq f(x)$ then $f(x+e_1+e_2)\leq f(x+2e_2)$, so a_2,a_1 can be eliminated. The

final boundary case, $x_1 = 0, x_2 > 0$ reduces to

$$2\min\{f(x+e_1), f(x+e_2)\} \le f(x) + \min\left\{\underbrace{f(x+2e_2)}_{a_1}, \underbrace{f(x+e_1+e_2)}_{a_2}\right\}.$$

Case a_1 follows from supermodularity and case a_2 follows from convexity in x_2. □

Our main result is rigorously proven for the case of only one bed, which is not trivial because all actions and states are possible and all costs play a role in the optimal policy. The numerical testing in the next section suggests that this result is general.

THEOREM 3.3. *If $h_1 \le \lambda_e \beta \tau$ and $h_2 \ge 0$, then for the special case of $B = 1$ server, a threshold policy in x_1 and x_2 is optimal for the backfill action, the cancel action, and the call-in action.*

3.6. Numerical Results

Because there are cases where $V_t \notin \Im$, we designed a test suite (shown in Table 1) composed of several representative cases (Cases 1–3) and a randomized test suite (Case 4) to investigate whether the threshold structure holds in a wide range of environments. We check these cases via the numerical computation of the MDP (value iteration) with a large finite state space. From Theorem 3.1, we know that the value iteration method will converge to the infinite-horizon optimal average cost value and policy that we are interested in. Case 1 describes a typical community hospital. Cases 2 and 3 consider extremely high and low values for the expedited call-in queue holding cost (h_2), respectively. To test the sensitivity of our model to the arrival process, within each case, we consider subcases (denoted $x-1$ in the second line of Table 1) where 60% of patient arrivals are emergencies and 40% are scheduled and its reverse (subcases $x-2$) where 40% of arrivals are emergencies and 60% are scheduled. In the first randomized test suite (Case 4-1), scheduled and emergency arrival rates are generated from a uniform distribution between 0 and 1, then the call-in arrival rate is taken to be a random, $U[0,1]$, percentage of the smaller of the two. This ensures that the call-in arrival rate is not higher than either of the two primary inputs to the hospital. In the second test suite (Case 4-2), this assumption is relaxed and the three rates are all generated from a $U[0,1]$ distribution and then normalized with respect to the random utilization. In Table 1, ρ represents the hospital utilization determined by $(\lambda_e+\lambda_s+\lambda_q)/(B\mu)$.

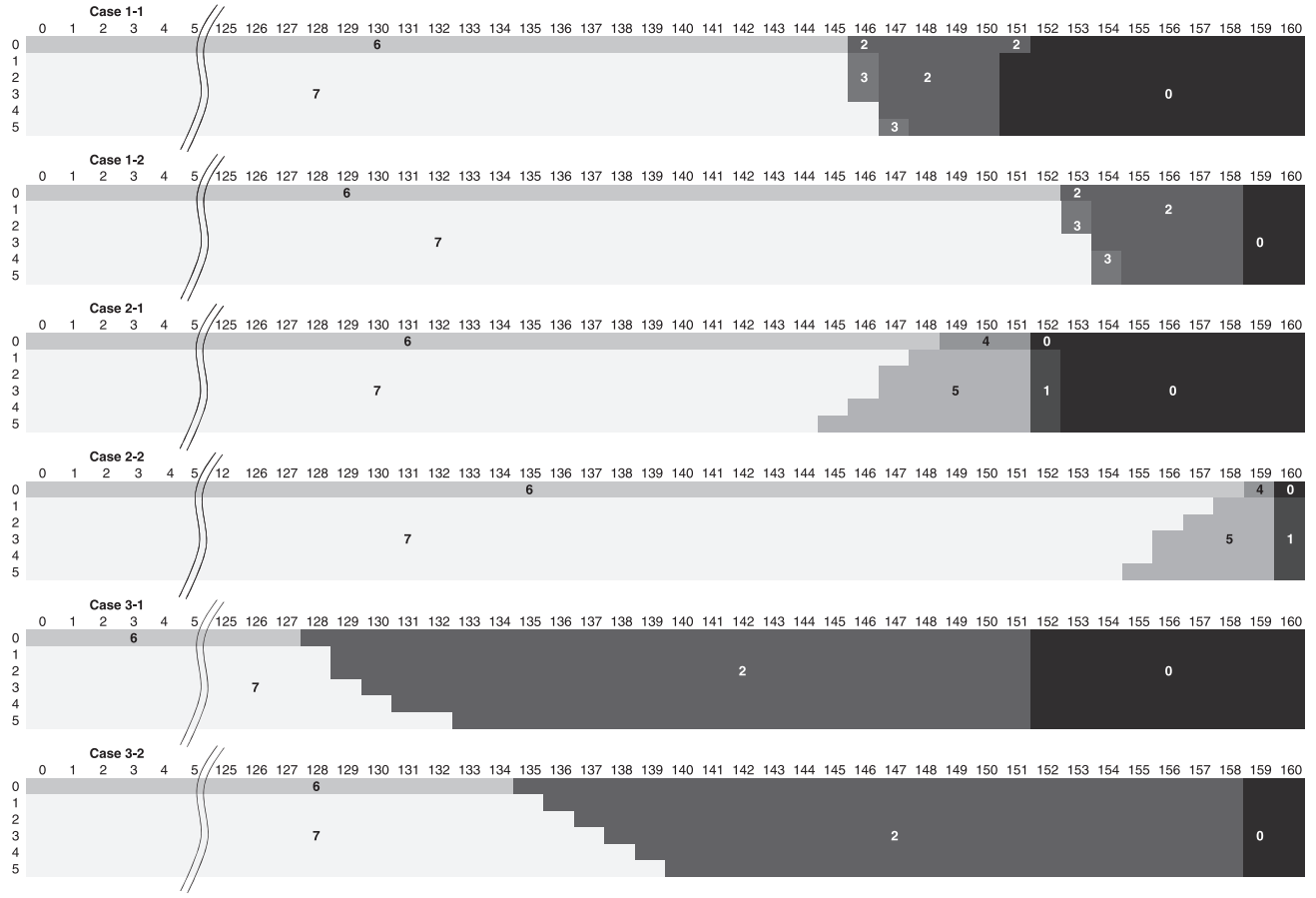
For each data set given in Table 1, we recursively solve the average cost per unit time MDP and record the optimal actions at each state. To illustrate the thresholds, we plot the minimizing action vs. number of beds (x_1, the horizontal axis) and number in the call-in queue (x_2, the vertical axis). Figure 4 exhibits switching curves as a function of the state for Cases 1, 2, and 3, where Table 2 explains the definition of each action.

In Figure 4, when hospital occupancy is low, the optimal policy is to call-in patients from the expedited queue in addition to admitting regular scheduled patients (actions 6 and 7). At higher occupancies, the optimal policy is to admit the scheduled patients but hold off the new call-in arrivals by placing them on the queue (actions 2 and 3). Finally, at critically high occupancy levels, the optimal policy seeks to avoid congestion by canceling scheduled patients and placing new call-in arrivals on the queue (actions 0 and 1). Note that as the emergency arrival rate increases, the control system starts canceling scheduled patients at lower occupancy levels (i.e., the black area in Figure 4 expands to the left). Furthermore, when the holding cost of a patient on the call-in queue increases, the system gives more attention to new call-in arrivals, cancels scheduled patients and admits new call-in arrivals (actions 4 and 5) at medium–high occupancy levels. On the other hand, when the holding cost of patients on the call-in list is zero (Case 3), scheduled patients are given more attention and action 2 becomes optimal at medium–high occupancy levels.

Table 1 Parameters Used for the Test Suite

	Case 1		Case 2		Case 3		Case 4	
Parma	1-1	1-2	2-1	2-2	3-1	3-2	4-1	4-2
λ_e	57%	38%	57%	38%	57%	38%	$U[0,1]$	$U[0,1]$
λ_s	38%	57%	38%	57%	38%	57%	$U[0,1]$	$U[0,1]$
λ_q	5%		5%		5%		$\min(\lambda_s, \lambda_e) \times U[0,1]$	$U[0,1]$
ρ	82%		82%		82%		$U[0,1]$	$U[0,1]$
B	160		160		160		$U[100,250]$	$U[100,250]$
h_1	1		1		1		$U[0,100]$	$U[0,100]$
h_2	1.5		100		0		$U[0,100]$	$U[0,100]$
c	34		34		34		$U[0,100]$	$U[0,100]$
τ	40		40		40		$U[0,100]$	$U[0,100]$

Figure 4 Optimal Actions



The action plots of the optimal policies in Figure 4 all clearly show that a threshold policy is optimal in both x_1 and x_2 for both the cancelation and call-in actions. Even though in some cases $V_t \notin \Im$, in all 2000 random problem instances (1000 each from Cases 4-1 and 4-2), the threshold policy for both actions in both dimensions is optimal. Thus, a threshold structure policy is at the very least optimal for a broad class of parameters.

4. Simulation Study of a Partner Hospital

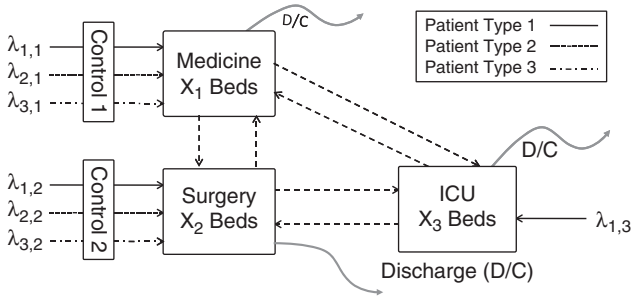
The MDP model discussed in the previous sections provides intuition into the double threshold properties of a hospital admission control system. This section develops a practical admission control mechanism based on the zone-based admission control policy introduced in section 3, but specifically modeling three bed units. We demonstrate the benefits of such a policy with a custom-designed C++ simulation study based on historical data from a partner hospital. We use the patient flow simulation framework described in Helm et al. (2008), so the features are explained at a high level to outline our approach. Figure 5 presents an abstraction of the simulation patient flow model.

The primary building block of our patient flow simulation model is the set of units, U, each having multiple beds. We allow for an arbitrary number of patient types, n, with patients of type $y \in \{1, \ldots, n\}$ arriving exogenously to these units according to a non-stationary Poisson process with rate $\lambda_{y,d,t}$, where

Table 2 Definition of Control Actions

Action code	Arriving call-in patient	Scheduled patient	Backfill	Action code	Arriving call-in patient	Scheduled patient	Backfill
7	Admit to hospital	Admit	Yes	3	Put in the queue	Admit	Yes
6	Admit to hospital	Admit	No	2	Put in the queue	Admit	No
5	Admit to hospital	Cancel	Yes	1	Put in the queue	Cancel	Yes
4	Admit to hospital	Cancel	No	0	Put in the queue	Cancel	No

Figure 5 Abstract Hospital Patient Flow Simulation Model



d represents the day of the week, t represents time of day, and $y \in \{1, \ldots, n\}$. After completing their initial segment of treatment, patients are transferred between units u_i and u_j with probability $p^y_{u_i,u_j}$ or are discharged from the hospital with probability $1 - \sum_{k \in U} p^y_{u_i,u_k}$. The empirical LOS distributions are based on patient type. If the unit to which a patient seeks admission is full, the patient is placed on a non-preferred unit that has the capability to care for the patient (e.g., a surgical patient placed on a medicine unit), as is common practice in most hospitals.

The patient flow simulation framework allows for custom controls to be designed and attached to framework elements. For our admission control study, we attach the zone-based admission control module to the arrival streams as shown in Figure 5. Our model's purpose allows us to exclude detailed modeling of the ED and OR, allowing us to focus on the admission control and bed occupancy dynamics.

In order to verify and validate the accuracy of our simulation model, we use strategies suggested by Sargent (2005). Verification proceeded via a series of white-box and black-box testing schemes. First we verified the correct operation of each of the components shown in Figure 5. Using detailed patient location/transition output that we generated for each unit for every turn of the simulation clock, we are able to verify that the correct number of patients are flowing through the system on a daily basis.

We validate our model of the system by comparing it against actual "real-world" hospital operations. That is, given a year's worth of hospital admissions data, we are able to extract the weekly scheduling pattern for elective medical and surgical patients and subsequently implement this policy in our model. Comparing the key features of the system (average daily census by day of week, volume of emergency and scheduled patients, and so forth) as in Lowery (1996), we find that our simulation closely matches the actual hospital operations.

Our simulated hospital is a medium-sized non-teaching hospital with approximately 50% emergency and 50% elective volume. There are three primary interacting units: surgery, medicine, and ICU. The majority of the hospital's patients can be divided into four surgical types and three medical types. Arrival rates, LOS distributions, and transfer probabilities are determined from 1 year of historical data.

In early discussions, several systemic problems were identified in this hospital so this study focuses primarily on mitigating these issues: (1) ED crowding, (2) excessive surgical cancelations, (3) difficulty admitting medicine patients in the middle of the week, and (4) too many patients placed off the preferred unit for their condition (e.g., a surgical patient on a medicine unit).

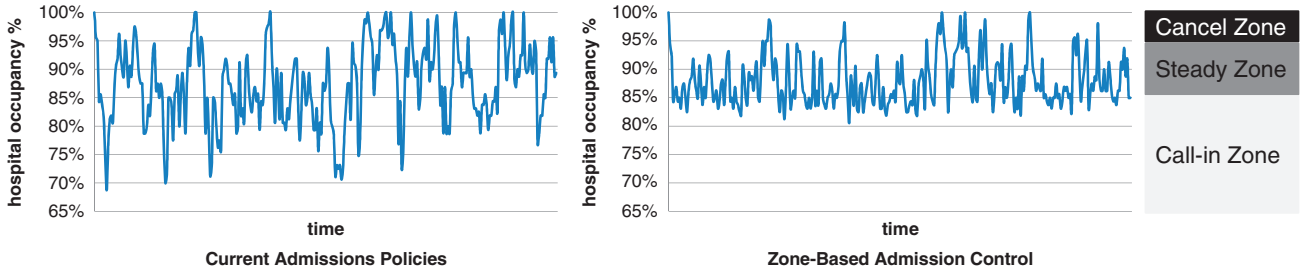
4.1. A Zone-Based Admission Control Mechanism and Call-in Queue Operation

Based on the analysis in section 3, we propose a zone-based admission control mechanism in which the hospital state is divided into three zones (cancel, steady, and call-in) as in Figure 3. However, the model of our partner hospital assigns each bed unit its own zone and also allowed the zones to vary by day of week to add refinement. The thresholds from the MDP model in section 3.5 gave a rough estimate for initial zone control values, which were refined using a heuristic simulation-based search. A single run of the simulation for 100 weeks takes less than a minute on a regular office PC. The simulation-based neighborhood search to find the admission zones takes approximately a half an hour, depending on the quality of the initial policy generated by the MDP. While the MDP model is simplified, it is very useful for approximating the more detailed control policy in the simulation.

An important practical feature of the call-in queue is that it should admit the patients on the queue within an advertised amount of time, which we set to 3 days in our simulation. In order to uphold this guarantee, we add a feature to the call-in queue that monitors each patient's length of time on the queue. When a patient reaches their maximum LOS on the queue, their priority is upgraded and they are automatically admitted into the hospital; we call this a "force-in." Because this mechanism limits the maximum waiting time on the call-in queue, our zone-based admission control can focus purely on the occupancy level. Of course, a significant number of call-in queue force-ins in practice signals that the system needs to be revised.

For practicality of implementation, we activate the control once per day (11 AM, for example) to allow enough time to call in patients from the call-in queue or cancel surgeries scheduled for later that day. This occupancy evaluation should include, if possible, projected discharges, surgeries scheduled and in progress, emergency patients boarding in the ED, and waiting for a bed and possibly other factors. At

Figure 6 Simulation Results—Comparing Current System with Zone-Based Admission Control



each review point, the hospital evaluates the predicted occupancy and applies the control policy; calling in patients up to the call-in threshold and canceling if census is above the cancelation threshold.

The expedited patients we identified are currently part of our partner hospital's emergency patient stream. The arrivals to the ED are modeled as a Poisson process with rate λ, with a proportion, q, of those arrivals being expedited patients. We can use Bernoulli splitting of the emergency Poisson process to generate two independent Poisson processes with rate $(1-q)*\lambda$ for true emergency arrivals and $q*\lambda$ for call-in queue arrivals. This splitting models the reality that a certain portion of ED arrivals occur because of the lack of an expedited mechanism for hospital admission within a short time frame.

4.2. Simulation Study Analysis

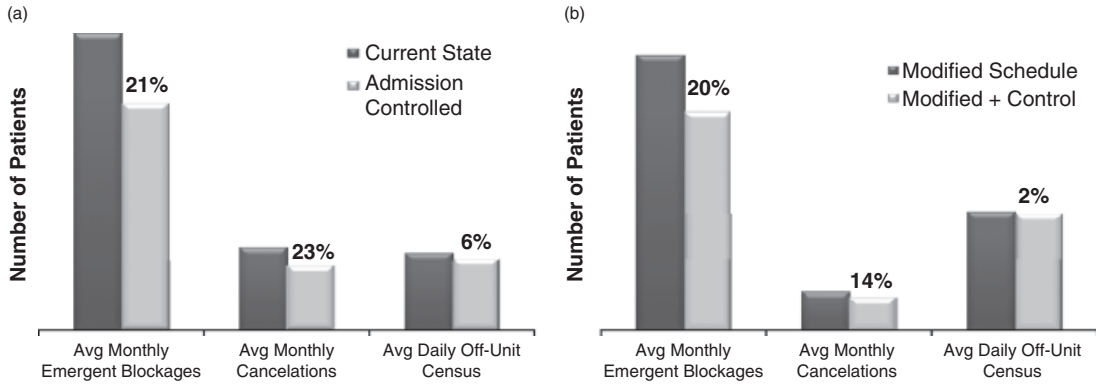
Using the current admission policies, Figure 6 shows one simulation sample path to visually demonstrate the effect of the expedited queue and zone-based admissions compared the sample path of the current admissions system.

Next we quantify the benefits of zone-based admission control in terms of the following key metrics: cancelations, emergency blockages, and off-unit census. Figure 7(a) compares 1 year of operation under the current admission policies (current state) with the same system using zone-based admission control (admission controlled). It conservatively models only 5% of the current emergency population as eligible for the expedited call-in queue.

Figure 7 clearly demonstrates that admission control can improve all three hospital metrics simultaneously. The most significant decreases are seen in cancelations and emergency blockages. The majority of cancelations and blockages occur in the middle of the week when the hospital is already full from a surgery schedule that emphasizes high volumes on Monday and Tuesday and few on Friday (which is a common practice). By placing call-in eligible candidates on a call-in queue during this occupancy spike we can avoid both emergency blockages and cancelations.

In general, an environment with a period of high volume scheduled electives followed by a period of low volume scheduled electives allows the expedited queue to improve the system in several ways. The expedited call-in queue delays a portion of patients who would otherwise exacerbate the ED crowding. This reduces the volume of non-electives and better accommodates the period of high elective volume, which in turn reduces the chance of emergency diversion and elective cancelation. Then as the occupancy starts to drop during the low volume period,

Figure 7 Simulation Results: Comparing the Key Hospital Metrics for (a) Current System vs. Zone-Based Admission Control and (b) A System with Improved Elective Schedule with and without Admission Control



expedited patients can be called in to bolster and smooth the occupancy level; with a host of benefits outlined in the introduction.

To further reduce the emergency blockages, the cancelation threshold can provide a small buffer on days where emergency blockages are frequent. This type of cancelation would probably be used sparingly (the cancelation zone is small) due to the lost revenue and organizational pain associated with canceling elective admissions, but even a small buffer properly placed provides significant relief for emergency blockages. Furthermore, it is likely that many hospitals serve elective patients at the expense of beds for emergency patients, given the prevalence of ambulance diversions (Olshaker and Rathlev 2006), so our approach helps to rationalize this system and protect emergency patients.

There are many possible zone schemes that provide different proportions in the trade-off between the three metrics. We chose the current scheme to provide benefit to all metrics and attempted to achieve fairly equal reductions in cancelations and emergency blockages. By manipulating the zone scheme one can achieve other reduction proportions in accordance with the desires of the partner hospital.

These results suggest that zone-based admission control can effectively improve hospital *system-level metrics*; however, one must also consider the implications for *service* to the expedited patients. Before introducing the call-in queue, expedited patients using the ED as a work around to gain admission were bed blocked at a rate of 11%. During blockage, these patients experienced excessive waiting in the ED, frequently being boarded in the hall, and the added stress of spending long periods of time in an overcrowded ED. After adding the call-in queue, the blockage rate for these patients dropped almost to zero. From a qualitative perspective, call-in patients are able to wait in the comfort of their own home and get some of the necessary tests *before* admission rather than in the ED. This potentially decreases their LOS, care costs, and they have a bed *ready* for them when they arrived at the hospital. Hence the call-in queue not only improves system performance, it also offers a valuable functionality to the expedited patients.

4.3. Sensitivity Analysis

In the previous section, we assumed that only 5% of emergency patients would be eligible for the call-in queue and the patient mix was 50% emergency and 50% scheduled. Under the heading ''call-in volume,'' Table 3 compares the improvement over the current system for the case where 5%, 10%, and 15% of the emergency patients are eligible for the expedited call-in queue. The three columns on the right labeled ''emergency volume'' test the system by varying the mix between emergency and scheduled patients (while keeping the original 5% call-in fraction of the emergency load fixed). The values are in terms of percent reduction vs. the current system in the three key hospital metrics we consider in this section. The system improves in all metrics as the volume of call-ins increases and the hospital gains more control over its random arrivals, but the system, as expected, exhibits decreasing marginal returns as the percent of call-in eligible patients increases. As might be expected, in the right three data columns of Table 3, when there is a higher volume of emergency patients, the zone-based control mechanism provides greater relative benefits than lower emergency volumes.

Table 3 Sensitivity Analysis of Call-In Queue and Emergency Volume

	Call-in volume (%)			Emergency volume with call-ins at 5% fraction (%)		
% Reduction vs. current	5	10	15	40	50	60
Cancelation	22	42	55	13	22	25
Emergency blockage	17	28	45	13	17	25
Off-unit census	6	7	8	6	6	2

In this paper, we have emphasized reactive admission control and not scheduling. One may conjecture that reactive control would not be necessary if elective schedules are properly managed. On the contrary, our analysis in Figure 7(b) shows that there are still significant benefits to using a zone-based admission control mechanism even when the elective admission schedule is controlled to improve performance. The improved elective admission schedule that we consider still has no elective admissions on the weekends but attempts to level the average daily census (or workload) across the weekdays to accommodate the prevailing practice. Note that the dark gray in Figure 7(b) represents the improved schedule without admission control, while the light gray represents the same improved schedule but using zone-based admission control as well.

Figure 7 demonstrates that zone-based admission control improves efficiency of health care delivery both in a hospital that does not control its own surgical schedule *and* in a hospital that has worked to improve its scheduling to increase operational efficiency. In fact, the sensitivity analysis of Table 3 shows that, though the degree of efficacy varies, in all the cases considered zone-based admission significantly benefits multiple aspects of hospital care delivery.

5. Conclusion and Future Research

By analyzing the structure of an admission control system, we have proven the optimality of a two threshold policy in a stylized model based only on census, and a special case of a more general model based on

census and length of the call-in queue. Though the MDP modeling in this paper considers a hospital with only one type of bed, the threshold concept can be extended to include multiple bed types as we have done in the simulation study of section 4 with very good results. Using simulation based on historical hospital data, we showed that a practical zone-based admission control policy can simultaneously reduce the number of emergency patients blocked, the number of canceled elective patients, and the amount of off-unit census. In all of our test cases, a zone-based admission control policy significantly improved key hospital performance metrics compared with current practice.

For future work, one may wish to consider if using more than three zones is worth the additional complexity. For example, one can further differentiate expedited patients into different queues based on severity of their condition. One can also consider different cancelation levels for different classes of scheduled patients.

For a hospital systems solution to be implementable and sustainable, the overhead and complexity should be minimized. The simple structure of a threshold/zone-based system fits this paradigm nicely, because it is easily understood and implemented. Analysis of historical hospital data and hospital dynamics/protocols can determine the threshold values, so the admissions department only needs to monitor the census and take action. This requires a centralized admissions decision-making body and census information. If the underlying dynamics of the hospital system change significantly, the zones will need to be adjusted; for example, a change in patient population, adding a new hospital wing, hiring a new surgeon, or holiday patterns.

Acknowledgments

Here, we gratefully acknowledge the constructive comments of the anonymous referees and the AE as well as the help of University of Michigan colleagues Walton M. Hancock, Hoda Parvin, and Arthur Adisusanto of the IOE Department and Joseph W. Norman. We also thank Paul Walter of Lean Care Systems and Jivan Deglise-Hawkinson from Université Joseph Fourier, Grenoble, and the many unnamed partners at multiple hospitals who shared their experience and data with us. The work of the first author was supported by a National Science Foundation Graduate Research Fellowship and was performed as part of his dissertation under the supervision of the third. The work of the third author was supported in part by NSF grant DMI-0542063.

References

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

Bowers, J., G. Mould. 2002. The deferrable elective patient: A means of reducing waiting-lists in orthopaedics. *J. Manage. Med.* **16**(2/3): 150–158.

Bretthauer, K. M., H. S. Heese, H. Pun, E. Coe. 2011. Blocking in healthcare operations: A new heuristic and an application. *Prod. Oper. Manag.* **20**(3): 375–391.

Connors, M. M. 1970. A stochastic elective admissions scheduling algorithm. *Health Serv. Res.* **5**(4): 308–319.

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

Dunn, R. G. 1967. Scheduling elective admissions. *Health Serv. Res.* **2**(2): 181–215.

Forster, A. J., I. Stiell, G. Wells, A. J. Lee, C. VanWalraven. 2003. The effect of hospital occupancy on emergency department length of stay and patient disposition. *Acad. Emerg. Med.* **10**(2): 127–133.

Gerchak, Y., D. Gupta, M. Henig. 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Manage. Sci.* **42**(3): 321–334.

Green, L. V., S. Savin, B. Wang. 2006. Managing patient service in a diagnostic medical facility. *Oper. Res.* **54**(1): 11–25.

Griffith, J. R., W. M. Hancock, F. C. Munson. 1976. *Cost Control in Hospitals*. Health Administration Press, Ann Arbor, MI.

Gupta, D. 2007. Surgical suites? Operations management. *Prod. Oper. Manag.* **16**(6): 689–700.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Trans.* **40**(9): 800–819.

Ha, A. Y. 1997. Optimal dynamic scheduling policy for a make-to-stock production system. *Oper. Res.* **45**(1): 42–53.

Hall, R., D. Belson, P. Murali, M. Dessouky. 2006. Modeling patient flows through the healthcare system. Hall, R. ed., *Patient Flow: Reducing Delay in Healthcare Delivery, International Series in Operations Research & Management Science*, Vol. 91. Springer, New York, NY, 1–44.

Han, J., C. Zhou, D. J. France. 2007. The effect of emergency department expansion on emergency department overcrowding. *Acad. Emerg. Med.* **14**(4): 338–343.

Hancock, W. M., P. F. Walter. 1979. The use of computer simulation to develop hospital systems. *SIGSIM Simul. Dig.* **10**(4): 28–32.

Haraden, C., R. Resar. 2004. Patient flow in hospitals: Understanding and controlling it better. *Front. Health Serv. Manage.* **20**(4): 3–15.

Harrison, G. W., A. Shafer, M. Mackay. 2005. Modelling variability in hospital bed occupancy. *Health Care Manage. Sci.* **8**(4): 325–334.

Helm, J. E., S. AhmadBeygi, M. P. Van Oyen. 2008. The flexible patient flow simulation framework. Proceedings of the 2009 Industrial Engineering Research Conference, Institute of Industrial Engineers, Vancouver, British Columbia, Canada.

Helm, J. E., M. P. Van Oyen. 2010a. Design and optimization methods for elective hospital admissions. Technical Report 11-01. Department of IOE, Univeristy of Michigan, Ann Arbor, MI.

Helm, J. E., M. P. Van Oyen. 2010b. Infinite horizon analysis of a hospital admission control model. Technical Report 10-01, Department of IOE, University of Michigan, Ann Arbor, MI.

Jacobson, S., S. Hall, J. R. Swisher. 2006. Discrete-event simulation of health care systems. Hall, R. ed., *Patient Flow: Reducing Delay in Healthcare Delivery, International Series in Operations Research & Management Science*, Vol. 91. Springer, New York, NY, 211–252.

Jun, J. B., S. H. Jacobson, J. R. Swisher. 1999. Application of discrete-event simulation in health care clinics: A survey. *J. Oper. Res. Soc.* **50**(2): 109–123.

Klassen, K. J., T. R. Rohleder. 2004. Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment. *Int. J. Serv. Ind. Manage.* **15**(2): 167–186.

Koole, G. 1998. Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Syst.* **30**(3): 323–339.

Kroneman, M., J. J. Siegers. 2004. The effect of hospital bed reduction on the use of beds: A comparative study of 10 European countries. *Soc. Sci. Med.* **59**(8): 1731–1740.

Lim, T., D. Uyeno, I. Vertinsky. 1975. Hospital admissions systems: A simulation approach. *Simul. Gaming* **6**(2): 188–201.

Lippman, S. A. 1975. Applying a new device in the optimization of exponential queuing systems. *Oper. Res.* **23**(4): 687–710.

Lowery, J. C. 1996. Design of hospital admissions scheduling system using simulation. Proceedings of the 28th Conference on Winter Simulation, IEEE Computer Society, pp. 1199–1204.

Machlin, K., S. R. Carper. 2007. Expenses for hospital inpatient stays, 2004. Statistical Brief #164, Agency for Healthcare Research and Quality.

May, J. H., W. E. Spangler, D. P. Strum, L. G. Vargas. 2011. The surgical scheduling problem: Current research and future opportunities. *Prod. Oper. Manag.* **20**(3): 392–405.

Munson, F. C., W. M. Hancock. 1972. Problems of implementing change in two hospital settings. *IIE Trans.* **4**(4): 256–266.

Olshaker, J. S., N. K. Rathlev. 2006. Emergency department overcrowding and ambulance diversion: The impact and potential solutions of extended boarding of admitted patients in the emergency department. *J. Emerg. Med.* **30**(3): 356.

Patrick, J., M. L. Puterman. 2007. Improving resource utilization for diagnostic services through flexible inpatient scheduling: A method for improving resource utilization. *J. Oper. Res. Soc.* **58**(2): 235–245.

Patrick, J., M. L. Puterman, M. Queyranne. 2008. Dynamic multipriority patient scheduling for a diagnostic resource. *Oper. Res.* **56**(6): 1507–1525.

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

Proudlove, N. C., S. Black, A. Fletcher. 2007. OR and the challenge to improve the NHS: Modelling for insight and improvement in in-patient flows. *J. Oper. Res. Soc.* **58**(2): 145–158.

Sargent, R. G. 2005 Verification and validation of simulation models. Proceedings of the 37th Conference on Winter Simulation, IEEE Computer Society, Orlando, FL, 130–143.

Shonick, W., J. R. Jackson. 1973. An improved stochastic model for occupancy-related random variables in general-acute hospitals. *Oper. Res.* **21**(4): 952–965.

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

White, D. L., C. M. Froehle, K. J. Klassen. 2011. The effect of integrated scheduling and capacity policies on clinical efficiency. *Prod. Oper. Manag.* **20**(3): 442–455.

Worthington, D. 1991. Hospital waiting list management models. *J. Oper. Res. Soc.* **42**(10): 833–843.

Zhang, B., P. Murali, M. M. Dessouky, D. Belson. 2008. A mixed integer programming approach for allocating operating room capacity. *J. Oper. Res. Soc.* **60**(5): 663–673.