

Submitted to
manuscript (Please, provide the manuscript number!)

Maximizing Throughput of Hospital Intensive Care Units with Patient Readmissions

Carri W. Chan

Department of Electrical Engineering, Stanford University cwchan@stanford.edu

Vivek F. Farias

Sloan School of Management, Massachusetts Institute of Technology vivekf@mit.edu

Nicholas Bambos

Departments of Electrical Engineering and Management Science & Engineering, Stanford University bambos@stanford.edu

Gabriel J. Escobar

Kaiser Permanente Division of Research, gabriel.escobar@kp.org

This work examines the impact of discharge decisions under uncertainty in a capacity-constrained high risk setting: the intensive care unit (ICU). New arrivals to an ICU are typically very high priority patients and, should the ICU be full upon their arrival, discharging a patient currently residing in the ICU may be required to accommodate a newly admitted patient. Patients so discharged risk physiologic deterioration which might ultimately require readmission; models of these risks are currently unavailable to providers. These readmissions in turn impose an additional load on capacity-limited ICU resources.

The present work studies the impact of different ICU discharge strategies on total readmission load. Our study focuses on a certain index policy for discharge that is predicated on a model of readmission risk. We use empirical data from over 6000 actual ICU patient flows to calibrate our model and judge the efficacy of our approach relative to several benchmark strategies. The empirical study suggests that a predictive model of the readmission risks associated with discharge decisions in tandem with simple index policies of the type proposed can provide very meaningful throughput gains in actual ICUs. In addition to our empirical work, we conduct a rigorous performance analysis for our discharge policy. We show that our policy is optimal in certain regimes, and is otherwise guaranteed to incur readmission loads no larger than a factor of $(\hat{\rho} + 1)$ of an optimal discharge strategy, where $\hat{\rho}$ is a certain natural measure of system utilization.

1. Introduction

The intensive care unit (ICU) is the designated location for the care of the sickest and most unstable patients in a given hospital. These units are among the most richly staffed in the hospital: for example, in California, licensed ICUs must maintain a minimum nurse-to-patient ratio of one-to-two. Critically ill patients, who may be admitted to a hospital due to multiple illnesses, including trauma, need urgent admission to the ICU. While it is possible to hold these patients in other areas (e.g., the emergency department) pending bed availability, this is quite undesirable, since delays in providing intensive care are associated with worse outcomes (Chalfin et al. 2007). Consequently, in such situations, clinicians may elect to discharge a patient currently in the ICU to make room for

a more acute patient. For the sake of precision, we will refer to this as a demand-driven discharge. In theory, the patient selected for such discharge would be one who was sufficiently stable to be transferred to a less richly staffed setting¹, and, ideally, the term ‘stable’ would be one based on ample clinical data. In practice, since predictive models of patient dynamics are not readily available, clinicians must make these transfer decisions based entirely on clinical judgment. At the same time, patients so discharged potentially face additional risks of physiological deterioration which might ultimately require readmission. These readmissions in turn impose an additional load on capacity-limited ICU resources. Even worse, readmitted patients tend to require longer stays in the ICU and have a higher mortality rate than first-time patients (see Snow et al. (1985), Durbin and Kopel (1993)). The present work thus examines the potential benefits of a quantitative decision support system for clinicians when faced with the requirement to identify a patient for discharge in order to make room for a more acute patient. The hope is that the availability of such a system could lead to both increased efficiencies in the use of scarce ICU resources and implicitly, better patient outcomes.

More formally, we study the problem of ‘optimally’ discharging patients when faced with the need to do so in order to accommodate more critical patients. Our study is predicated on a predictive model of the likelihood that a given patient so discharged will require readmission; we will eventually estimate such a model from actual patient data. Our goal is to maximize ICU throughput, i.e. treat the most ICU patients back to health as possible. We consider a stylized model of an actual ICU where the number of ICU beds is fixed². Patients arrive to the ICU at random times. All new arrivals must be given an ICU bed immediately; they cannot queue up and wait for a bed to become available. This models the aforementioned fact that new ICU patients are typically extremely high priority. If no beds are vacant upon the arrival of a new patient, a current patient will have to be discharged in order to accommodate the newly arriving patient. This discharged patient may subsequently deteriorate and return to the ICU, imposing an additional load on the ICU beds. The objective is to maximize the number of patients who enter, exit and benefit from the ICU. Equivalently, the goal is to maximize the throughput of the ICU. Due to the complexity of the associated stochastic model whose state space requires we track all arrivals over time, identifying the throughput region and the corresponding optimal discharge policy is difficult and so we settle for a related objective that is somewhat more tractable (but nonetheless still yields a hard problem).

¹ such as the Transitional Care Unit (TCU) or Medical Surgical Floor (Floor)

² Since a strict (one-to-two in California) nurse-to-patient ratio must be maintained, it is often the size of the nursing staff that determines the number of available ICU beds rather than the actual number of physical beds which are available.

In particular, we study the problem of minimizing the total expected readmitted load incurred due to demand-driven patient discharge over some finite horizon.

We make the following key contributions:

- We identify a simple ‘myopic’ discharge strategy that corresponds to an index policy: every patient class³ is associated with a class specific index. The index for a given class can be computed from historical patient flow data in a robust fashion. When a patient must be discharged in order to accommodate new patients, the strategy simply discharges an existing patient of a class with the lowest possible index.

- We calibrate our model to empirical data from over 6000 patient flows at a large privately owned partnership of hospitals and identify parameters for patient dynamics. We examine the impact of using our discharge rule in place of a number of alternatives, some of which resemble the status quo. We show that our policy can have a substantial impact on ICU throughput, providing an increase of about 10% even under modest assumptions on patient traffic; clinicians currently do not have access to the type of predictive models we estimate nor the sort of decision support tool we develop.

- Our index policy is ‘robust’: In particular the indices we compute are oblivious to patient traffic intensities which are highly variable and difficult to estimate. Rather, they rely on a relatively simple to estimate model that yields the likelihood that a demand-driven patient discharge will result in readmission given the class of the patient, and the average load imposed by such a readmission. For the data set under consideration, relative changes of estimated parameters greater than 100% were typically required to impact a change in the patient class priorities induced by the corresponding index rule.

- We demonstrate via a theoretical analysis that our index policy is, for a certain class of problems, optimal and in general incurs total expected readmission load that is no more than $1 + \hat{\rho}$ times that incurred under an optimal discharge rule, where $\hat{\rho}$ is a certain natural measure of ICU utilization.

As such, this study identifies a discharge procedure that allows us to utilize available ICU resources as effectively as possible. A second related benefit is that we may use our procedure to hypothetically determine staffing levels that would allow an ICU to meet target service levels measured in total expected readmission load. At a high level, our analyses suggest that investments in providing clinicians with more decision support (e.g., severity of illness scores and the associated risks of physiological deterioration) could translate into tangible benefits both in terms of improved

³ There exist a number of proprietary classification systems; patients within a class are relatively homogenous.

patient outcomes, increased efficiency, and decreased costs; we present a concrete path to achieving these benefits.

1.1. Related Literature

In a recent econometric study (Diwas and Terwiesch 2007), patient discharge was shown to be a legitimate cause of patient readmission thereby effectively reducing peak ICU capacity due to the additional load the readmitted patients bring. The empirical data we have analyzed in calibrating our ICU model corroborates this fact.

When a new patient arrives to the ICU, either after experiencing some trauma or completing surgery, s/he must be admitted. If there are not enough beds available, space must be allocated by transferring current patients to units with lower levels of staffing and care. In Swenson (1992) and related works, the authors examine how to allocate ICU beds from a qualitative perspective that is not based on analysis of patient data but rather on philosophical notions of ‘fairness’. The authors propose a 5-class ranking system for patients based on the amount of care required by the patient as well as his/her risk of complications. Our approach may be seen as a quantitative perspective on the same problem wherein decisions are motivated by the analysis of relevant quantitative patient data. To date, the work (particularly in the medical community) on how to determine discharge decisions has been rather subjective due to the lack of information-rich models which attempt to capture patient dynamics. Thus, these works have not considered that by discharging a patient from the ICU in order to accommodate new patients may result in readmission, further increasing demand for the limited number of beds. We not only propose such a model, but also show the efficacy of discharge policies which utilize this previously unavailable information.

Dobson et al. (2009) consider a setup quite similar to ours but ignore the readmission phenomenon; rather they simply seek to quantify the total expected number of patients discharged in order to accommodate new, more critical patients. To this end they analyze a policy that chooses to discharge patients with the shortest remaining service time (which are modeled as deterministic quantities). As will be seen in Section 4, which presents an empirical performance evaluation using a real patient flow data-set, a distinct heuristic is desirable when one does account for patient readmission.

A number of modeling approaches have been used to make capacity, staffing and other tactical decisions in the healthcare arena (see for instance Huang (1995), Kwak and Lee (1997), and Green et al. (2003)). Queueing theory has been particularly useful to study the question of necessary staffing levels in hospitals. As examples of this work, Green et al. (2006) and Yankovic and Green (2008) consider a number of staffing decisions from a queueing perspective. The goal is to provide

patients with a particular service level (in terms of timeliness, and also nurse-to-patient ratio) while at the same time addressing issues such as temporal variations in arrival rates of patients of different types. See also Green (2006) for an overview of the use of OR models for capacity planning in hospitals. Murray et al. (2007) consider different factors such as age, gender, physician availability and number of visits per patient per year to determine the largest patient panel size that may be supported by available resources. In Savin (2006), the authors consider how to reduce delay in primary care settings by varying the number of patients served by the particular primary care office. When a patient wishes to make an appointment, s/he may be delayed before the physician is able to see him/her. Two significant differences separate the problem we consider from those considered in the above streams of work: arriving patients to an ICU must receive service immediately (which thus necessitates discharging current patients). This in turn requires that we consider individual patient dynamics, and in particular model the impact of discharging a patient to accommodate new ones on the discharged patient's likelihood of revisiting the ICU. We can then make staffing decisions in much the same way as the aforementioned work.

From a modeling and optimization stand-point, this work bears some similarity to the stochastic depletion framework introduced in Chan and Farias (2009). There are subtle differences in the dynamics necessitated by the model we study here and our cost criterion that do not allow for the application of the results presented in that work.

The rest of the paper proceeds as follows. In Section 2, we formally introduce a model of patient dynamics and an associated queueing system and proceed to state our problem precisely. In Section 3, we consider the performance of an index policy which selects patients to discharge in the greedy order of the expected readmission loads they impose. We present conditions under which the proposed greedy policy is, in fact, optimal. Moreover, we show that in general the greedy policy is guaranteed to be within a factor of $(\hat{\rho} + 1)$ of optimal, where $\hat{\rho}$ is a measure of the system load; a brief numerical study in Section 3.4 suggests that this gap is likely to be smaller in parameter regimes of interest. In Section 4, we discuss the calibration of our model using a proprietary ICU patient flow data-set from a group of private hospitals. Having calibrated our model, we show in Section 5 that the greedy policy outperforms a number of benchmarks of interest by a meaningful margin. We conclude in Section 6.

2. Model

We begin by proposing a stylized model of the patient flow dynamics in a hospital ICU. We account for the fact that discharging a current ICU patient in order to accommodate a new one results in

increased risks of the discharged patient requiring readmission. Such a readmission would, in turn, result in increased consumption of ICU resources down the road. Since arriving patients cannot be queued or blocked, the model we consider is distinct from a typical queueing/ loss system model. A natural goal is to find a patient discharge policy that maximizes ICU ‘throughput’. This proves to be a somewhat daunting task, and we settle for the related, but simpler task, of minimizing the total expected workload incurred due to patient readmission.

Preliminaries: We consider time to be discrete and indexed by $t \in [0, T]$. In each time-slot, we must determine if a patient must be discharged and, if so, which one. If there are enough available beds to accommodate all current and arriving patients, discharge of current patients is not required.

We assume that patients may be classified into one of M classes, each potentially corresponding to the particular ailment/health condition of the ICU patient. Let $m \in \mathcal{M} = \{1, 2, \dots, M\}$ denote the type of a particular patient. Patients from a given class are assumed to have identical statistics for their initial lengths of stay, the likelihood of readmission upon a demand-driven discharge, and their length-of-stay upon readmission. Specifically, we assume that the initial length-of-stay for a patient of class m is a geometric random variable with mean $1/\mu_m^0$. If such a patient is discharged prior to completing treatment due to the arrival of a more acute patient, he will return to the ICU with probability p_m and his expected length-of-stay upon readmission is a geometric random variable with mean $1/\mu_m^b$. Thus, such a demand-driven discharge of a patient of type m results in an additional expected workload of p_m/μ_m^b due to potential readmission. Such a patient model ignores the possibility that upon relapse the patient may not survive prior to being readmitted; our model can, however, be extended to capture this effect. The patient length-of-stay distribution is assumed to be geometric and thus memoryless. While crude, this serves as a reasonable approximation (see the empirical study in Section 4). That said, in Section 3.3, we discuss an extension to our model which is able to capture a patient’s evolution and changing condition during his ICU stay by using a ‘phase’-type length-of-stay distribution.

At most one new patient can arrive in each time-slot and an arrival occurs with probability λ . We let $a_{t,m}$ denote the probability that a newly arriving patient at time t is of type m . We allow an optimal discharge policy access to this information; the greedy index policy we study will require neither knowledge of λ nor the probabilities $a_{t,m}$.

We assume that the ICU has B beds. If all B beds are full and a new patient arrives, then a patient must be discharged prior to completing service in order to accommodate the newly arrived patient. We let $x_{t,m} \in \{0, 1, \dots, B\}$ denote the number of class m patients currently in the ICU at

the beginning of time-slot t and let $y_{t,m} \in \{0, 1\}$ be an indicator for the arrival of a type m patient at the start of the t th epoch. Note that because at most one new patient can arrive in each time-slot, $\sum_{m=1}^M y_{t,m} \leq 1$ for all t . A current patient must be discharged if $\sum_{m=1}^M x_{t,m} + \sum_{m=1}^M y_{t,m} = B + 1$ —we refer to this type of discharge as a demand-driven discharge. The natural departure (or service completion) of patient type m occurs at the end of the t th time-slot with probability μ_m^0 after any admissions and/or demand-driven discharges occur.

State and Action Space: The dynamic optimization problem we will propose is conveniently studied in a ‘state-space’ model. We define our state-space as the set:

$$\mathcal{S} = \left\{ (x, y, t) : x \in \{0, 1, \dots, B\}^M, \sum_{m=1}^M x_m \leq B, y \in \{0, 1\}^M, \sum_{m=1}^M y_m \leq 1, 0 \leq t \leq T \right\}$$

In particular, the state of the system is completely described by the number of patients of each type currently in the ICU, the type of the arriving patient at that state if any, and the epoch in question. We denote by $x(s)$ the projection of s onto its first coordinate and similarly employ the notation $y(s)$ and $t(s)$. We let the random variable $s_t \in \mathcal{S}$ denote the state in the t th epoch. Note that because the $\{a_{t,m}\}$ process is assumed to be deterministic and given a-priori, the current time slot t completely specifies the arrival probabilities for each patient class.

For each state s , let $\mathcal{A}(s) \subset \mathcal{M}$ denote the set of feasible actions that can be taken in time-slot $t(s)$. For states wherein a demand-driven discharge is required, i.e. states s for which $\sum_m x(s)_m + y(s)_m > B$, we have $\mathcal{A}(s) = \{m : x(s)_m > 0\}$. At all other states s , $\mathcal{A}(s) = \{m : x(s)_m > 0\} \cup \{\emptyset\}$. Thus, an action $A \in \mathcal{A}(s)$ specifies the class of the patient, if any, to be discharged in time-slot $t(s)$; since only one patient can arrive in each time slot, at most one demand-driven patient discharge is required to accommodate a new patient. We will henceforth suppress the dependency of the set of feasible actions, $\mathcal{A}(s)$, on s .

Dynamics: Let $s' = S(s, A)$ denote the random next state encountered upon employing action A (demand-driven discharge of patient type A) in state s . A random number, $X_{t(s),m}$, of class m patients will complete treatment and depart naturally, where $X_{t(s),m}$ is a Binomial- $(x(s)_m + y(s)_m - \mathbf{1}_{\{A=m\}}, \mu_m^0)$ random variable. Let R_t be independent random variables defined for each t indicating the type of an arriving patient at the start of the t th epoch. R_t takes values in $\{1, 2, \dots, M\} \cup \{\emptyset\}$; $R_t = m$ with probability $\lambda a_{t,m}$ for $m \in \{1, 2, \dots, M\}$ and $R_t = \emptyset$ with the remaining probability. The vector denoting arrivals at the next state, $Y_{t(s)+1}$ is then given by $Y_{t(s)+1,m} = \mathbf{1}_{R_{t(s)+1}=m}$. Thus, $s' = S(s, A)$ is defined as:

$$\begin{aligned} x(s')_m &= x(s)_m + y(s)_m - \mathbf{1}_{\{A=m\}} - X_{t(s),m}, \\ y(s')_m &= Y_{t(s)+1,m}, \\ t(s') &= t(s) + 1. \end{aligned}$$

The cost incurred for taking action A is defined by the cost function $C : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}_+$, where $C(s, A) = \frac{p_A}{\mu_A^b}$ for $A \in \{1, 2, \dots, M\}$, and $C(s, \emptyset) = 0$. Recall that p_A is the probability of readmission of patient class A and μ_A^b is her expected service rate upon readmission. Hence, the cost incurred by action A is the expected readmission load due to demand-driven discharge of patient type A .

It is worth remarking that the model we have described is stylized since the impact of our discharge policy on future arrival statistics is ignored. We do this primarily for tractability; a model that accounted for such dependencies will require that we track every patient that ever enters the system. Even with this simplifying feature, the model we are left with remains high-dimensional.

Objective: Let Π denote the set of feasible discharge policies, π which map the state space \mathcal{S} to the set of feasible actions \mathcal{A} . Define the expected total cost-to-go under policy π as:

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

We let $J^*(s) = \min_{\pi \in \Pi} J^\pi(s)$ denote the minimum expected total cost-to-go under any policy. We denote by π^* a corresponding optimal policy, i.e. $\pi^*(s) = \arg \min_{\pi \in \Pi} J^\pi(s)$.

The optimal cost-to-go function (or *value* function) J^* and the optimal discharge policy π^* can in principle be computed numerically via dynamic programming: In particular, define the dynamic programming operator \mathcal{H} according to:

$$(\mathcal{H}J)(s) = \min_{A \in \mathcal{A}} E [C(s, A) + J(S(s, A))]. \quad (1)$$

for all $s \in \mathcal{S}$ with $t(s) \leq T-1$. J^* may then be found as the solution to the Bellman equation $\mathcal{H}J = J$, with the boundary condition $J(s') = 0$ for all s' with $t(s') = T$. The optimal policy π^* may be found as the greedy minimizer with respect to J^* in (1). The size of \mathcal{S} precludes this straightforward dynamic programming approach. Even if optimal solution were possible, the robustness of such an approach and its implementability remain in question since it relies on detailed patient arrival statistics which are typically not stationary and difficult to estimate. As such, our goal will be to design simple, robust heuristics for the load minimization problem at hand.

3. A Greedy Heuristic

This section introduces a myopic policy for the dynamic optimization problem proposed. Under such a policy, the patient selected for a demand-driven discharge is chosen from a patient class that would incur the minimal expected load due to readmission. This readmission load is simply the product of the probability a patient of that class is likely to be readmitted and his expected

length-of-stay should he be readmitted. In particular, such a policy states that the patient (class) $\pi^g(s)$ chosen for discharge satisfies:

$$\pi^g(s) \in \arg \min_{m \in \mathcal{A}(s)} \frac{p_m}{\mu_m^b} \quad (2)$$

It is easy to see that such a policy has a natural implementation as an ‘index’ policy. In particular, each patient class may be associated with an index corresponding to its expected readmission load, and should a patient arrival necessitate the demand-driven discharge of a current patient, one simply discharges a patient from a class with the highest index of the patients present. It is interesting to note that implementing such a policy requires data about particular patient classes, but does not require the estimation of arrival rates of the various classes. This latter information is highly dynamic and difficult to estimate.

Since the policy we have proposed ignores the effect of future arrivals and the expected length-of-stay of the current occupants it is natural to expect such a policy to be sub-optimal. The following example shows what can go wrong:

Example 1 (Greedy Sub-optimality) Consider the case with $B = 2$ beds and a time horizon of $T = 2$. There are 2 patient types, $i \in \{1, 2\}$. The parameters for each patient type are as follows for some small $\epsilon > 0$:

$$\begin{aligned} \text{For } i = 1: & \quad \mu_1^0 = 1/2, \quad p_1 = p, \quad \mu_1^b = p \\ \text{For } i = 2: & \quad \mu_2^0 = 1, \quad p_2 = p, \quad \mu_2^b = \frac{p}{1-\epsilon} \end{aligned}$$

Therefore, patient type 1 has nominal expected length-of-stay of 2 and readmission load of 1. Similarly, patient type 2 has nominal expected length-of-stay of 1 and readmission load of $1 - \epsilon$.

Consider an initial state at $t = 0$ such that there exists 2 ICU patients: one of each type. Hence, $x_{1,0} = 1$ and $x_{2,0} = 1$. Also, a new patient of type 1 arrives at $t = 0$ and $t = 1$, i.e. $y_{1,0} = y_{1,1} = 1$ while $y_{2,0} = y_{2,1} = 0$.

At $t = 0$, there are already 2 patients in the ICU, and a new patient arrives. Therefore, a current patient must be discharged in order to accommodate the new patient. The greedy policy discharges patient type 2 at $t = 0$ because its readmission load is less than that of patient type 1. This comes at a cost of $1 - \epsilon$. Now, with this demand-driven discharge and the admission of the new patient there are 2 type 1 patients occupying the ICU. With probability $1/4$ neither type 1 patient completes service and departs by $t = 1$ and with the second new arrival, a patient must be discharged to accommodate this new arrival at a cost of 1. With probability $3/4$ at least one of the type 1 patients completes service prior to the second new arrival and no demand-driven discharge is required at $t = 1$. Hence, the expected cost of the greedy policy is $1 - \epsilon + 1/4 = 5/4 - \epsilon$.

On the other hand, the optimal policy recognizes that patient type 2 has a very short length-of-stay and decides not to discharge this patient at $t = 0$. Instead the optimal policy discharges patient type 1 to accommodate the new patient, incurring a cost of 1. Now with this demand-driven discharge and the admission of the new patient, there is one type 1 patient and one type 2 patient occupying the ICU. At the end of time slot $t = 0$, the type 2 patient completes service and departs naturally with probability 1. Regardless of whether the type 1 patient departs naturally, when the second new arrival comes at $t = 1$, it can immediately be accommodated without requiring a demand-driven discharge of a current patient. Hence, the expected cost of the optimal policy is 1.

Taking $\epsilon \rightarrow 0$ we see that $J^*(s_0) \leq \frac{4}{5}J^g(s_0)$ here.

In light of the sub-optimality of our proposed greedy policy, the remainder of this section is devoted to establishing performance guarantees for this policy. In particular, we identify a setting where the greedy policy is, in fact, optimal. More generally we establish that the greedy policy incurs expected readmission costs that are at most a factor of $(\hat{\rho} + 1)$ times the expected costs incurred by an optimal policy, i.e. the greedy policy is a ‘ $\hat{\rho} + 1$ -approximation’. Here $\hat{\rho} = \lambda/\mu_{\min}^0$ ($\mu_{\min}^0 \equiv \min_m \mu_m^0$) is a measure of the utilization of the ICU. This latter bound is independent of all other system parameters.

3.1. Greedy Optimality

In this section, we consider a special case of the general model presented in Section 2 for which a greedy discharge rule is optimal. In particular we have the following theorem:

Theorem 1 (*Greedy Optimality*) Assume that for any two patient classes i, j , if $\frac{p_i}{\mu_i^b} \leq \frac{p_j}{\mu_j^b}$, then $1/\mu_i^0 \geq 1/\mu_j^0$. Then, we have that the greedy policy is optimal, i.e.

$$J^g(s) = J^*(s), \forall s \in \mathcal{S}$$

PROOF: We will without loss consider states s at which all feasible actions require the demand-driven discharge of a current patient (who has not yet completed treatment); i.e. $\sum_m x(s)_m = B$ and $y(s) \neq 0$. For the sake of a contradiction, we will assume that under any optimal policy π^* , $\pi^*(s) \notin \arg \min_{m: x(s)_m > 0} \frac{p_m}{\mu_m^b}$, i.e. the patient selected for the demand-driven discharge under any optimal policy is not among the set of patient types that minimizes one-period costs at state s . For notational convenience, we take $\pi^*(s) = j$, and $i = \pi^g(s) \in \arg \min_{m: x(s)_m > 0} \frac{p_m}{\mu_m^b}$. Thus, by assumption we have that

$$J^*(s) = C(s, j) + E[J^*(S(s, j))] < C(s, i) + E[J^*(S(s, i))]. \quad (3)$$

Now, let $s_i = S(s, j)$, and $s_j = S(s, i)$. We may assume that $x(s_i)_k = x(s_j)_k \forall k \neq i, j$. Moreover, since $C(s, i) < C(s, j)$, we have $1/\mu_i^0 \geq 1/\mu_j^0$ so that we may couple sample paths in the system which used the optimal policy in state s (demand-driven discharged patient j) with the system which used the greedy policy at state s (demand-driven discharged patient i) so that patient i finishes service and departs in the epoch subsequent to $t(s)$ in the former system only if j finishes service and departs naturally in that same epoch in the latter system. Thus, in time slot $t(s) + 1$ we have either that: (i) $x(s_i)_i - x(s_j)_i = 1$ and $x(s_j)_j - x(s_i)_j = 0$, (ii) $x(s_i)_i - x(s_j)_i = 0$ and $x(s_j)_j - x(s_i)_j = 0$ or (iii) $x(s_i)_i - x(s_j)_i = 1$ and $x(s_j)_j - x(s_i)_j = 1$. In case (i), Lemma 1 implies that $J^*(s_i) \geq J^*(s_j)$. In case (ii), we clearly have $J^*(s_i) = J^*(s_j)$ since $s_i = s_j$.

Let us consider case (iii), which says that neither patient i nor j have departed by time slot $t(s) + 1$. We couple the systems starting at states s_i and s_j so that they see identical arrivals and identical service times (departures) for the patients they have in common. Moreover, we couple the service times of the additional type i patient in the s_i system and the additional type j patient in the s_j system as follows: If after any required demand-driven discharges in a particular time step, patient i and j both remain in their respective systems, patient j will complete/depart with probability μ_j^0 . If patient j departs, patient i will depart in the same time step with probability μ_i^0/μ_j^0 ; if patient j does not complete, then neither will patient i . If only one of i or j are present, they will complete with probability μ_i^0 and μ_j^0 respectively.

Now let us consider using the following sub-optimal policy for the system starting at state s_j : we assume that the additional type j patient is in fact a type i patient, and apply the optimal policy for this transformed state. If at some point the type j patient completes service naturally, we choose to register this departure with probability μ_i^0/μ_j^0 , and with the remaining probability assume a ‘virtual’ additional type i patient that will complete service in subsequent periods with probability μ_i^0 . If at some point the discharge policy chooses the additional type j patient (which it regards as a type i patient) for the demand-driven discharge, we charge ourselves $C(s, j)$ (notice that this may occur after the actual patient has already departed and correspond to the demand-driven discharge of the virtual patient), so that the costs incurred here are certainly higher than under an optimal policy for the s_j system. Call this policy π' . We use the optimal policy in the s_i system.

Let \bar{p}_i be the probability that the additional type i patient will have to be demand-driven discharged in the s_i system. Now we have that $J^*(s_i) = \bar{C} + \bar{p}_i C(s, i)$ where \bar{C} is the total readmission costs incurred for patients excluding the additional type i patient. Notice that under our coupling,

$J^{\pi'}(s_j) = \bar{C} + \bar{p}_i C(s, j) = J^*(s_i) + \bar{p}_i [C(s, j) - C(s, i)]$. Consequently, we have that $J^*(s_j) - J^*(s_i) \leq \bar{p}_i (C(s, j) - C(s, i))$.

Cases (i), (ii), and (iii) together yield $E[J^*(S(s, i)) - J^*(S(s, j))] \leq C(s, j) - C(s, i)$ which contradicts (3) (since $C(s, i) \neq C(s, j)$) and yields our result. \square

The above theorem considers problems for which patients with lower readmission loads also have lower nominal lengths-of-stay. In this case, since eliminating a low readmission load patient also frees up capacity that would have otherwise been occupied for a relatively longer time, it is intuitive to expect the greedy policy to be optimal. However, the assumptions of the theorem are likely to be restrictive in practice. In the next section, we consider the performance of the greedy policy without any assumptions on problem primitives.

3.2. A General Performance Guarantee

Our objective in this section is to demonstrate that the greedy heuristic incurs expected costs that are within $\hat{\rho} + 1$ times that incurred by an optimal policy. In particular, we will show that for any state $s \in \mathcal{S}$, $J^g(s) \leq (\hat{\rho} + 1)J^*(s)$, where $\hat{\rho} = \lambda/\mu_{\min}^0$ is a utilization ratio which measures the amount of strain on the ICU: a higher $\hat{\rho}$ implies a more stressed ICU while a lower value implies more able bed resources.

To show the desired bound, we begin with a few preliminary results for the optimal value function J^* . The first result is a natural monotonicity result which says that having an ICU with higher occupancy levels is less desirable than having lower occupancy levels. In particular:

Lemma 1 (*Value Function Monotonicity*) *For all states $s, s' \in \mathcal{S}$ satisfying $x(s) \geq x(s'), y(s) = y(s'), t(s) = t(s')$,*

$$J^*(s) \geq J^*(s').$$

PROOF: Consider a coupling of the systems starting at state s and s' wherein both systems witness identical sample paths for patient arrivals and identical service times for the patients they have in common. More precisely, assuming that at time t , the systems are in states s_t and s'_t respectively, the patients who arrive in both systems are coupled so that $y(s_t) = y(s'_t)$. Let $z(s_t)$ and $z(s'_t)$ be the patient vectors in both systems after these arrivals and any potential demand-driven discharges. Then the number of service completions in both systems over the remainder of the t th epoch are coupled as follows: If $z(s_t)_m \geq z(s'_t)_m$, then the number of patients of type m that finish service and depart naturally from the s' system is given by the Binomial- $(z(s'_t)_m, \mu_m^0)$ random variable $X'_{t,m}$ while the number of patients of type m that finish service and depart naturally from the s

system is given by $X'_{t,m} + Z_{t,m}$ where $Z_{t,m}$ is a Binomial- $(z(s_t)_m - z(s'_t)_m, \mu_m^0)$ random variable. A symmetric situation must hold if $z(s'_t)_m \geq z(s_t)_m$.

Now assume that the system starting at s uses an optimal policy whereas the system starting at state s' ‘mimics’ the actions of the s system (call this policy $\bar{\pi}$), so that if the s system chooses to demand-driven discharge a patient of a particular type, the s' system will also choose to discharge a patient of that type should such a patient be available, whether or not this demand-driven discharge is called for (i.e. whether or not a new patient has arrived and there are no available beds). In the event that the s' system needs to make a demand-driven patient discharge and the s system either does not need make a demand-driven discharge or else selects to demand-driven discharge a patient of a class not available in the s' system, the s' system discharges a randomly chosen patient from among those available. It is easy to see that $\bar{\pi}$ is an admissible randomized non-anticipatory policy: starting at state s' one adds ‘virtual’ patients so that the total number of patients (real and virtual) of a given type in the s' system are identical to the number in the s system. One then employs an optimal policy, and simulates service completion for virtual patients. We now show that under our coupling, $x(s_t) \geq x(s'_t)$ for all t .

The proof is based on induction in time. The base case follows from our assumption that $x(s) \geq x(s')$. We assume that for all $t \leq k$, $x(s_t) \geq x(s'_t)$ and will show this implies the same is true for $t = k + 1$. Let $A_k = \pi^*(s_k)$ and A'_k be the patient discharged at time k under the $\bar{\pi}$ policy. Note that $A'_{k,m} \leq A_{k,m}$ by our definition of $\bar{\pi}$ and the induction hypothesis. We have

$$\begin{aligned} x(s_{k+1})_m - x(s'_{k+1})_m &= [(x(s_k)_m + y(s_k)_m - A_{k,m})^+ - X_{k,m}] - \\ &\quad [(x(s'_k)_m + y(s'_k)_m - A'_{k,m})^+ - X'_{k,m}] \\ &\geq x(s_k)_m - x(s'_k)_m + X'_{k,m} - X_{k,m} \\ &= x(s_k)_m - x(s'_k)_m - Z_{k,m} \\ &\geq 0 \end{aligned}$$

The first inequality comes from our coupling and the definition of the two policies. The second inequality follows from the definition of $Z_{t,m}$; $Z_{t,m} \leq x(s_t)_m - x(s'_t)_m$.

We may thus establish that for all $t(s) \leq t \leq T$, $A_t \geq A'_t$, so that $C(s_t, \pi^*(s_t)) \geq C(s'_t, \bar{\pi}(s'_t))$ for all such t . Taking expectations over the random patient arrivals and departures, we have $J^*(s) \geq J^{\bar{\pi}}(s') \geq J^*(s')$, which is the result. \square

Define the mapping $\hat{S}_k : \mathcal{S} \rightarrow \mathcal{S}$ according to $\hat{S}_k(s) = s'$ with $t(s') = t(s), y(s') = y(s), x(s')_m = x(s)_m$ for all $m \neq k$ and $x(s')_k = (x(s)_k - 1)^+$. In particular, $\hat{S}_k(s)$ differs from the state s only

in that it has potentially one less admitted patient of type k . As we have seen in Lemma 1, it is beneficial to have fewer patients occupying the ICU. The next lemma shows that the benefit of removing a single patient is bounded by the cost of the demand-driven discharge of that patient. Recall that $C(s, k) = \frac{p_k}{\mu_k^b}$ is the expected readmission load of demand-driven discharge of patient type k .

Lemma 2 (*Free Demand-Driven Discharge*) *Given patient type k and $\alpha = \frac{\lambda}{\lambda + \mu_{\min}^0} = \frac{1}{\frac{1}{\rho} + 1}$, for all states s such that $x(s)_k > 0$:*

$$E[J^*(s)] \leq \alpha C(s, k) + E[J^*(\hat{S}_k(s))]$$

PROOF: Let $s^k = \hat{S}_k(s)$ so that $x(s) \geq x(s^k)$. Consider a coupling of the systems starting at state s and s^k wherein both systems witness identical sample paths for patient arrivals and identical service times for the patients they have in common. More precisely, assuming that at time t , the systems are in states s_t and s_t^k respectively, the patients who arrive in both systems are coupled so that $y(s_t) = y(s_t^k)$. Let $z(s_t)$ and $z(s_t^k)$ be the patient vectors in both systems after these arrivals and any potential demand-driven discharges. Then the number of service completions in both systems over the remainder of the t th epoch are coupled as follows: If $z(s_t)_m \geq z(s_t^k)_m$, then the number of patients of type m that finish service and depart naturally in the s^k system is given by the Binomial- $(z(s_t^k)_m, \mu_m^0)$ random variable $X_{t,m}^k$ while the number of patients of type m that finish service and depart naturally in the s system is given by $X_{t,m}^k + Z_{t,m}$ where $Z_{t,m}$ is a Binomial- $(z(s_t)_m - z(s_t^k)_m, \mu_m^0)$ random variable. A symmetric situation must hold if $z(s_t^k)_m \geq z(s_t)_m$.

Given the above coupling, we assume that the system starting at state s^k uses an optimal policy, $\pi^*(\cdot)$, whereas the system starting at state s ‘mimics’ the actions of the s^k system (call this policy $\tilde{\pi}$). In particular, $\tilde{\pi}(s_t) = \pi^*(s_t^k)$ if this is a feasible action at state s_t . Else, if $x(s_t)_k > 0$, $\tilde{\pi}(s_t) = k$; if $x(s_t)_k = 0$, we select $\tilde{\pi}(s_t)$ uniformly at random from $\{m : x(s_t)_m > 0\}$. It is not difficult to see that $\tilde{\pi}$ is an admissible randomized policy (since one may simply simulate the s^k system, while respecting our coupling above).

Now, as in Lemma 1, we may establish that under our coupling, $x(s_t^k) \leq x(s_t)$. Since an optimal policy will not make a demand-driven patient discharge in state s' unless $x(s') + y(s') = B + 1$, it follows that the random time $\tau = \inf\{t > t(s) \mid \sum_m x(s_t)_m + y(s_t)_m = B + 1\}$ corresponds to the first time a patient must be demand-driven discharged to accommodate a new patient in the s system. Moreover, there will have been no demand-driven discharges in either the s^k or s systems for all $t < \tau$ (though there may have been departures).

Call D_k the event that patient k has departed the s system prior to time τ . Given our coupling, if event D_k occurs, then $s_\tau = s_\tau^k$, and $s_t = s_t^k$ for all $t > \tau$. Under the complementary event, D_k^c , s_τ will have one more patient (namely, patient k) than s_τ^k . Thus, we will have $\pi^*(s_\tau^k) = \emptyset$ and $\tilde{\pi}(s_\tau) = k$, and both systems will be in identical states at time $\tau + 1$; i.e. $s_{\tau+1} = s_{\tau+1}^k$. It follows that $s_t = s_t^k$ for all $t > \tau$.

The total cost incurred by the s system under policy $\tilde{\pi}$, $J^{\tilde{\pi}}(s)$, is equal to the cost incurred in time slot τ and the future costs incurred for all $t > \tau$ (since no demand-driven discharges are made prior to τ in either system). Now, from our argument thus far,

$$E[C(s_\tau, \tilde{\pi}(s_\tau)|D_k] = E[C(s_\tau^k, \pi^*(s_\tau^k)|D_k],$$

while

$$E[C(s_\tau, \tilde{\pi}(s_\tau)|D_k^c] = C(s, k) \text{ and } E[C(s_\tau^k, \pi^*(s_\tau^k)|D_k^c] = 0.$$

Moreover, since the event D_k contains the event that k departs prior to the next arrival (which has probability $\frac{\mu_k^0}{\lambda + \mu_k^0}$), we know that $P(D_k^c) \leq \frac{\lambda}{\lambda + \mu_k^0}$. We consequently have:

$$\begin{aligned} E[C(s_\tau, \tilde{\pi}(s_\tau))] &= P(D_k)E[C(s_\tau, \tilde{\pi}(s_\tau)|D_k] + P(D_k^c)E[C(s_\tau, \tilde{\pi}(s_\tau)|D_k^c] \\ &= P(D_k)E[C(s_\tau^k, \pi^*(s_\tau^k)|D_k] + P(D_k^c)C(s, k) \\ &\leq P(D_k)E[C(s_\tau^k, \pi^*(s_\tau^k)|D_k] + \frac{\lambda}{\lambda + \mu_k^0}C(s, k) \end{aligned} \quad (4)$$

It follows that

$$\begin{aligned} J^*(s) &\leq J^{\tilde{\pi}}(s) \\ &= E[C(s_\tau, \tilde{\pi}(s_\tau)) + J^{\tilde{\pi}}(s_{\tau+1})] \\ &= E[C(s_\tau, \tilde{\pi}(s_\tau)) + J^*(s_{\tau+1}^k)] \\ &\leq P(D_k)E[C(s_\tau^k, \pi^*(s_\tau^k)|D_k] + \frac{\lambda}{\lambda + \mu_k^0}C(s, k) + E[J^*(s_{\tau+1}^k)] \\ &= \frac{\lambda}{\lambda + \mu_k^0}C(s, k) + E[J^*(s_\tau^k)] \\ &\leq \frac{\lambda}{\lambda + \mu_{\min}^0}C(s, k) + E[J^*(s_\tau^k)] \\ &= \alpha C(s, k) + E[J^*(s^k)] \end{aligned} \quad (5)$$

The first inequality comes from the optimality of J^* . The first equality comes from the definition of τ . The second equality follows since under our coupling $\tilde{\pi}(s_t) = \pi^*(s_t^k)$ for $t > \tau$. The second inequality comes from (4). The third equality follows from the fact that $E[C(s_\tau^k, \pi^*(s_\tau^k)|D_k^c] = 0$.

□

Now suppose in state s we were able to take the greedy action for ‘free’, i.e. without incurring any readmission costs for the demand-driven discharge of the patient. Then, the following result shows that this is more beneficial than taking the optimal action and paying the cost for the demand-driven discharge. More precisely,

Lemma 3 (*Greedy-for-Free*) For any state $s \in \mathcal{S}$ and $\alpha = \frac{1}{\frac{1}{\beta} + 1}$,

$$E[J^*(S(s, \pi^g(s)))] \leq \alpha C(s, \pi^*(s)) + E[J^*(S(s, \pi^*(s)))]$$

PROOF: This is a direct result of Lemmas 1 and 2. There are two cases to consider: (i) the greedy and optimal policies demand-driven discharge the same patient type in state s : $\pi^g(s) = \pi^*(s)$ and (ii) the greedy and optimal policies demand-driven discharge different patient types in state s : $\pi^g(s) \neq \pi^*(s)$. Let $s_{t(s)+1}^g = S(s, \pi^g(s))$ be the state in time slot $t(s) + 1$ given the greedy policy is used in state s . Similarly, $s_{t(s)+1}^* = S(s, \pi^*(s))$.

Let’s first consider case (i). By our coupling, if both policies select the same patient to demand-driven discharge, then the state in time slot $t(s) + 1$ for the s^g system is the same as that for the s^* system: $s_{t(s)+1}^g = s_{t(s)+1}^*$. By the non-negativity of costs:

$$\begin{aligned} E[J^*(s_{t(s)+1}^g)] &= E[J^*(s_{t(s)+1}^*)] \\ &\leq \alpha C(s, \pi^*(s)) + E[J^*(s_{t(s)+1}^*)] \end{aligned} \quad (6)$$

Now, let’s consider case (ii). Since the greedy and optimal policies do not coincide in state s , the states in time slot $t(s) + 1$ for the s^g and s^* systems do not coincide: $s_{t(s)+1}^g \neq s_{t(s)+1}^*$. We divide this into 2 subcases: a) there are no type $\pi^*(s)$ patients in state $s_{t(s)+1}^g$, i.e. $x(s_{t(s)+1}^g)_{\pi^*(s)} = 0$ (this would occur if in state s there is only one type $\pi^*(s)$ patient and it departs naturally at the end of time slot $t(s)$) and b) there is at least one type $\pi^*(s)$ patient in state $s_{t(s)+1}^g$, i.e. $x(s_{t(s)+1}^g)_{\pi^*(s)} > 0$. Call the corresponding events D_* and D_*^c respectively.

In case a) we have,

$$\begin{aligned} E[J^*(s_{t(s)+1}^g)|D_*] &= E[J^*(\hat{S}_{\pi^*(s)}(s_{t(s)+1}^g))|D_*] \\ &\leq E[J^*(s_{t(s)+1}^*)|D_*] \\ &\leq \alpha C(s, \pi^*(s)) + E[J^*(s_{t(s)+1}^*)|D_*] \end{aligned} \quad (7)$$

The first equality comes from the definition of the \hat{S} mapping and from the fact that there are no $\pi^*(s)$ type patients in state $s_{t(s)+1}^g$ so that $s_{t(s)+1}^g = \hat{S}_{\pi^*(s)}(s_{t(s)+1}^g)$. In other words, state $s_{t(s)+1}^g$ is no different than if patient $\pi^*(s)$ were *also* demand-driven discharged in time slot $t(s)$, despite

only needing to discharge one patient to accommodate the new one. The first inequality comes from Lemma 1, by putting the demand-driven discharged $\pi^g(s)$ patient back into the ICU to get state $s_{t(s)+1}^*$ from state $\hat{S}_{\pi^*(s)}(s_{t(s)+1}^g)$. It is permissible to do this because under case a) at least one patient (of type $\pi^*(s)$) has naturally departed by time slot $t(s) + 1$, leaving an available bed in state $\hat{S}_{\pi^*(s)}(s_{t(s)+1}^g)$. The third inequality comes from the non-negativity of costs.

In case b) we have,

$$\begin{aligned} E[J^*(s_{t(s)+1}^g)|D_*^c] &\leq \alpha E[C(s_{t(s)+1}^g, \pi^*(s))|D_*^c] + E[J^*(\hat{S}_{\pi^*(s)}(s_{t(s)+1}^g)|D_*^c] \\ &= \alpha C(s, \pi^*(s)) + E[J^*(\hat{S}_{\pi^*(s)}(s_{t(s)+1}^g)|D_*^c] \\ &\leq \alpha C(s, \pi^*(s)) + E[J^*(s_{t(s)+1}^*)|D_*^c] \end{aligned} \quad (8)$$

The first inequality comes from applying Lemma 2 to all possible $s_{t(s)+1}^g$ with $x(s_{t(s)+1}^g)_{\pi^*(s)} > 0$ and taking $k = \pi^*(s)$. The second inequality comes from Lemma 1 by putting the demand-driven discharged $\pi^g(s)$ patient back into the ICU. It is permissible to do this because two patient types ($\pi^g(s)$ and $\pi^*(s)$) have been demand-driven discharged by time slot $t(s) + 1$, leaving an available bed in state $\hat{S}_{\pi^*(s)}(s_{t(s)+1}^g)$. Equations (6), (7), (8) together yield the result. \square

In words, Lemma 3 considers that in state s , we are given two options. The first is to take the greedy action without incurring any immediate costs. The second is to take the optimal action while incurring some of the costs. Lemma 3 states that the first option will result in lower costs and enables us to bound the cost savings in taking the optimal action as opposed to the greedy action. We are now in position to show our main result. Namely, the greedy heuristic is guaranteed to be within a factor of $\hat{\rho} + 1$ of optimal, where $\hat{\rho} = \frac{\lambda}{\mu_{\min}^0}$ is the utilization ratio of the ICU.

Theorem 2 For all $s \in \mathcal{S}$, $J^g(s) \leq (\hat{\rho} + 1)J^*(s)$.

PROOF: The proof proceeds by induction on the number of time steps that remain in the horizon, $T - t(s)$. The claim is trivially true if $t(s) = T - 1$ since both the myopic and optimal policies coincide in this case. Consider a state s with $t(s) < T - 1$ and assume the claim true for all states s' with $t(s') > t(s)$.

Now if $\pi^*(s) = \pi^g(s)$ then the next states encountered in both systems are identically distributed so that the induction hypothesis immediately yields the result for state s . Consider the case where $\pi^*(s) \neq \pi^g(s)$. Defining $\alpha = \frac{1}{\frac{1}{\hat{\rho}} + 1}$, we have:

$$\begin{aligned} J^*(s) &= C(s, \pi^*(s)) + E[J^*(S(s, \pi^*(s)))] \\ &\geq (1 - \alpha)C(s, \pi^*(s)) + E[J^*(S(s, \pi^g(s)))] \end{aligned}$$

$$\begin{aligned}
&\geq (1 - \alpha)C(s, \pi^g(s)) + E[J^*(S(s, \pi^g(s)))] \\
&\geq (1 - \alpha)C(s, \pi^g(s)) + E[(1 - \alpha)J^g(S(s, \pi^g(s)))] \\
&= (1 - \alpha)J^g(s) \\
&= \frac{1}{\hat{\rho} + 1}J^g(s)
\end{aligned} \tag{9}$$

The first equality comes from the definition of the optimal policy. The first inequality comes from Lemma 3. The second inequality comes from the definition of the greedy policy which minimizes single period costs. The third inequality comes from the induction hypothesis. The second equality comes from the definition of the greedy value function. This concludes the proof. \square

Our guarantee on performance loss suggests that in regimes where ICU utilization is low, the greedy policy is guaranteed to be close to optimal. At some level, this is an intuitive result—low levels of utilization should imply infrequent demand-driven discharges as there are likely to be available beds when new patients arrive; Theorem 2 makes this intuition precise by demonstrating a bound on how performance loss scales with utilization levels. Our guarantees are worst case; in subsequent sections we will consider a generative family of problems for which the performance loss is a lot smaller than predicted, even at high utilization levels. Moreover, we will demonstrate via an empirical study using patient flow data, that the greedy policy is superior to a number of benchmarks that resemble current practice.

3.3. Patient Evolution during ICU stay

Thus far, we have assumed the distribution for the length-of-stay of each patient is memoryless. Since the health of a patient will vary over the course of his stay, one may wish to employ a length-of-stay distribution that does not have a constant hazard rate. We now consider how to incorporate this more realistic scenario.

For each patient class m , consider a random progression of the state of their health condition. Let $h^m \in \{h_0^m, h_1^m, \dots, h_{n_m}^m\}$ denote the set of health condition states patient class m can achieve. Whenever a new patient of type m arrives, it begins with a health state of h_0^m . Assuming that a patient is in health state h_n^m in some epoch, the patient departs with probability $\mu_m^0(h_n^m)$. If he does not depart, he evolves to health state h_{n+1}^m with probability γ_n^m and remains in state h_n^m with probability $1 - \gamma_n^m$. The different health condition states and corresponding departure probabilities enable us to capture the changes (improvement or deterioration) in patient health as a patient spends time in the ICU. Note that there are no constraints on the relationship between the $\mu_m^0(h_n^m)$ so that the patient does not necessarily improve with time. Indeed, there have been studies which shows that patients likelihood of departure *decreases* the longer they have spent in the hospital Chalfin (2005).

The state space now needs to be expanded to incorporate the different health states each patient class can achieve. To do this, we can redefine $x(s)$ to be a 2-dimensional array where $x_{m,h}(s)$ denotes the number of class m patients in health condition state h . All the results in this section can be recovered in this extended model after having defined ICU utilization as

$$\hat{\rho} = \frac{\lambda}{\min_{m,h} \mu_m^0(h)}$$

The proofs of the results in this section for this extended model are essentially identical to what has already been shown and, consequently, are not replicated. The only difference is in the proof of Lemma 2 where the probability of event D_k^c must be modified to account for the different probabilities of departure over the course of patient evolution.

3.4. A Numerical Study of the Performance Gap

This section is devoted to a brief numerical study of the performance loss incurred by the greedy policy. We compare greedy and optimal performance over (smaller) problems whose size permits the computation of the optimal policy. In a subsequent section, we examine larger problems calibrated to empirical data and compare the performance of the greedy policy to a number of benchmark policies.

In order to enable computation of the optimal policy, we consider a ‘small’ scenario with $B = 10$ beds, $M = 2$ patient types and a time horizon of 240 time slots (assuming admission and discharge decisions are made every 6 minutes, or 10 times an hour, this corresponds to a time horizon of 24 hours). For each data point, we fix the probability of arrival of each patient type. We consider 100 different realizations for the nominal length-of-stay, the readmission probability and readmission length-of-stay of each patient type which we vary uniformly at random. For each fixed set of parameters— $a_{i,t}$, μ_i^0 , p_i , and μ_i^b —we calculate the optimal policy using dynamic programming. We compare the average performance of this optimal policy to the performance of the greedy policy over 100 iterations.

Figure 1 shows the greedy cost relative to the optimal cost ($J^g(s)/J^*(s)$) for a range of different arrival rates. As in Section 2, the probability of a patient arrival is given by λ while the probability an arrival is of patient type 1 is given by a_1 . Values above 1 show the loss in performance due to using the greedy policy. We can see that the greedy policy performs within 3% of optimal, which will typically be superior to what the bound in Section 3.2 suggests. In fact, for reasonable arrival rates ($\lambda < .05$ means 1 patient arrives every 2 hours) the performance loss of the greedy policy is less than 1% of optimal. These results are encouraging, and suggest substantially superior performance than do our worst case bounds. Subsequent sections will be devoted to a performance evaluation for parameters calibrated to actual ICU patient flow data.

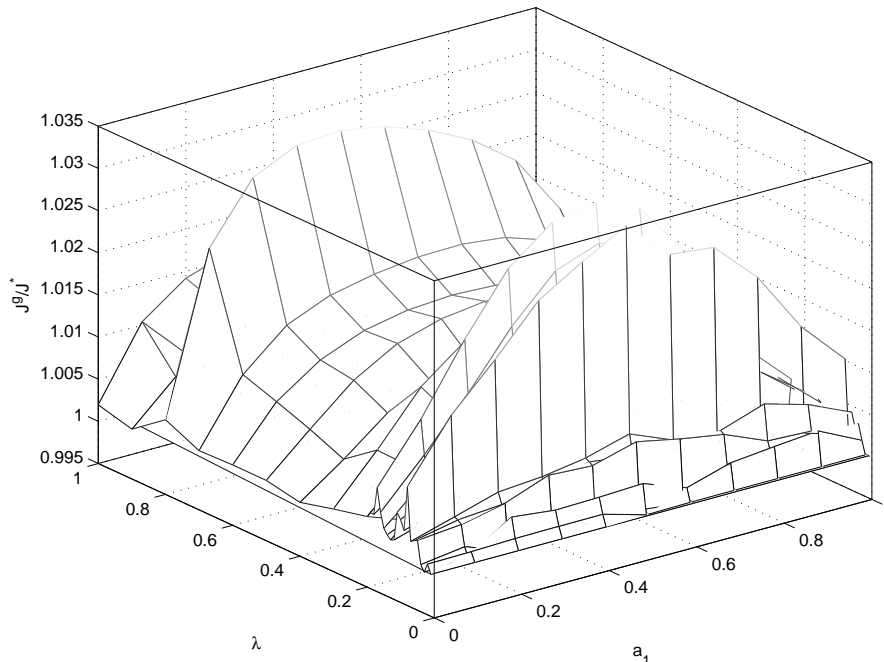


Figure 1 Performance of greedy policy compared to optimal for varying arrival rates.

4. Empirical Data

In this section, we analyze patient data from 7 different private hospitals for a total of 6640 surviving ICU patients over the course of 1 year. Of those patients, 6184 had sufficient data regarding their health indicators to be included in the study. Our goal is to calculate the main patient parameters of our model; namely, the nominal length-of-stay ($1/\mu_m^0$), the readmission probability (p_m), and the readmission length-of-stay ($1/\mu_m^b$).

Patient Classes: Our model requires that we classify patients into M classes based primarily on medical factors relevant to their length-of-stay. Here, we classified patients into 9 different classes based on ‘severity scores’ available in our data set. Of note, the hospital system from which the data are collected has developed a specific methodology for retrospective assignment of severity of illness scores to assess the severity of each patient (see Escobar et al. (2008)). This methodology assigns patients a probability of mortality based on data available immediately prior to admission to the hospital. It does have the important limitation of not providing such a probability for patients admitted to the ICU (e.g., in the situation where a patient was not admitted directly to the ICU). The severity scores are based on a number of different factors including age, primary condition (cardiac, pneumonia, GI bleed, seizure, cancer, etc.), lab results obtained 72 hours prior to hospital admission, chronic ailments (diabetes, kidney failure, etc.) and so on. These scores are used to predict the hospital length-of-stay and mortality rate for each patient. We quantize these severity

scores into one of nine different bins, one bin for each combination of expected length-of-stay (< 3 days, $3 - 4$ days, and > 4 days) and mortality rate ($< 1\%$, $1 - 3.5\%$, and $> 3.5\%$). Because these severity scores require a variety of patient information which is sometimes missing from records, we could not classify 456 patients. We do not use data corresponding to patients who die. This is recommended practice since length-of-stay data for such patients can be misleading; when a patient is unlikely to survive many or no medical interventions can be made to delay eventual death depending on the family's wishes.

ICU Occupancy Levels: Our data set indicates the utilization of the ICU upon patient discharge. This data is central to verifying our hypothesis that ICU occupancy levels upon patient discharge are an important predictor of the likelihood of readmission. We define the 'near capacity' or 'full' state as when the ICU occupancy level is at least 75% of its maximum. If the ICU occupancy is less than 75% of maximum, we say the ICU is in the 'low' state. This characterization is similar to that in Diwas and Terwiesch (2007) and acceptable from a medical perspective.

Sampling Bias: Our study rests on the assumption that the statistics governing a patient's length-of-stay in the ICU, the likelihood of their readmission and the lengths of any subsequent visits depend solely on their health condition as summarized by their severity scores, and whether or not they were discharged from a full ICU. Since we are interested in isolating the impact of demand-driven discharge to accommodate new patients on patient length-of-stay statistics and the likelihood of readmission, it is important to check that the distribution of severity scores for patients in the group of patients discharged from a full ICU is close to that of patients discharged from an ICU in the low state. To this end, we use the Kolmogorov-Smirnov two-sample test (see Smirnov (1939) and related references), which is the continuous version of the chi-squared test. For each pair of ICU occupancy levels (from 1 to 20), we compare the empirical distributions of severity using the Kolmogorov-Smirnov test to see if the samples come from the same distribution. We find that with significance level of 1%, the samples do come from the same distribution. Hence, we conclude with high probability, that the ICU occupancy level parameter and the severity scores of data points in our data set are independently distributed.

To summarize, a data point in our data set can be expressed as a tuple of the form $(S, (L_1, F_1), (L_2, F_2), \dots, (L_k, F_k))$ where S is a severity score, L_i is the patient length-of-stay on his i th visit to the ICU in the episode and F_i is an indicator for whether the ICU was full upon his i th discharge.

4.1. Estimation

Our estimator for μ_m^0 , the nominal length-of-stay for patient type m is simply the empirical average

$$\hat{\mu}_m^0 \triangleq \mu(\text{LOS}_{\text{low}}^0)_m = \sum_i L_1^i \mathbf{1}_{\{F_1^i \neq \text{full}\}} / \sum_i \mathbf{1}_{\{F_1^i \neq \text{full}\}}.$$

Similarly $\sigma(\text{LOS}_{\text{low}}^0)_m$ is an empirical standard deviation. We also calculate the fraction of these patients who return to the ICU during the same hospital stay to calculate a nominal probability of readmission, $P(\text{R}|\text{Low})$. Since a demand-driven discharge is unlikely if beds are available, these readmitted patients relapse for reasons unrelated unrelated to a demand-driven discharge. Thus,

$$P(\text{R}|\text{Low}) = \sum_i \mathbf{1}_{\{F_1^i \neq \text{full}, L_2^i > 0\}} / \sum_i \mathbf{1}_{\{F_1^i \neq \text{full}\}}.$$

Finally, of patients readmitted to the ICU from among those initially discharged from a non-full ICU, we compute an estimate of their expected length-of-stay upon readmission, according to:

$$\mu(\text{LOS}_{\text{low}}^R)_m = \sum_i L_2^i \mathbf{1}_{\{F_1^i \neq \text{full}, F_2^i \neq \text{full}, L_2^i > 0\}} / \sum_i \mathbf{1}_{\{F_1^i \neq \text{full}, F_1^i \neq \text{full}, L_2^i > 0\}}.$$

Notice that $\mu(\text{LOS}_{\text{low}}^R)_m$ is an estimate of patient length-of-stay upon readmission when the readmission is due to medical factors unrelated to demand-driven discharge. Table 1 states the values of the estimates for our data set including information about the relevant number of data points where relevant.

Patient Type	# data points	$\mu(\text{LOS}_{\text{low}}^0)$ (hours)	$\sigma(\text{LOS}_{\text{low}}^0)$	$P(\text{R} \text{Low})$	# data points	$\mu(\text{LOS}_{\text{low}}^R)$ (hours)	$\sigma(\text{LOS}_{\text{low}}^R)$
1	781	37.8	44.7	.070	44	44.2	52.0
2	197	50.2	69.7	.102	16	39.6	42.7
3	39	40.7	32.5	.026	1	44.2	0
4	335	49.5	57.9	.039	8	48.1	45.3
5	425	47.7	49.5	.066	23	59.5	63.3
6	234	54.1	60.4	.077	14	84.9	61.9
7	183	61.5	100.0	.131	17	54.8	56.5
8	355	63.2	71.3	.082	20	92.6	107.9
9	1207	88.3	132.6	.098	89	110.7	191.1

Table 1 Nominal patient parameters: parameters when patients naturally depart and are not discharged in order to accommodate new patients. Average initial length-of-stay ($\text{LOS}_{\text{low}}^0$), readmission probability $P(\text{R}|\text{Low})$ and readmission length-of-stay ($\text{LOS}_{\text{low}}^R$) when discharged from a ‘low’ occupancy ICU. Length-of-stay is given in hours.

We compute similar estimates for patients discharged from a full ICU; we assume these discharges

are demand-driven. Of particular interest is the probability of patient readmission when a patient is discharged from a full ICU, $P(R|Full)$. We estimate this probability according to:

$$P(R|Full) = \sum_i \mathbf{1}_{\{F_1^i=full, L_2^i>0\}} / \sum_i \mathbf{1}_{\{F_1^i=full\}}.$$

We have seen that patients who are not discharged in order to accommodate new patients may require readmission (Table 1); we expect that patients who are discharged from a full ICU may require readmission for those same reasons *in addition* to complications which arise due to being demand-driven discharged. Therefore, we expect the probability of readmission when discharged from a full ICU to be higher than when discharged from a low ICU. We also estimate the expected length-of-stay of such readmitted patients according to

$$\mu(LOS_{full}^R)_m = \sum_i L_2^i \mathbf{1}_{\{F_1^i=full, F_2^i \neq full, L_2^i>0\}} / \sum_i \mathbf{1}_{\{F_1^i=full, F_1^i \neq full, L_2^i>0\}}.$$

Table 2 states the values of these estimates for our data set including information about the relevant number of data points where relevant.

Several remarks are in order. First, we attempted to eliminate outliers from our data set; in total, we removed 3 data points whose lengths of stay were more than 6 standard deviations above their estimated means. Further information on the relevant patients in the data set revealed that their circumstances were indeed abnormal. Second, we note that our estimates suggest that the various lengths of stay random variables have coefficients of variation close to 1 so that our assumption that these quantities are geometrically distributed is potentially a reasonable first order approximation.

Patient Type	# data points	$\mu(LOS_{full}^0)$ (hours)	$\sigma(LOS_{full}^0)$	$P(R Full)$	# data points	$\mu(LOS_{full}^R)$ (hours)	$\sigma(LOS_{full}^R)$
1	422	38.1	37.3	.083	14	43.5	39.9
2	120	50.1	59.9	.125	5	43.6	52.0
3	26	42.7	28.0	.077	0	–	–
4	206	56.3	67.7	.044	2	108.3	42.1
5	300	55.7	83.1	.083	9	126.7	103.2
6	142	47.4	36.2	.077	2	58.4	4.9
7	123	57.0	52.3	.163	8	293.7	271.4
8	228	68.2	93.9	.110	10	62.5	38.8
9	856	82.1	117.0	.112	31	237.9	522.9

Table 2 Discharged patient parameters: parameters when patients are discharged in order to accommodate new patients. Average initial length-of-stay (LOS_{full}^0), readmission probability $P(R|Full)$ and readmission length-of-stay (LOS_{full}^R) when discharged from a ‘full’ ICU. Length-of-stay is given in hours.

4.2. Model Calibration:

Given the estimates from our data set, we next set out to calibrate our model. In particular, we would like to estimate for each class m , the probability of readmission due to complications arising from a demand-driven discharge rather than a natural departure, p_m , as well as the expected length-of-stay of a patient readmitted due to complications arising from the demand-driven discharge.

We take as our estimate of p_m , the quantity $P(R|Full) - P(R|Low)$. This is justified by the underlying assumption that for any severity score (i.e. health condition) S , a given patient with that score and discharged from a full ICU will be more likely to require readmission than if discharged otherwise. In addition our observation that severity scores and ICU occupancy levels upon discharge/departure were (empirically) independent in our data set guarantee that the distributions over S for the data points we use to estimate $P(R|Full)$ and $P(R|Low)$ respectively, are consistent. Now, the demand-driven discharged patients that require readmission may have required readmission due to factors unrelated to the demand-driven discharge or due to complications arising from the demand-driven discharge. Since the fraction of patients in the former category is simply $P(R|Low)$, the remaining fraction of patients, $P(R|High) - P(R|Low)$ belong to the later category. This implies the following structural relationship

$$\frac{P(R|Low)}{P(R|Full)} \mu(LOS_{low}^R)_m + \frac{P(R|Full) - P(R|Low)}{P(R|Full)} \frac{1}{\mu_m^b} = \mu(LOS_{full}^R)_m$$

which may be used to estimate μ_m^b . Alternatively, this relationship allows us to estimate $\frac{p_m}{\mu_m^b}$ directly as $P(R|Full)\mu(LOS_{full}^R)_m - P(R|Low)\mu(LOS_{low}^R)_m$.

These estimates are summarized in Table 3. Only patient types with more than 2 data points for each parameter are considered, which eliminates 3 patient types: 3, 4, and 6.

Patient Type	Nominal LOS ⁰ (1/ μ_m^0)	p_m	$\frac{p_m}{\mu_m^b}$ (hours)
1	37.8	0.013	.52
2	50.2	0.023	1.41
5	47.7	0.017	6.59
7	61.5	0.032	40.69
8	63.2	0.028	-.71
9	88.3	0.014	15.79

Table 3 Estimated Model

Notice in Table 3 that patient type 8 yields a negative estimate of $\frac{p_m}{\mu_m^b}$ which we attribute to statistical errors in our estimates due to small sample sizes and potential model error; it may be the

case that for class 8 patients, our variates are explained by more than simply severity scores and ICU occupancy upon release. As a result, we do not include this patient type in our simulations and consider a total of 5 patient types.

5. Performance Evaluation

Our goal is to evaluate the performance of the greedy discharge policy relative to several relevant benchmarks. To this end, we consider the model described in Section 2 calibrated to the parameters extracted from our data set in the previous section (see Table 3). In the medical community, the decision over which patient to demand-driven discharge is made by assessing which patient is the ‘least critical’ (see, for instance, Swenson (1992)) but what determines criticality is left open to interpretation. Each of the discharge policies studied below can be interpreted in that vain:

- **Probability of readmission (p_m) index:** Under this policy, one selects a patient from that class with the smallest probability of readmission, p_m , of the patients currently in the ICU. For our data set, this translates to the order 1, 9, 5, 2, 7. Readmitted patients tend to be more critical (see Durbin and Kopel (1993)), so that the rationale here is that a lower likelihood of readmission translates to lower patient criticality.

- **Length-of-stay (LOS) index:** Under this policy, one selects a patient from that class with the smallest nominal length-of-stay, $(1/\mu_m^0)$, of the patients currently in the ICU. For our data set, this translates to the order 1, 5, 2, 7, 9. This policy thus equates criticality with the nominal length-of-stay of a patient. This policy is analyzed in Dobson et al. (2009) albeit for a model that is agnostic to readmission loads.

- **The Greedy index:** This is the policy that has been the focus of our study and analysis thus far. The policy prioritizes patients in increasing order of readmission load p_m/μ_m^b which, for the present study, translates to the order 1, 2, 5, 9, 7.

Because physicians currently do not utilize models which model patient dynamics, these policies are based on information which may be available to physicians in the future. Hence, The LOS and p_m index policies serve to resemble the best-effort policies used in current practice. In addition to the index rules above, we also consider a random discharge policy.

We consider a time horizon of 1 week where admission and discharge decisions are made every 6 minutes, or 10 times within an hour and consider an ICU with $B = 10$ beds. Discharge policy simulations are over 100 sample paths each. We use the parameters estimated in Table 3 for nominal length-of-stay, μ_m^0 , probability of readmission, p_m , and expected readmission load, p_m/μ_m^b . We vary the probability of an arrival, λ between 0.01 and 0.1 (i.e. between 1 and 10 arrivals on average

every 10 hours). It remains to specify the traffic mix across patient classes. Below we consider detailed simulation results for three separate compositions of patient traffic.

Uniform Traffic Mix: Our first set of experiments assume that the arrival probability of each patient type is identical so that given an arrival, the patient type is uniformly distributed across the 5 patient types. Figure 2 shows the expected increased readmission load in hours for the four discharge policies in this scenario. We can see that the greedy policy outperforms each benchmark by nearly 10% in some instances. The next best policy is the index policy based on the probability of readmission, i.e. the p_m index policy. However, because there are some patient types with low probability of readmission, but high readmission load (see for instance patient type 9), this policy still results in higher readmission load. As expected, the random policy performs very poorly. What is interesting is that the index policy based on the nominal length-of-stay does not do much better. Upon inspection, we can see from Table 3 that the nominal length-of-stay does little to predict the expected increase in readmission load due to demand-driven discharge which can partially explain the poor performance of this benchmark. Thus, although the problem of minimizing readmission load (as a proxy for maximizing throughput) due to required demand-driven discharges is a hard one, the greedy index appears to substantially outperform the benchmarks studied here. As the arrival rate increases more patients will need to be demand-driven discharged, consequently increasing the expected readmission load. We can see that the performance degradation is more gradual under the p_m -index policy and the greedy index policy than the random and LOS index policies.

In order to appreciate the physical meaning of the costs estimated in these experiments, we note that with 24 hours in a day, an additional cost of $24 \times 7 = 168$ hours corresponds to the loss of an entire bed for 1 week since it will be occupied by readmitted patients. What we see for the uniform traffic mix is that for an arrival probability of $\lambda = .05$ the greedy policy incurs readmission load that is 87.1 hours lower than the next best policy (the p_m -index policy) which corresponds to the loss of a single ICU bed (in a 10 bed ICU) for half a week. The savings relative to the LOS index and random policies are substantially higher. Interestingly, the total number of demand-driven patient discharges by each policy are all within 5% of each other so that while we are not sacrificing much in terms of the number of patients who must be discharged in order to accommodate new patients, the greedy policy significantly reduces the resultant readmission load.

Increased type 1 traffic: We now increase the arrival rate of patient type 1 which is the patient type which is demand-driven discharged first by all benchmark policies other than the random policy. Given an arrival, the patient is of type 1 with probability .5. If it is not of type

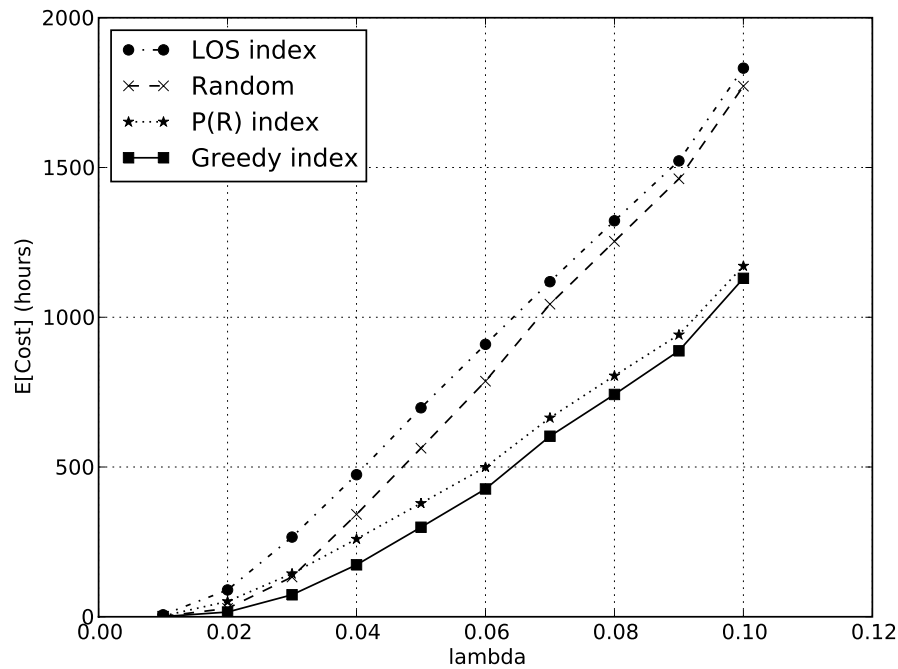


Figure 2 Performance of greedy policy compared to benchmarks for various arrival rates and uniform distribution across patient types.

1, it's type is uniformly distributed across the remain 4 patient types. Because patient type 1 has the shortest nominal length-of-stay, lowest readmission load and probability of readmission, one might expect with a higher density of these patients, the greedy policy, the length-of-stay index policy and the probability of readmission index policy will all have similar performances since they will mostly be discharging the same patient. We can see in Figure 3 that the greedy policy still comfortably outperforms these two policies. The savings relative to the next best policy corresponds to 71.7 hours over one week at a net patient arrival rate of $\lambda = 0.05$ (or 1 bed for 3 days every week in a 10 bed ICU). Additionally, in this scenario the performance of the random policy degrades significantly. This is likely because the random policy keeps missing this 'cheap' patient and demand-driven discharges the expensive ones while the other policies do not.

Increased type 9 traffic: Finally we increase the arrival rate of patient type 9. Given an arrival, the patient is of type 9 with probability .5. If it is not of type 9, it's type is uniformly distributed across the remain 4 patient types. We can see in Figure 4 that the greedy policy continues to outperform our benchmark policies. The savings relative to the next best policy correspond to 105.8 hours over one week at a net patient arrival rate of $\lambda = 0.05$ (or 1 ICU bed out of 10 for 4.5 days over one week).

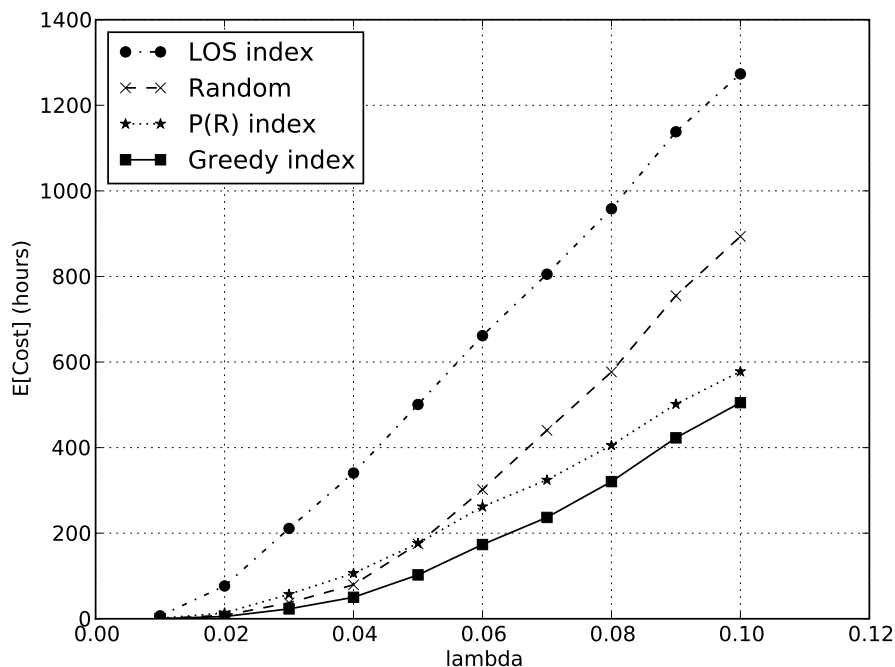


Figure 3 Performance of greedy policy compared to benchmarks for various arrival rates. With probability .5 an arrival is a type 1 patient, and with probability .5 the arrival is uniformly distributed across the remaining patient types.

5.1. A Remark on the Sensitivity of the Greedy index to Problem Data:

Since the parameters that determine the readmission policy we use must be estimated from patient data which, as we have seen here, can be sparse and potentially prone to corruption for various reasons, it is of particular interest to understand whether our proposed policy is highly sensitive to the accurate estimation of these parameters as far as the present empirical study is concerned. To this end we perturb the p_m and μ_m^b parameters until we induce a change in the resulting greedy index. We observe that most parameters would need to be changed by at least 50% to induce a change in the resulting greedy index policy. For the data set under consideration, *no* perturbation of parameters under 20% of their nominal values would induce a change in the resulting index policy. This is comforting, as it suggests that the greedy index policy is relatively robust to errors in estimation of problem parameters.

The empirical study we have presented in the preceding two sections offer a sense of the impact the greedy index policy we have proposed can have on ICU throughput. In particular, we have shown the following:

1. The readmission load phenomenon that the present work seeks to exploit is certainly reflected

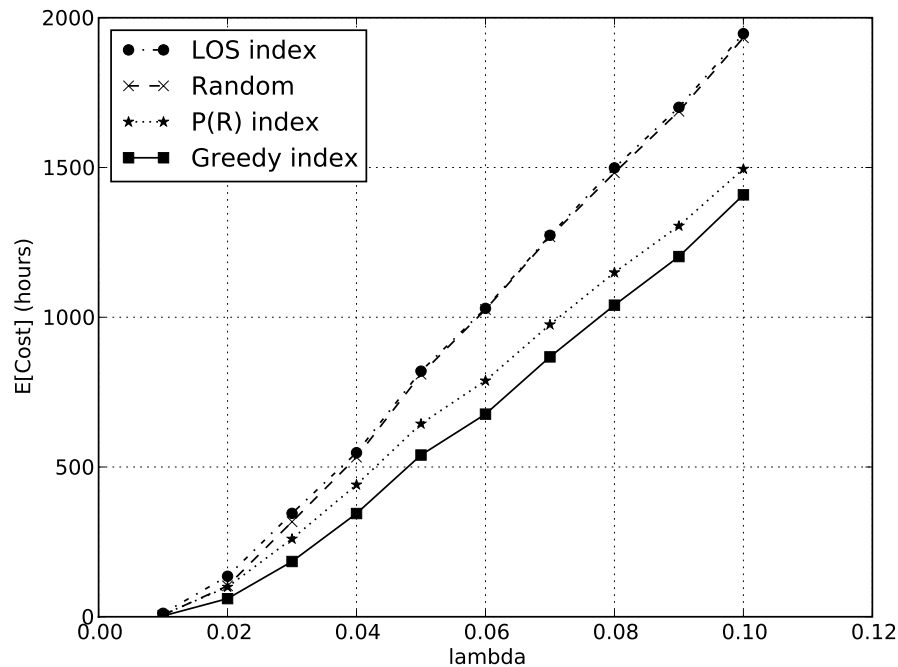


Figure 4 Performance of greedy policy compared to benchmarks for various arrival rates. With probability .5 an arrival is a type 9 patient, and with probability .5 the arrival is uniformly distributed across the remaining patient types.

in the empirical data set considered.

2. Discharge policies (such as the greedy index policy proposed here) that acknowledge this phenomenon in making demand-driven discharge decisions can have a substantive impact of ICU throughput. These gains are on the order of a 10% increase in throughput relative to benchmark schemes for the data set under consideration.

3. The greedy index policy is robust to errors in estimation and its implementation necessitates only modest data collection requirements. In particular, the parameters p_m and μ_m^b can be estimated from parameters available in existing historical data.

6. Conclusion

Faced with the need to accommodate an acute newly admitted patient, a clinician may select from among patients currently in the ICU, a relatively ‘stable’ patient for transfer to a less richly staffed hospital unit. A patient so discharged from the ICU faces risks of physiological deterioration that may ultimately require readmission to the ICU. This is, of course, not an ideal situation either from an efficiency standpoint or the standpoint of ideal patient outcomes. The present work studied the *feasibility* of developing a decision support tool to aid clinicians in these difficult decisions. We

have attempted to gauge the *value* of such a support tool using a large patient flow data set and quantified this value in terms of potential throughput gains.

The model we have developed revolves around simple estimates of the likelihood of a demand-driven patient discharge eventually resulting in a readmission. We estimated our model from actual patient-flow data. Given our model, we developed a simple index based policy to serve as a decision support tool to a physician making the aforementioned discharge decisions. Our support tool is, by its structure, easy to implement from a clinical standpoint, and highly robust to estimation errors. The latter point is well reflected in our empirical study. Our study suggests that implementation of our support tool could result in throughput gains of close to 10% even under modest assumptions on patient traffic, at least in the context of the hospital system from which we collected the data for the study.

This work suggests several future potential research directions, including:

1. Our predictive model for the likelihood of readmission is, from a clinical standpoint, relatively simple. For example, it does not include diagnostic or physiologic data available at the time that a patient was discharged. A richer data set will permit a relatively straightforward extension of our current work; a higher fidelity predictive model will yield a clearer picture of the impact of our discharge policies on throughput and potentially bring to light new phenomena.

2. Developing more complex predictive models of patient dynamics that recognize the evolution of patients over the course of their stay. We believe that the present study is sufficient motivation to collect data that would allow us to identify such a model. Such data could be employed to assign patients a “readiness for discharge” severity score similar in concept to other existing severity of illness scores. This is also key to practical deployment of a decision support tool.

3. Theoretically, we have shown that our index policy is optimal in certain regimes and guaranteed to incur readmission loads of no greater than a factor of $(\hat{\rho} + 1)$ of an optimal policy in general. It would be interesting to understand traffic regimes where this bound could be made tighter – this is, of course, a somewhat secondary pursuit but nonetheless very interesting from a theoretical perspective.

4. It would be interesting to initiate a study of ICU *admissions* so as to move towards a more holistic view of equitable and optimal allocation of hospital resources.

References

- Chalfin, D. B. 2005. Length of intensive care unit stay and patient outcome: The long and short of it all. *Critical Care Medicine* **33** 2119–2120.

- Chalfin, D. B., S. Trzeciak, A. Likourezos, B. M. Baumann, R. P. Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* **35** 1477–1483.
- Chan, C. W., V. F. Farias. 2009. Stochastic depletion problems: Effective myopic policies for a class of dynamic optimization problems. *Mathematics of Operations Research* **34**(2) 333–350.
- Diwas, K.C., C. Terwiesch. 2007. An econometric analysis of patient flows in the cardiac ICU. *Working Paper, University of Pennsylvania, Wharton School of Business*.
- Dobson, G., H.-H. Lee, E. Pinker. 2009. Patient flow in an ICU. *Working Paper, University of Rochester, William E. Simon Graduate School of Business Administration*.
- Durbin, C.G., R.F. Kopel. 1993. A case-control study of patients readmitted to the intensive care unit. *Critical Care Medicine* **21** 1547–1553.
- Escobar, G. J., J. D. Greene, P. Scheirer, M. N. Gardner, D. Draper, P. Kipnis. 2008. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Medical Care* **46** 232–239.
- Green, L. V. 2006. *Queueing Analysis in Healthcare*, chap. Patient Flow: Reducing Delay in Healthcare Delivery. Springer, New York, N.Y.
- Green, L. V., S. Savin, B. Wang. 2003. Managing patient service in a diagnostic medical facility. *Operations Research* **54** 11–25.
- Green, L. V., J. Soares, J. F. Giglio, R. A. Green. 2006. Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13** 61–68.
- Huang, X. A. 1995. A planning model for requirement of emergency beds. *Journal of Mathematics Applied in Medicine Biology* **12** 345–353.
- Kwak, N., C. Lee. 1997. A linear programming model for human resource allocation in a health-care organization. *Journal of Medical Systems* **21** 129–140.
- Murray, M., M. Davies, B. Boushon. 2007. Panel size: how many patients can one doctor manage? *Family Practice Management* **14** 44–51.
- Savin, S. 2006. *Patient Flow: Reducing Delay in Healthcare Delivery, International Series in Operations Research & Management Science*, vol. 91, chap. Managing patient appointments in primary care. Springer, US, 123–150.
- Smirnov, N. 1939. Estimating the deviation between the empirical distribution functions of two independent samples. *Moscow University Mathematics Bulletin* **2** 3–16.
- Snow, N., K.T. Bergin, T.P. Horrigan. 1985. Readmission of patients to the surgical intensive care unit: Patient profiles and possibilities for prevention. *Critical Care Medicine* **13** 961–985.

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

Yankovic, N., L. Green. 2008. A queueing model for nurse staffing. *Working Paper, Columbia University, Columbia Business School*.