# The ICU Will See You Now: Efficient–Equitable Admission Control Policies for a Surgical ICU with Batch Arrivals

Muer Yang
Opus College of Business, University of St. Thomas, Minneapolis, MN, 55403 yangmuer@stthomas.edu

Michael J. Fry
Carl H. Lindner College of Business, University of Cincinnati, Cincinnati, OH 45221, mike.fry@uc.edu

Corey Scurlock
Director, Cardiothoracic ICU, Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, NY 10011, corey.scurlock@mountsinai.org

Intensive care units (ICU) are frequently the bottleneck in the hospital system, limiting patient flow and negatively impacting profits. We examine admission control policies for a surgical ICU where patients arrive in batches. We formulate this problem as a Markov-decision process (MDP) with the objective of balancing system efficiency and equity. Equity concerns are driven by the decision maker's desire to provide equivalent service to different classes of patients (or equivalently to the surgeons who operate on those patients). We propose a simple and efficient heuristic solution method related to our MDP formulation that provides a performance guarantee. We also apply our admissions policy to a real setting motivated by the cardiothoracic-surgical ICU at Mount Sinai Medical Center in New York; the results demonstrate that our policy significantly outperforms competing policies and that the ICU can achieve large equity gains with limited efficiency losses.

*Key words*: Intensive Care Unit; Equity; Markov-decision process; Scheduling

## 1. Introduction

Intensive medical care is an expensive healthcare service. Hospital intensive care units (ICUs) are often among the most specialized, most staffed and best equipped hospital departments. The annual cost of hospital ICUs in the United States is over $90 billion, accounting for more than 20% of total hospital acute-care costs (Pronovost et al. 2004). ICUs serve 8% of the patient population but account for 30% of total inpatient expenditures (Cahill and Render 1999). Given these factors, hospitals attempt to fully utilize ICU resources, which often results in overcrowded ICUs. The

average bed occupancy in a hospital ICU generally ranges from 77% to 90% (Pronovost et al. 2004).

Heavily utilized ICUs cause numerous problems. ICUs frequently become the bottleneck to patient flow which limits the responsiveness throughout the hospital healthcare-delivery system (McManus et al. 2004, KC and Terwiesch 2007). Many surgeries must be postponed or even canceled due to the limited downstream capacities of ICUs (Green 2002, Min and Yih 2010). Postponed surgeries delay necessary treatment for critically-ill patients and they result in higher patient-care costs for the hospital. Given the often low profit margins for hospitals (Tor Schoenmeyr et al. 2009) and high ICU utilization, it is essential that ICU beds are allocated to patients in the best way possible.

We consider the case where the decision maker (DM) must decide which surgeries to schedule during a period given limited surgical-ICU beds (here we refer to an ICU that accepts only patients arriving from scheduled surgeries). In our scenario, each period the DM (who is generally an intensivist in charge of the ICU) is presented with a list of surgeries that are desired to be performed that period. However, the number of surgeries submitted for that period may be greater than the current number of available ICU beds. In such instances, the DM must decide which surgeries to perform that period and which to postpone. Complicating matters is the fact that the surgeries can differ in terms of procedures required during surgery. Different surgeries can differ in terms of the costs incurred for postponement and also may result in different lengths of stay (LOS) for the patients in ICU.

The DM wishes to choose surgeries to postpone that will increase the overall efficiency of the ICU, but s/he may also be concerned with functions related to equity among the different surgeries. This is because a specific type of surgery (or subset of surgeries) is often performed by a single surgeon. The DM needs to treat all surgeons 'fairly' by not consistently postponing the same surgeon's operations (which can effect both the surgeon's schedule and income). This adds to the modeling complications of our scenario, but is an important consideration for the DM in reality, particularly when surgeons are not hospital employees. Surgeon incomes are often tied directly to

procedures performed (Gosden et al. 1999); hence, surgeons demand equitable treatment in terms of postponed surgeries. Equivalently, the DM may wish to ensure patient equity due to metrics related to individual patient types or types of surgery. In terms of our model, either patient equity and surgeon equity can be included in our formulations; thus, we refer to "patient types" when discussing equity concerns.

We model this setting as a Markov-decision process (MDP) with the goal of minimizing the total cost of ICU admission rejections under the consideration of equity among different patient types. We desire to identify optimal, or near-optimal, admission control policies for the ICU. These policies should be easy-to-implement in practice and we also seek to provide intuition for hospital administrators.

The rest of this paper is arranged as follows. Section 2 reviews the literature on ICU-operations management. Section 3 introduces the MDP model formulation. Section 4 develops a myopic policy to solve the MDP model with a performance guarantee. Section 5 contains our numerical experiments to compare and examine the admission control policies based on our models and suggested solution methods. Section 6 presents a case study from a cardiothoracic-surgical ICU at Mount Sinai Medical Center using empirical data for patient arrivals and lengths of stay. Finally, Section 7 presents conclusions and possible extensions to our work.

## 2. Literature Review

There is a rich body of literature using operation-research tools to study patient flow and capacity management in healthcare settings (e.g., Smith-Daniels et al. 1988, Marshall et al. 2005, Green 2006). Numerous papers have studied how to improve operational efficiency in hospital ICUs. The existing operations literature mainly studies three aspects of ICU operations: the admission control process, ICU capacity, and the discharge process.

When the number of new admission requests to the ICU exceeds the number of available beds in the ICU, one approach is to reject the "extra" requests and put those patients on a waiting list. We consider such approaches to be admission control policies. Shmueli et al. (2003) compare three

different admission control policies where the objective is to maximize the expected number of survivals within a traditional queueing model to estimate the probability distribution of occupied beds in ICU. Kolker (2009) formulates a model to determine the maximum number of elective surgeries that should be scheduled each day given fixed ICU capacity in order to meet the stochastic demand from emergency surgeries.

Admission control policies can be considered as queue disciplines if we model the ICU as a queueing system. Queue disciplines (or nonpreemptive priority queues) have been well studied in the scheduling literature. For instance, the $c_i\mu_i$ rule (which is equivalent to the shortest processing time rule when $c_i = c_j, \forall i, j$) minimizes the average number in queue for a $G/G/1$ queue with only two types of customers (Shanthikumar and Yao 1992). The $c_i/\rho_i$ rule (which is equivalent to the lowest utilization rule when $c_i = c_j, \forall i, j$) minimizes the average waiting time in queue for the same queue system (Shanthikumar and Yao 1992) and also for an $M/G/c$ queue with multiple classes of customers who have identical service time distributions (Federgruen and Groenevelt 1988). A mixed dynamic rule (Federgruen and Groenevelt 1988) can also minimize the average waiting time for the $M/G/c$ queue with multiple classes of customers with identical service time distributions. To the best of our knowledge, no known policies can guarantee optimality under these metrics for the most general $G/G/c$ queues. Our models focus on the admission control policy for a hospital ICU.

The capacity management decision for ICU beds generally encompasses determination of the total number of ICU beds and the allocation of ICU beds to different types of patients (most often classified as either elective or emergency patients). Swenson (1992) emphasizes the idea of 'justice' to distribute ICU beds when there are fewer ICU beds than admission requests. The author suggests that the DM should rank patients based on available prognostic data and allocate ICU beds to the patients who are most in need based on his/her rank. Kim et al. (1999) develop models to determine the minimum number of ICU beds for given arrival rates. In a later paper (Kim et al. 2000), the authors investigate ICU-bed allocation policies that reserve a certain number of beds exclusively for elective-surgery patients. Kim and Horowitz (2002) examine a quota system which

allocates a different number of ICU beds to elective-surgery patients each day. De Bruin et al. (2007) determine the optimal bed allocation policy over the entire cardiac care chain (from the First Cardiac Aid, to the Coronary Care Unit, and to the normal care clinical ward) subject to a maximum number of rejected admissions. In our models, we assume that the total number of ICU beds is fixed a priori; and that all beds can be allocated to any patient regardless of surgery type.

Another strategy to increase throughput in the ICU is to discharge patients currently in the ICU earlier than they would be otherwise in order to allow admission of new patients to the ICU. Such a strategy is known as "ICU bumping" (e.g. Lowery 1993, Dobson et al. 2010, Chan et al. 2010). The drawback of ICU bumping is that the quality of care may be compromised for the ICU patients being bumped. ICU bumping increases the throughput at the risk of deteriorating bumped patients' physical conditions who then may require readmission to the ICU at a later time. KC and Terwiesch (2007) show that the patients being bumped out of ICU are 18% more likely to return to the ICU and have significantly longer length of stays in the ICU upon their return; however, the authors also show that the net peak capacity in the ICU can be increased. Dobson et al. (2010) provide a Markovian model to evaluate the steady-state bumping rates of patients in the ICU given arrival patterns and LOS distributions. The early discharge policy they employ is to discharge the patients who have the shortest expected remaining LOS in ICU. Chan et al. (2010) identify a greedy discharge policy to reduce the total expected readmission load due to early discharge where discharged patients return to the ICU with some known probability.

The ICUs in many hospitals do not want to increase the ICU throughput rate at the risk of patients' quality of care. Thus, they may refuse to consider ICU-bumping policies. The hospitals that we have worked with most closely choose to not engage in ICU bumping. Therefore, we do not include such considerations in our models.

None of the existing papers explicitly consider issues of equity in determining ICU-management policies as we do here. For the patients, 'equity' means that we should not keep rejecting the same type of patients; for the surgeons, 'equity' means that each surgeon should perform a similar load of surgeries based on his/her specialties and the number of patients requiring that surgery. Our

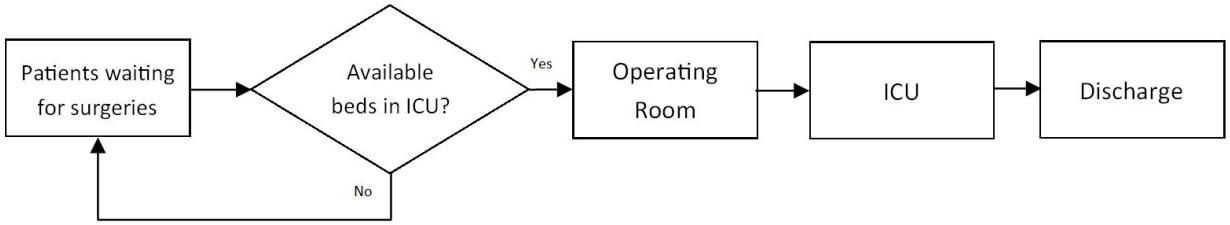


**Figure 1 The Current ICU Process for Surgical-Based ICUs**

paper, therefore, focuses on developing an optimal admission control policy for the ICU to minimize rejected admissions under the considerations of equity among patient types, given stochastic admission requests, batch arrivals, and fixed capacity in the ICU.

## 3. Model Formulation

A common ICU admission and discharge process for an ICU receiving patients only from scheduled surgeries is shown in Figure 1. The ICU admissions director (the DM in our setting) receives a list of surgery requests each weekday (usually late in the day for surgeries to be performed the following day). The number of surgery requests received each day is random in the view of the admissions director (in these settings the admissions director does not have control of, nor prior visibility of, the arriving surgeries). The DM then must decide which surgeries to allow and which to postpone for the following day. This decision is based on the current state of the ICU and knowledge of the arrival process distribution and length-of-stay distributions for patient types. (In reality, the DM receives the list of surgery requests the evening prior to the date of surgery and must approve/postpone surgeries at this time to allow patients to fast prior to surgery. However, patient discharges for the following day are generally known with certainty at this time, so that we can assume, without loss of generality, that the DM makes the approval/postpone decisions on the same day as the surgery.) Once a patient receives surgery they are sent to the ICU. Once in the ICU, patients recover according to some random process and then are discharged at a later date, freeing an ICU bed.

From a modeling perspective we are facing batch arrivals with multiple customer types. Batch sizes are random, but the interarrival time of batches is fixed. The LOS in ICU for each surgery type is allowed to follow a different distribution. There is a finite number of ICU beds available in total. The ICU beds can be allocated to any type of surgery patient. If too few ICU beds are available on a day to satisfy all admission requests, the DM must reject some requests. Those rejected surgeries must be postponed and will face an admission decision the following day. Similar to Chan et al. (2010), we formulate our model using a discrete time horizon $t \in [0, T]$ Markov-decision process. In each time period, the DM must determine which (if any) surgery requests should be rejected.

### 3.1. State Space

Each arriving surgery request (patient) can be characterized as one of $m \in \mathcal{M}$ types according to the specific characteristics of the patient and surgery request. Additionally, each surgery request can also be assigned to class $d \in \mathcal{D} = \{1, 2, ..., D\}$ based on the characteristics of the arriving patient's class, or more directly for our concern, the surgeon which will be performing the surgery. In the most general case, the type $m \in \mathcal{M}$ assignment could be a function of the overall health of the patient, the surgery type to be performed and the operating surgeon. The patient type then governs the transition probabilities for our MDP. The surgery-request (or patient) class, $d \in \mathcal{D}$, assignment is used in our cost functions and in calculating equity metrics, where here we assume that the equity of importance is among the operating surgeons; thus each class $d$ would correspond to a different operating surgeon.

We also define function $\delta(\cdot)$ which returns the class for a given surgery-request type. For example, for a type $m$ surgery request, $\delta(m)$ returns the class assignment (e.g., which surgeon is doing this surgery). The length of stay in ICU of a type-$m$ patient, $LOS_m$, follows a distribution, $F_m$, with mean $1/\mu_m$.

We assume that there are $B$ beds in the ICU, and that each bed can be assigned to any type-$m$ request. Denote $x_{t,m} \in \{0, 1, ..., B\}$ as the number of type-$m$ patients currently in the ICU at time $t$, and $y_{t,m}$ as the number of type-$m$ surgery requests arriving at time $t$ including those rejected during

time period $t-1$. If $B - \sum_{m=1}^{M} x_{t,m} < \sum_{m=1}^{M} y_{t,m}$, then $\sum_{m=1}^{M} (y_{t,m} + x_{t,m}) - B$ surgery requests must be rejected. This implies that we always use empty ICU beds, and we reject surgery requests only if all ICU beds are occupied.

The state space is therefore defined as

$$\mathcal{S} = \left\{ s_{x,y,t} : x \in \{0, 1, ..., B\}^M, \sum_{m=1}^{M} x_m \leq B, y_m \geq 0, t \in [0, T] \right\}.$$

*Remark*: We assume that we never reject a patient if an ICU bed is available. However, such a policy is not necessarily optimal for all objective functions. Consider the case where the DM wishes to minimize the total number of rejections over the time horizon of $T$ days. Denote any two consecutive arrivals of surgery requests as type $m_0$ and type $m_1$ (where $m_0$ and $m_1$ could be the same type). Suppose that $m_0$ and $m_1$ arrive at time $t_0$ and $t_1$, respectively, where $t_1 > t_0$. At time $t_0$, there is one empty bed in the ICU. Further, assume that the LOS for each patient is deterministic with values $LOS_{m_0}$ and $LOS_{m_1}$ for patients $m_0$ and $m_1$, respectively. If $LOS_{m_0} \leq LOS_{m_1}$, it is always better to admit $m_0$ at $t_0$ and not to leave the bed empty. If $t_1 - t_0 > LOS_{m_0} > LOS_{m_1}$, then it is always better to admit request $m_0$ at time $t_0$ and not to leave the ICU bed empty. If $t_1 - t_0 \leq LOS_{m_0}$, then leaving the ICU bed empty and rejecting $m_0$ may be optimal. To see this, assume there are no arrivals or departures during time $t \in \{t_0, t_0 + 1, \ldots, T\}$ other than those associated with $m_0$ and $m_1$. If $m_0$ is admitted into the ICU at $t_0$, then $m_1$ will be rejected once per day for $t_0 + LOS_{m_0} - t_1$ days. Alternatively, if $m_0$ is rejected at $t_0$, $m_1$ is admitted at $t_1$, and $m_0$ will be rejected once per day for $t_1 + LOS_{m_1} - t_0$ days. Thus, when $t_0 + LOS_{m_0} - t_1 > t_1 + LOS_{m_1} - t_0$, i.e., $LOS_{m_0} - LOS_{m_1} > 2(t_1 - t_0)$, it is better to leave the ICU bed empty and force patient $m_0$ to wait until $m_1$ departs the ICU. However, a policy of rejecting patients while keeping ICU beds open is considered infeasible by both the surgeons and hospital administrators – mainly due to the negative responses anticipated by both surgeons and patients to such a policy. In particular, hospital administrators are sensitive to the possibility of a patient experiencing adverse health effects while waiting for surgery if ICU beds are available.

### 3.2. Actions

For each state $s_{x,y,t}$, we define $A_t^M$ as the feasible actions that can be taken in time period $t$.

$$A_t^M = \left\{ A_{t,m} \in \{0, 1, ..., y_{t,m}\}, \forall m : \sum_{m=1}^{M} A_{t,m} = (\sum_{m=1}^{M} (x_{t,m} + y_{t,m}) - B)^+ \right\}$$

$A_{t,m}$ is the number of rejected surgery requests of type $m$ in state $s_{x,y,t}$. For any state $s_{x,y,t}$ where $\sum_{m=1}^{M} x_{t,m} + y_{t,m} > B$, we reject $A_{t,m}$ type-$m$ surgery requests, where $\sum_{m=1}^{M} A_{t,m} = \sum_{m=1}^{M} (x_{t,m} + y_{t,m}) - B$; for all other states, $A_{t,m} = 0, \forall m$. The surgery requests rejected at time $t$ return as surgery requests of the same type at time $t+1$.

### 3.3. System Dynamics

Let $X_{t,m}$ denote the number of type-$m$ patients leaving the ICU during time period $t$; and let $Y_{t,m}$ denote the number of new type-$m$ surgery requests arriving at the beginning of time period $t$, which does not include those surgery requests rejected in previous time periods. The batch size of each surgery type follows a discrete distribution. Let $p_{i,t,m}$ denote the probability that $i$ new type-$m$ surgery requests arrive at time $t$, where $i \in \{0, 1, ..., I_m\}$. These probabilities are known a priori. The $\{p_{i,t,m}\}$ process specifies the distribution of random variable $Y_{t,m}$, so that $Y_{t,m} \leq I_m$. We define $s_{x',y',t'} = S(s_{x,y,t}, A_t^M)$ as

$$x'_{t',m} = x_{t,m} + y_{t,m} - X_{t,m} - A_{t,m},$$

$$y'_{t',m} = Y_{t,m} + A_{t,m},$$

$$t' = t + 1.$$

### 3.4. Transition Cost Function

A cost is incurred when the system transitions from state $s_{x,y,t}$ to state $s_{x',y',t'}$ by taking action $A_t^M$. While many possible cost functions could be considered, in our setting we seek to provide both efficiency and equity in our admission policies. Let $\lambda_d = \frac{1}{T} \sum_{t=1}^{T} \sum_{\delta(m)=d} \sum_{i=1}^{I_m} i p_{i,t,m}, \forall m \in \mathcal{M}$, be the average batch size of class-$d$ surgeries over time. Let $A_{t,d} = \sum_{\delta(m)=d} A_{t,m}, \ \forall m \in \mathcal{M}$; and define vector $A_t^D = \left\{ A_{t,d}, \forall d \in \mathcal{D} \right\}$. Thus, $A_t^D$ can be used to identify which surgeon's surgeries are

postponed by taking actions $A_t^M$. Define set $\mathcal{A}_t^D$ which contains all feasible $A_t^D$. Finally, we define $C(s_{x,y,t}, A_t^D)$ as the transition cost function,

$$C(s_{x,y,t}, A_t^D) = \max_{d \in \mathcal{D}} \frac{A_{t,d}}{\lambda_d},$$

which simultaneously considers both efficiency and equity. It can also be viewed as an approximation of the maximum average surgery-request rejection rate among all patient classes over time.

### 3.5. Objective and Optimality Equation

The expected total cost (i.e., value function) under policy $\pi$ is

$$J^\pi(s) = E\left[\sum_{t'=t}^{T-1} C(s_{x,y,t'}, \pi(s_{x,y,t'})) \mid s = s_{x,y,t}\right]. \tag{1}$$

Denote $J^*(s) = \min_{\pi \in \Pi} J^\pi(s)$ and $\pi^*(s) \in \arg\min_{\pi \in \Pi} J^\pi(s)$. According to Bellman's equation, we can theoretically minimize (1) through a dynamic programming approach,

$$J(s) = \min_{A \in \mathcal{A}} E[C(s_{x,y,t}, A_t^D) + J(S(s_{x,y,t}, A_t^D))]. \tag{2}$$

The state-space size in practice, however, is too large for traditional dynamic programming approaches to solve (2). We therefore aim to design an efficient and easy-to-implement heuristic to generate near-optimal solutions to practical problems.

## 4. A Myopic Policy

The nature of our system dictates that surgery requests are rejected only when the number of available beds is less than the number of surgery requests. There are two extreme situations: no surgery requests are rejected whenever there are enough beds available in the ICU, and all surgery requests are rejected whenever the ICU is completely full. These two situations can be expressed mathematically as: if $\sum_{m=1}^M x_{t,m} + y_{t,m} \leq B$, then $A_{t,m} = 0$; if $\sum_{m=1}^M x_{t,m} = B$, then $A_{t,m} = y_{t,m}$.

For the general situation where the ICU is not full but there are not enough beds available to accommodate all surgery requests, we could, in theory, determine the optimal values of $A_t^D$ by solving (2). Given the size of the problem in realistic settings, it is impractical to solve this equation optimally. We thus develop a myopic policy, $\omega$, that can be solved much more efficiently.

Denote $\hat{A}_t^D$ as the surgery-request-rejection action according to policy $\omega$ given a state $s_{x,y,t}$. If $\sum_{m=1}^{M} x_{t,m} + y_{t,m} > B$ and $\sum_{m=1}^{M} x_{t,m} < B$, then define

$$\hat{A}_t^D \in \arg\min_{A_t^D \in \mathcal{A}_t^D} \max_{d \in \mathcal{D}} \frac{A_{t,d}}{\lambda_d}. \tag{3}$$

**Myopic Policy**

**Step 1.** Given initial state $s_{x,y,t}$.

**Step 2.1** If $\sum_{m=1}^{M} x_{t,m} + y_{t,m} \leq B$, then $A_{t,m} = 0$.

**Step 2.2** If $\sum_{m=1}^{M} x_{t,m} = B$, then $A_{t,m} = y_{t,m}$.

**Step 2.3** If $\sum_{m=1}^{M} x_{t,m} + y_{t,m} > B$ and $\sum_{m=1}^{M} x_{t,m} < B$, then determine $\hat{A}_t^D$.

**Step 3.** $t = t + 1$. If $t = T$, stop; otherwise go to Step 1.

Notice that Step 2.3 (i.e., solving (3)) itself is an optimization problem, and how to find $\hat{A}_t^D$ still remains in question. We therefore develop the following Min-Max Algorithm to solve (3) efficiently and optimally.

**Min-Max Algorithm**

**Step 1.** Initialize $Counter = 0$ and $A_{t,m} = 0, \forall m$.

**Step 2.** (Re)define $\mathcal{M}' = \{m : y_{t,m} - A_{t,m} \geq 1\}$.

**Step 3.** Let $A_{t,m} = A_{t,m} + 1$ for type-$m$ surgery request with the smallest $\frac{A_{t,\delta(m)}}{\lambda_{\delta(m)}}$, where $\forall m \in \mathcal{M}'$.

**Step 4.** $Counter = Counter + 1$. If $Counter = \sum_{m=1}^{M} x_{t,m} + y_{t,m} - B$, stop. Otherwise, go to Step 2.

The Min-Max Algorithm is easy to implement and, as stated in Theorem 1, it guarantees a global optimal solution to (3).

THEOREM 1. *The Min-Max Algorithm generates a global optimal solution to* (3).

*Proof* Denote $z(A_{t,d}) = \frac{A_{t,d}}{\lambda_d}$, $Z(A_t^D) = \max_{d \in \mathcal{D}} \frac{A_{t,d}}{\lambda_d}$, $\hat{A^*}_t^D$ as the global optimal solution to (3) and $\hat{A'}_t^D$ as the solution generated by the Min-Max Algorithm. Suppose $\hat{A^*}_{t,d} \neq \hat{A'}_{t,d}$ for some $d \in \mathcal{D}$. Let $z(\hat{A^*}_{t,d^*}) = Z(\hat{A^*}_t^D)$ and $z(\hat{A'}_{t,d'}) = Z(\hat{A'}_t^D)$.

Suppose $Z(\hat{A^*}_t^D) < Z(\hat{A'}_t^D)$. (Otherwise, nothing needs to be proved.)

Clearly, $z(A_{t,d})$ is an increasing function of $A_{t,d}$. Because $\hat{A^*}_{t,d} \geq 0$, $z(\hat{A^*}_{t,d}) \geq z(0) = 0$, and so $Z(\hat{A^*}_t^D) \geq 0$. If $Z(\hat{A^*}_t^D) = 0$, it means that $A_{t,d} = 0$, $\forall d$. Thus, there is no action to take.

Suppose $Z(\hat{A^*}_t^D) > 0$. Because $z(\hat{A^*}_{t,d'}) \leq z(\hat{A^*}_{t,d^*}) = Z(\hat{A^*}_t^D) < Z(\hat{A'}_t^D) = z(\hat{A'}_{t,d'})$, and $z(A_{t,d})$ is an increasing function in $A_{t,d}$, it follows that $\hat{A^*}_{t,d'} < \hat{A'}_{t,d'}$. Thus, since $\sum_d \hat{A^*}_{t,d} = \sum_d \hat{A'}_{t,d}$, $\exists l \in \mathcal{D}$ and $l \neq d'$ such that $\hat{A^*}_{t,l} \geq \hat{A'}_{t,l} + 1$, which implies that $z(\hat{A^*}_{t,l}) \geq z(\hat{A'}_{t,l} + 1)$. Due to the nature of the myopic policy, $z(\hat{A'}_{t,d} + 1) \geq z(\hat{A'}_{t,d'}), \forall d \neq d'$, so we have $z(\hat{A^*}_{t,l}) \geq z(\hat{A'}_{t,d'}) = Z(\hat{A'}_t^D)$. However, our original assumption states that $z(\hat{A^*}_{t,l}) \leq Z(\hat{A^*}_t^D) < Z(\hat{A'}_t^D)$, which is a contradiction.

Therefore the Min-Max Algorithm generates a global optimal solution to (3). □

Note that the Min-Max Algorithm guarantees a global optimal solution whenever the objective is a non-increasing (or non-decreasing) function.

### 4.1. Performance Bound for the Myopic Policy

The transition cost function, $C(s_{x,y,t}, A_t^D)$, includes costs related to both equity and efficiency. The total cost function (1) therefore forces the solution to be the optimal tradeoff between equity and efficiency. The "equity" here is represented by the maximum function; while the "efficiency" is represented by the number of rejections, $A_t^D$. Due to the nature of our problem, the decisions made in the current time period affect the future states visited and future costs, i.e., the surgeries rejected today will impact the total number of surgeries that have to be rejected in the future. The myopic policy ignores the future, focusing only on the present. The myopic policy minimizes $C(s_{x,y,t}, A_t^D)$ for a given $\sum_m A_{t,m}$ in the current time period $t$, but it may result in larger values of $\sum_m A_{t',m}$ and $C(s_{x',y',t'}, A_{t'}^D)$, $\forall t' > t$. In other words, the myopic policy minimizes the immediate cost at the possible expense of higher future costs.

To illustrate, assume that Patient A is rejected according to the myopic policy, but not according to the optimal policy. Under the myopic policy, Patient B, who would not have been admitted to the ICU under the optimal policy, enters the ICU under the myopic policy, and may stay in the ICU longer than patient A would have stayed. Thus, the bed taken by Patient B is occupied longer

than it would have been under the optimal policy. Therefore, the ICU may reject more patients in the future under the myopic policy than it would have under the optimal policy.

This intuition guides us to identify the worst-case scenario for the myopic policy. We define the worst case as the case where any patients rejected under the myopic policy will have shorter LOS than the patients rejected under the optimal policy. In this worst case, we assume that each patient rejected according to the myopic policy is not the one that should have been rejected according to the optimal policy. The cost of rejecting the wrong patient or admitting the wrong patient into the ICU is that this patient will occupy an ICU bed longer than the patient under the optimal policy, and therefore will potentially block a newly-arrived patient to be admitted to the ICU in the future. With the worst case identified, we derive a performance bound for the myopic policy in Theorem 2. Recall that $\omega$ represents the myopic policy.

THEOREM 2. *Assume that $LOS_m$ $\forall m$ follows an exponential distribution with mean $\mu_m$. For all state $s$, $J^{\omega}(s) \leq J^{*}(s) + \frac{BT}{\lambda_{min}}(\frac{1}{\mu_{min}} - \frac{1}{\mu_{max}})$.*

*Proof* Denote set $\mathcal{P}^{\pi}(s)$ as the set containing all patients rejected under policy $\pi$ in state $s$. Assume $\mathcal{P}^{*}(s) \neq \mathcal{P}^{\omega}(s)$, otherwise the myopic policy is the optimal policy. Let $\hat{\mathcal{P}}^{*}(s) = \mathcal{P}^{*}(s) - \mathcal{P}^{*}(s) \bigcap \mathcal{P}^{\omega}(s)$ and $\hat{\mathcal{P}}^{\omega}(s) = \mathcal{P}^{\omega}(s) - \mathcal{P}^{*}(s) \bigcap \mathcal{P}^{\omega}(s)$. Let $p_i^{\omega}(s) \in \hat{\mathcal{P}}^{\omega}(s)$ be a patient rejected under the myopic policy but not under the optimal policy, and let $p_j^{*}(s) \in \hat{\mathcal{P}}^{*}(s)$ be a patient rejected under the optimal policy but not under the myopic policy. Note that $\sum_i p_i^{\omega}(s) = \sum_j p_j^{*}(s) \leq B$. All patients $p_i^{\omega}(s)$ should have been admitted to the ICU according to the optimal policy, while all patients $p_j^{*}(s)$ actually are admitted to the ICU according to the myopic policy. Given the worst-case situation where $\mu_{p_j^{*}(s)} \leq \mu_{p_i^{\omega}(s)}$, in expectation, the ICU bed is occupied by patient $p_j^{*}(s)$ $\frac{1}{\mu_{p_j^{*}(s)}} - \frac{1}{\mu_{p_i^{\omega}(s)}}$ more days than the bed should have been occupied by patient $p_i^{\omega}(s)$. Under the myopic policy, each patient $p_j^{*}(s)$ can cause at most one more patient to be rejected a total of $\frac{1}{\mu_{p_j^{*}(s)}} - \frac{1}{\mu_{p_i^{\omega}(s)}}$ times, because the ICU bed could have been freed during these periods if the optimal policy had been followed at state $s$. One more patient rejected a total of $\frac{1}{\mu_{p_j^{*}(s)}} - \frac{1}{\mu_{p_i^{\omega}(s)}}$ times results in an extra cost of at most $\frac{1}{\lambda_m}(\frac{1}{\mu_{p_j^{*}(s)}} - \frac{1}{\mu_{p_i^{\omega}(s)}})$, for a patient of type $m \in \mathcal{M}$. Because $B$ is the maximum

number of patients, $p_i^{\omega}(s)$, who should not have been rejected under the optimal policy, the highest additional cost that could occur by following the myopic policy at state $s$ is $\frac{B}{\lambda_m}(\frac{1}{\mu_{p_j^*(s)}} - \frac{1}{\mu_{p_i^{\omega}(s)}})$. Thus, over the entire finite time horizon from periods 1 to $T$, the largest additional cost that could be incurred by following the myopic policy in each period is $\sum_{t=1}^{T} \frac{B}{\lambda_m}(\frac{1}{\mu_{p_j^*(s)}} - \frac{1}{\mu_{p_i^{\omega}(s)}})$. Therefore, we have $J^{\omega}(s) \leq J^*(s) + \sum_{t=1}^{T} \frac{B}{\lambda_m}(\frac{1}{\mu_{p_j^*(s)}} - \frac{1}{\mu_{p_i^{\omega}(s)}}) \leq J^*(s) + \frac{BT}{\lambda_{min}}(\frac{1}{\mu_{min}} - \frac{1}{\mu_{max}})$. □

From Theorem 2, we see that the myopic policy will perform best when there is little difference between the shortest and longest expected length-of-stays (i.e., $\mu_{max} \approx \mu_{min}$) and when arrival rates are high (i.e., $\lambda_{min} >> 0$). As $\lambda_{min}$ increases (or, equivalently, $B$ decreases), the utilization of the ICU increases. For very high utilizations, surgery requests will be rejected regardless of the admission policy and the immediate costs of rejecting surgeries dominates. Thus, the myopic policy performs closer to the optimal policy.

Theorem 2 also suggests cases where the myopic policy can provide optimality. As $\epsilon \to 0$ for $\mu_{max} = \mu_{min} + \epsilon$, $\frac{BT}{\lambda_{min}}(\frac{1}{\mu_{min}} - \frac{1}{\mu_{max}}) \to 0$. We, therefore, have the following corollary stating that the myopic policy is optimal if all patient-types have the same mean LOS in the ICU.

COROLLARY 1. *If $\mu_1 = \mu_2 = ... = \mu_M$, the myopic policy is optimal.*

*Proof* Assume $\mu_1 = \mu_2 = ... = \mu_M$, thus, $\mu_{max} = \mu_{min}$. According to Theorem 2, for all states $s$, $J^{\omega}(s) \leq J^*(s) + \frac{BT}{\lambda_{min}}(\frac{1}{\mu_{min}} - \frac{1}{\mu_{max}}) = J^*(s)$. Because $J^*(s)$ is optimal, $J^*(s) \leq J^{\omega}(s)$. Thus, $J^*(s) = J^{\omega}(s)$, i.e., the myopic policy is optimal. □

*Remark*: Intuitively, we also expect that the cost of the myopic policy will be very close to the cost of the optimal policy when the arrival rate is low and the number of ICU beds is large (i.e., ICU utilization is low). This is because very few patients will be rejected in such a setting which limits the opportunities for the myopic policy to differ from the optimal policy. This is not reflected in our expression in Theorem 2 because there our worst case assumes that patients are rejected each period regardless of ICU utilization.

## 5. Performance Evaluation

The results in Section 4.1 demonstrate the situations where the myopic policy is optimal and gives worst-case bounds on solution quality of the myopic policy. We suspect that the worst-case bounds

are very conservative, and that the actual performance of the myopic policy is much better in reality. In this section, we evaluate the actual performance of the myopic policy through numerical experiments. We compare the myopic policy with the optimal policy in a set of small examples where the optimal policy can still be computed using traditional dynamic programming solution techniques.

We consider a small example with two types of surgery requests, each performed by a different surgeon ($M = D = 2$), two ICU beds ($B = 2$), a time period of ten days ($T = 10$), and where the largest arrival batch size for each patient type is two ($I_1 = I_2 = 2$). We also assume that the surgical arrival types and patient classes are identical so that $m = d$. (Note that there are over 25,000 states even in this small example.) We assume that the batch arrival sizes of type-1 and type-2 patients follow Poisson distributions with means $\lambda_1$ and $\lambda_2$ respectively, and that the LOS of type-1 and type-2 patients are exponentially distributed with means $\mu_1$ and $\mu_2$, respectively. We assume that the ICU begins the time horizon with all beds available. We allow $\lambda_m \in (0, 1.2]$ and $\mu_m \in [1, 5]$ $\forall m \in \mathcal{M}$, and we randomly draw 100 scenarios with different values of arrival rates and LOS, so that we can capture a spectrum of situations including similar/different patients (in terms of arrival and LOS distributions) and highly/poorly utilized ICUs. We use the conventional value-iteration algorithm to identify optimal solutions.

Figure 2 demonstrates the percentage differences in terms of expected total costs of the optimal and the myopic policies over the 100 trials. For illustration purposes, we construct an empirical cumulative distribution function for the observed percentage difference between the myopic policy and the optimal policy. As seen in Figure 2, 94% of the scenarios result in myopic-policy performance that is within 1% of the optimal solution cost, and the worst case among the 100 scenarios is within 5.38% of the optimal.

## 6. Case Study: ICU in Mount Sinai Medical Center

In this section, we apply our myopic solution method for ICU admissions to an actual surgical-ICU setting. The director of the cardiothoracic-surgical ICU at Mount Sinai Medical Center receives

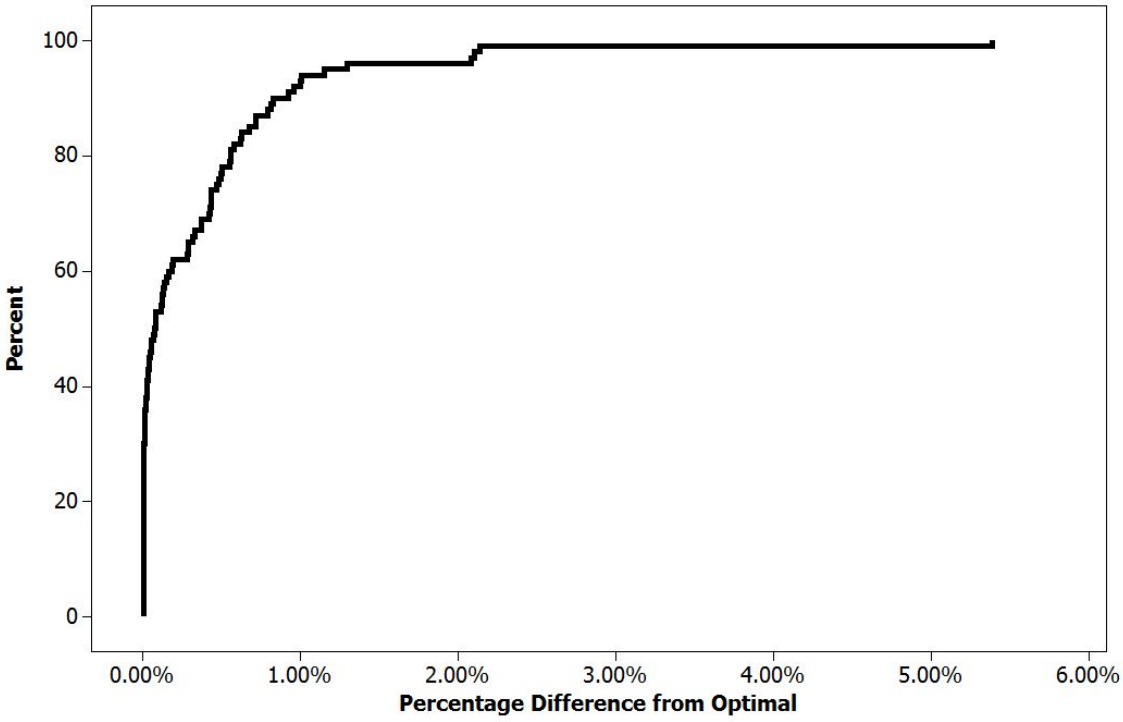


**Figure 2 Empirical CDF of Percentage Difference from Optimal**

surgery requests at the beginning of each weekday, Monday through Friday. More than 20 different types of cardiothoracic-surgery procedures are performed in the cardiothoracic operating rooms. Through a combination of statistical analysis and expert opinion, we group all cardiothoracic-surgical procedures into six major types that have similar arrival and LOS characteristics: Mitral Valve, Aortic Valve, Coronary Artery Bypass Graft Surgery, Ascending Aorta Surgery, Major Thoracic, and Cardiac Other. The surgery procedures within a type are performed by a same surgeon. Table 1 shows the six surgery types and the corresponding procedures in each type.

We collected data for this cardiothoracic-surgical ICU at Mount Sinai Medical Center for six consecutive months. The data collected included the number of each type of surgery request on each day and each individual actual LOS in the ICU. The surgery requests arrive as a batch each weekday. The batch size of each surgery type follows a different distribution that is also dependent on the day of week, Monday through Friday. Different surgery types have different distributions of LOS. The empirical data show that the distributions of batch size are non-Poisson and the

**Table 1 Six Major Cardiothoracic Surgery Types**

| Surgery Types | Procedures |
|---|---|
| Mitral Valve (MV) | Mitral Valve, Mitral Valve/CABG |
| Aortic Valve (AV) | Aortic Valve, Aortic Valve/CABG, Aortic Valve/Mitral Valve/Tricuspid Valve |
| CABG | Coronary Artery Bypass Graft Surgery |
| Ascending Aorta Surgery (AA) | Circ Arrest, Bentall, Arch, Ascending Aneurysm, Elephant Trunk Stage I |
| Major Thoracic (MT) | VATS, Pneumonectomy, Thymectomy |
| Cardiac Other (CO) | Atrial Myxoma, Myocardial Resection, Pulmomary Embolectomy, ASD Resection, VSD resection, etc. |

distributions of LOS are non-exponential. In fact our analysis failed to find any existing distribution that fit the data acceptably. We, therefore, fit empirical distributions to the data of batch sizes. For the LOS data, we employ mixed empirical distributions (see Shanker and Kelton 1991) in order to capture the tail of the LOS distribution. For each surgery type, we arrange all the LOS data points in ascending order, and we use the first 75% percent of the data to construct the empirical distribution and fit an exponential distribution to the remainder of the 25% of data points.

There are 12 beds in the cardiothoracic-surgical ICU at Mount Sinai Medical Center. Because all cardiothoracic-surgery patients are moved to ICU after surgery, the director cannot approve surgery requests until he ensures that there are enough available ICU beds on that day. If the ICU does not have enough open beds, the director needs to make a decision on which surgery requests should be approved and which should be postponed. Those approved surgeries are performed on the same day, and the patients are admitted to the ICU afterwards. Patients will not be discharged from the ICU until they reach the satisfactory heathy level or decease. Postponed surgeries will reappear as surgery requests on the following day. At Mount Sinai Medical Center, nearly all cardiothoracic surgeries are performed eventually regardless of waiting time. In very rare cases, some surgical patients may go to other healthcare organizations if they wait too long. This event is so rare that we do not consider it; however, our models can be easily modified to include balking patients.

Using these input data, we validated our models based on the observed ICU utilizations for the

time periods of interest. We note that the patient waiting times (i.e., rejection rates) predicted by our model exceed those observed at Mount Sinai Medical Center. Discussion with administrators at Mount Sinai leads us to believe that this is because Mount Sinai will sometimes take "exceptional measures" to make sure that a patient can receive surgery in a timely manner. These measures include explicitly adjusting the surgery schedule and "finding an available ICU bed" (generally from some other ICU in the hospital that can provide equivalent care). However, these actions are not considered desirable and are outside the standard operating procedure at Mount Sinai, thus, we have not included such actions in our model.

To evaluate our myopic admission policy, we introduce two competing admission policies that are common in practice and in the literature: a random policy and a shortest-processing time policy. A random admission policy is currently used in the cardiothoracic-surgical ICU of Mount Sinai Medical Center. Due to the batch arrivals of all types of surgery requests at the same time on each day, a simple First-Come-First-Serve rule cannot be applied in this situation. To be fair to all surgery requests, the director of the cardiothoracic-surgical ICU randomly rejects surgery requests whenever the ICU can not accommodate all requests. The shortest-precessing-time (SPT) rule is popular in manufacturing settings, which gives the job with shortest expected processing time the highest priority. The SPT rule has been proven to maximize the throughput of a system in a deterministic environment, and usually performs well in stochastic environments even though in only very limited settings can it be proven to be optimal. The downside of the SPT rule is that jobs with the longest processing time are consistently rejected. The random admission rule can be considered as a rule emphasizing equity among different surgery requests, whereas the SPT rule can be considered as a rule emphasizing the efficiency of the ICU. We use these two rules as benchmarks to compare our myopic rule, which seeks to balance equity and efficiency in the ICU.

In addition to our suggested performance metric in (1), we introduce two other common performance metrics: the average waiting time $\bar{W}$ and the maximum observed waiting time $W_{max}$. Let $w_i$ be the waiting time of the $i^{th}$ arrival and $v$ be the total arrivals in a sample path. Note that because patients can be rejected a maximum of once per time period, the waiting time of patient $i$,

$w_i$, also gives the number of rejections for surgical request $i$. The average waiting time in a sample path is thus

$$\bar{W} = \frac{\sum_{i=1}^{v} w_i}{v};$$

and the maximum observed waiting time in a sample path is

$$W_{max} = \max_i w_i.$$

The overall average waiting time measures the efficiency of the ICU, whereas the maximum waiting times measure the equity among surgeries requests and the expected total cost reflects the balance between equity and efficiency.

We perform 100 replications of a simulation for each admission policy. Each replication is a realization of ten-years of operations in the ICU based on our historical data, i.e, 3650 days, with a warmup period of one year. Different streams of random number are used for each admission policy to maintain independence. Table 2 shows the average values of the three performance metrics with the corresponding 95% confidence intervals under each admission policy.

**Table 2 Admission Control Policies Comparison**

| Admission Policy | 10-year Total Cost | $W$ (Days) | $W_{max}$ (Days) |
|---|---|---|---|
| Myopic | 2321.78±44.76 | 0.5632±0.01 | 18.26±1.12 |
| Random | 3737.19±75.88 | 0.5061±0.01 | 31.71±2.35 |
| SPT | 4833.38±113.42 | 0.4612±0.01 | 30.5±2.44 |

Table 2 shows that the myopic policy is approximately 38% and 52% better than the other policies in terms of the 10-year total cost as expressed by (1)—a measure of the balance between efficiency and equity. As expected, the myopic policy outperforms the random policy and the SPT policy under this metric. As we focus on the metric of average waiting time — representing the system overall efficiency, the myopic policy is about 11% and 22% worse than the random policy and the SPT policy. Certainly, this is the price that the myopic policy must pay to seek the trade-off between efficiency and equity. The 22% "efficiency loss" indicated here means that the myopic policy results in a surgery being delayed, on average, approximately 2.5 hours longer than it would

be under the SPT policy. Recall that the minimum unit of time considered by our model for this ICU is one day. Thus, 2.5 hours on average is most likely tolerable in practice. We believe that the maximum waiting time can be treated as an equity metric. Table 2 provides the observed maximum waiting time among all six surgery types across the three admission policies. The myopic policy is about 40% better than the other policies, which means that the longest delay for a surgery request can be shortened by 13 days; we call this the "equity gain." Comparing the "efficiency loss" and the "equity gain," the myopic policy provides the best trade-off between efficiency in the ICU and equity among surgery types.

**Table 3 Average Waiting Time by Surgery Type (Days)**

| Policy | MV | AV | CABG | AA | MT | CO |
|---|---|---|---|---|---|---|
| Myopic | 0.7459±0.02 | 0.5686±0.01 | 0.6389±0.01 | 0.0305±0.00 | 0.0500±0.01 | 0.3530±0.01 |
| Random | 0.5094±0.01 | 0.5160±0.01 | 0.5137±0.01 | 0.4339±0.02 | 0.5260±0.02 | 0.4820±0.01 |
| SPT | 0.2421±0.01 | 0.8607±0.01 | 0.0400±0.00 | 0.4030±0.01 | 0.1092±0.00 | 1.5404±0.05 |

The random policy used in practice is expected to give more weight to providing equity among patient types. From Table 3, we see that the random policy results in average waiting times for patient types between 0.4339 to 0.5260 days (about 2.2 hours). On the contrary, the average waiting time for each patient type ranges from 0.0305 to 0.7459 days (about 17 hours) under the myopic policy and ranges from 0.0400 to 1.5404 days (about 36 hours) under the SPT policy. Due to the nature of the SPT policy—constantly rejecting the patient type with the longest average LOS—it is not surprising that one patient type has much longer delay than the others.

**Table 4 Maximum Observed Waiting Time by Patient Type (Days)**

| Policy | MV | AV | CABG | AA | MT | CO |
|---|---|---|---|---|---|---|
| Myopic | 17.55±1.17 | 16.59±1.22 | 18.26±1.12 | 4.17±0.81 | 5.16±0.68 | 14.42±1.07 |
| Random | 31.71±2.35 | 29.34±2.32 | 29.91±2.26 | 19.49±2.03 | 20.91±2.05 | 26.17±2.06 |
| SPT | 6.14±0.31 | 14.97±0.77 | 1.90±0.15 | 7.21±0.51 | 2.92±0.16 | 30.51±2.44 |

The inequity resulting from different admission control policies is even more evident when examining the maximum observed waiting time by patient type (see Table 4). The range of the maximum observed waiting times among patient types are approximately 14 days, 12 days, and 28 days under

the myopic policy, the random policy, and the SPT policy respectively. The SPT policy increases the system efficiency at the cost of a particular patient type (type CO in this setting). The random policy, however, pursues system equity by prolonging every patient type's waiting time. The myopic policy effectively balances these two issues.

**Table 5 Average Waiting Time by Patient Type (Days) with 13 beds**

| Policy | MV | AV | CABG | AA | MT | CO |
|---|---|---|---|---|---|---|
| Myopic | 0.4088±0.01 | 0.2892±0.01 | 0.3393±0.01 | 0.0075±0.00 | 0.0142±0.01 | 0.3530±0.01 |
| Random | 0.2590±0.01 | 0.2757±0.01 | 0.2736±0.01 | 0.2153±0.02 | 0.2731±0.01 | 0.2527±0.01 |
| SPT | 0.1317±0.00 | 0.4592±0.01 | 0.0236±0.00 | 0.2165±0.01 | 0.0634±0.00 | 0.8013±0.02 |

As mentioned previously, most ICUs operate at very high utilization, making the ICU the bottleneck of patient flow throughout the hospital system. To increase patient flow, the most straight forward action would be to increase the capacity of the ICU, i.e., to increase the bottleneck rate. Unfortunately, it is often impractical for hospitals to increase the capacity of ICUs do to the extreme expense of increased staffing and equipment to add ICU beds as well as due to space constraints in most hospitals. Additionally, ICUs are generally seen as a cost center and hospitals often make decision to increase higher revenue generating subsystems. However, it is still informative to examine the effect of ICU capacity (or, equivalently, ICU utilization) on the choice of admissions policies and their performance.

Suppose the hospital is able to expand the capacity of the ICU from 12 to 13 beds. Tables 5 and 6, respectively, show the average waiting times and the maximum observed waiting times across all patient types for an ICU with 13 beds. All arrival and LOS processes remain unchanged from our previous analysis. The average and the maximum waiting times in the ICU with 13 beds are approximately 50% shorter than those in the ICU with 12 beds, regardless of the admission policy used. Otherwise, the three admission policies behave similarly in the 13-bed ICU as they did in the 12-bed ICU.

Figure 3 reinforces the effect of ICU capacity (utilization) on admission policy performance. Figure 3 shows the average observed maximum waiting time for each patient type under various

**Table 6 Maximum Observed Waiting Time by Patient Type (Days) with 13 beds**

| Policy | MV | AV | CABG | AA | MT | CO |
|---|---|---|---|---|---|---|
| Myopic | 11.56±0.74 | 10.43±0.74 | 11.33±0.84 | 1.58±0.35 | 2.15±0.36 | 7.78±0.69 |
| Random | 16.92±1.32 | 16.01±1.29 | 16.62±1.31 | 9.98±0.94 | 11.29±1.30 | 14.23±1.35 |
| SPT | 4.71±0.19 | 9.63±0.60 | 1.39±0.12 | 4.90±0.27 | 2.33±0.15 | 16.12±1.27 |

ICU capacities (from 12 beds to 15 beds). As shown in the figure, for any given number of ICU beds, the SPT policy results in the largest difference among maximum waiting times across patient types; the random policy generates the smallest difference of maximum waiting times by prolonging the waiting times of all patient types. Similar to our previous results, the myopic policy reduces the difference of maximum waiting times compared to the SPT policy and also reduces the waiting times of all patient types in comparison to the random policy. Notably, as the total ICU capacity increases (i.e., the ICU becomes less utilized), the advantages of the myopic policy compared to the SPT and random policies in terms of efficiency and equity diminish. As ICU capacity increases, the myopic policy is still the best among the three in terms of balancing equity and efficiency; however, the differences among the three policies are greatly reduced. Note also that the ordinal ranking of the patient types based on average maximum observed waiting times remains relatively constant for each admission policy across ICU capacities. In general, CABG and MV patients experience the longest waits in the myopic and random policies, while CO patients experience much longer waiting times under SPT, independent of the capacity of the ICU.

## 7. Conclusions and Future Work

Intensive care units—the most staffed and best equipped unit in most hospitals—are almost always heavily utilized and overcrowded. ICUs are frequently the bottleneck of the hospital system, limiting patient flow and negatively impacting profits. As healthcare organizations, however, hospitals not only pursue system efficiency, but also equity among patients and surgeons. This paper studies admission policies for surgical-based ICUs with the goal of balancing system efficiency and equity. We focus on the ICU admission control process and model it as a Markov decision process. We propose a simple and efficient algorithm to solve the MDP with a performance guarantee, which provides periodic admission decisions for ICUs. Our numerical results demonstrate that this

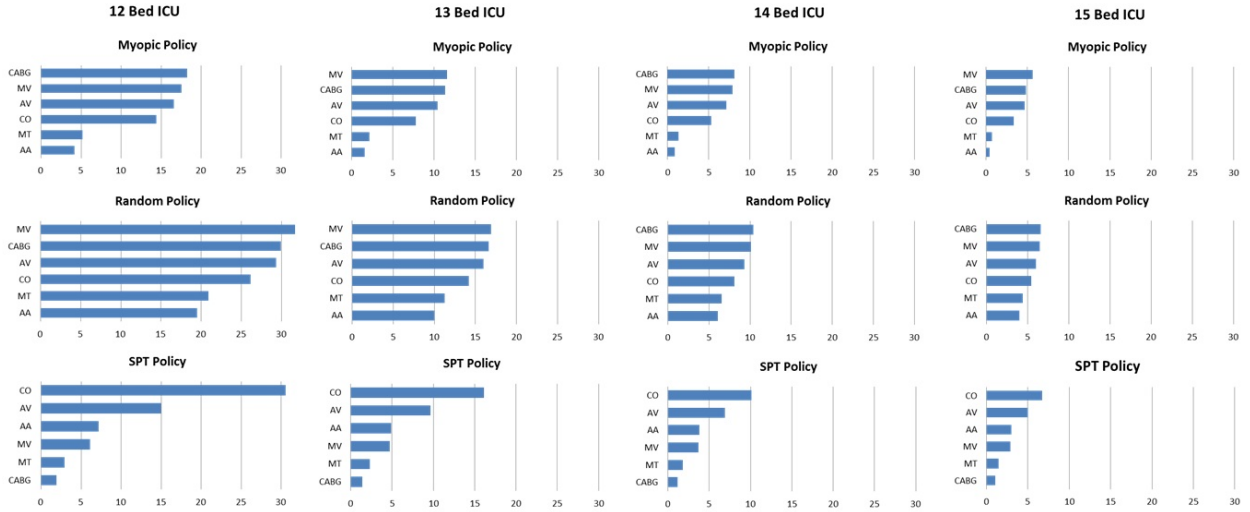


**Figure 3 Maximum Observed Waiting Time by Patient Type (Days) for ICU Capacities of 12 to 15 beds**

admission control policy effectively balances the efficiency and equity metrics facing hospitals and significantly outperforms competing policies.

At a more general level, we study a dynamic process with multiple customer classes and batch arrivals with the objective of providing efficiency and equity. The batch arrival process has fixed interarrival times but random arrival sizes for each customer class. Moreover, any rejected arrivals will return at the next arrival epoch, and will continue to do so until they complete service. We identify special cases where our myopic policy is optimal, and we provide worst-case bounds on solution quality. Our numerical experiments on small-sized problems (where the optimal solution is still computable) show that our policy is within 1% of the optimal 94% of the time.

Practically, we apply our admission policy to the cardiothoracic-surgical ICU at the Mount Sinai Medical Center using real data on daily surgery requests across patient types and actual LOS of every patient. Comparing this with the current admission policy used by Mount Sinai Medical Center, we show that our policy could have reduced the maximum wait time by 13 days at the cost of increasing the average waiting time by only 1.3 hours. Our results show that the ICU can achieve large equity gains with limited efficiency losses. Our policy is easy to understand and is easily implemented in a spreadsheet environment as a decision support tool available for practitioners.

An extension of this work could include a joint decision-making process that considers both

the scheduling of surgeries and admissions into the ICU. The current process at Mount Sinai's cardiothoracic ICU (and, indeed, at many ICU's that deal with patients from scheduled surgeries) is that the intensivist in charge of admissions to the ICU (our decision maker) does not have visibility or control of surgeries scheduled more than one day into the future. However, such an arrangement is clearly suboptimal and could be improved through better planning and communications. If the intensivist could schedule, say, one week's worth of surgeries for patients who will then be admitted to the ICU, then it is very likely that both efficiency and equity measures could be improved. The intensivist could use knowledge of the different surgeries' expected LOS in making the scheduling decisions; for instance, s/he should be able to utilize weekends more intelligently where patients can recover in the ICU, but there are no scheduled arrivals. This is not possible in the current decision making procedure used at Mount Sinai, but most parties realize that improved ICU performance could be gained by altering the cardiothoracic-surgery scheduling process. The resulting models are quite complex, but similar methodologies could be utilized to develop near-optimal policies in this setting.

## References

Cahill, W., M. Render. 1999. Dynamic simulation modeling of ICU bed availability. *Proceedings of the 1999 Winter Simulation Conference*. 1573–1576.

Chan, C. W., V. F. Farias, N. Bambos, G. J. Escobar. 2010. Maximizing throughput of hospital intensive care units with patient readmissions. Working Paper, http://www.columbia.edu/∼cc3179/ICU_2010.pdf.

De Bruin, A. M., A. C. Van Rossum, M. C. Visser, G. M. Koole. 2007. Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science* **10**(2) 125–137.

Dobson, G., H.-H. Lee, E. Pinker. 2010. A model of icu bumping. *Operations Research* **58**(6) 1564–1576.

Federgruen, A., H. Groenevelt. 1988. $M/G/c$ queueing systems with multiple customer classes: Characterization and control of achievable performance under nonpreemptive priority rules. *Management Science* **34**(9) 1121–1138.

Gosden, T., L. Pedersen, D. Torgerson. 1999. How should we pay doctors? A systematic review of salary payments and their effect on doctor behaviour. *QJM: An International Journal of Medicine* **92** 47–55.

Green, L. 2006. Queueing analysis in healthcare. R. W. Hall, ed., *Patient Flow: Reducing Delay in Healthcare Delivery*, *International Series in Operations Research and Management Science*, vol. 91. Springer US, 281–307.

Green, L. V. 2002. How many hospital beds? *Inquiry* **39**(4) 400–412.

KC, D., C. Terwiesch. 2007. An econometric analysis of patient flows in the cardiac ICU. Working Paper, The Wharton School, University of Pennsylvania, http://www.gsb.stanford.edu/facseminars/events/oit/documents/oit_03_08_diwas.pdf.

Kim, S., I. Horowitz. 2002. Scheduling hospital services: the efficacy of elective-surgery quotas. *Omega* **30**(5) 335–346.

Kim, S., I. Horowitz, K.K. Young, T.A. Buckley. 1999. Analysis of capacity management of the intensive care unit in a hospital. *European Journal of Operational Research* **115**(1) 36–46.

Kim, S., I. Horowitz, K.K. Young, T.A. Buckley. 2000. Flexible bed allocation and performance in the intensive care unit. *Journal of Operations Management* **18** 427–443.

Kolker, A. 2009. Process modeling of icu patient flow: Effect of daily load leveling of elective surgeries on icu diversion. *J Med Syst* **33** 27–40.

Lowery, J. C. 1993. Multi-hospital validation of critical care simulation model. *Proceedings of the 1993 Winter Simulation Conference*. 1207–1215.

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

McManus, M. L., M. C. Long, A. Cooper, E. Litvak. 2004. Queuing theory accurately models the need for critical care resources. *Anesthesiology* **100**(5) 1271–1276.

Min, D., Y. Yih. 2010. Scheduling elective surgery under uncertainty and downstream capacity constraints. *European Journal of Operational Research* **206** 642–652.

Pronovost, P. J., D. M. Needham, H. Waters, C. M. Birkmeyer, J. R. Calinawan, J. D. Birkmeyer, T. Dorman. 2004. Intensive care unit physician staffing: Financial modeling of the leapfrog standard. *Critical Care Medicine* **32**(6) 1247–1253.

Shanthikumar, J. G., D. D. Yao. 1992. Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. *Operations Research* **40**(2) S293–S299.

Shmueli, A., C. L. Sprung, E. H. Kaplan. 2003. Optimizing admissions to an intensive care unit. *Health Care Management Science* **6** 131–136.

Smith-Daniels, V., S. Schweikhart, D. Smith-Daniels. 1988. Capacity management in health care services: review and future research directions. *Decision Sciences* **19** 889–919.

Swenson, M.D. 1992. Scarcity in the intensive care unit: Principles of justice for rationing icu beds. *American Journal of Medicine* **92** 552–555.

Tor Schoenmeyr, M.S., P. F. Dunn, D. Gamarnik, R. Levi, D. L. Berger, B. J. Daily, W. C. Levine, W. S. Sandberg. 2009. A model for understanding the impacts of demand and capacity on waiting time to enter a congested recovery room. *Anesthesiology* **110**(6) 1293–1304.