Maximizing Throughput of Hospital Intensive Care Units with Patient Readmissions

Carri W. Chan  
Division of Decision, Risk and Operations, Columbia Business School cwchan@columbia.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 version: December 16, 2010  
Second version: May 14, 2010  
First version: October 19, 2009

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 the 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 setting (such as the Transitional Care Unit (TCU) or Medical Surgical Floor (Floor)), 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 the 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, associating a demand-driven discharge with some cost dependent on patient characteristics, our goal is to ‘optimally’ discharge patients so as minimize total expected costs associated with demand driven discharges over time. As an example of a demand-driven discharge related cost, one may consider the increase in expected readmission load associated with the increased likelihood of readmission due to a demand-driven discharge. We will eventually estimate such a cost metric from actual patient data. We consider a stylized model of an actual ICU where the number of ICU beds is fixed. 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. 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. We later consider an extension of our model which includes the additional option of blocking new patients. This discharged patient may subsequently deteriorate and return to the ICU, imposing an
additional load on the ICU beds; a demand-driven discharge might increase the likelihood of this deterioration and as such might contribute to a higher readmission load. Our primary goal will be to minimize the total expected increase in readmission load due to demand driven discharges. We will see that minimizing this objective is closely aligned with an appropriate notion of throughput maximization.

We make the following key contributions:

• We identify a simple ‘myopic’ discharge strategy that corresponds to an index policy: every patient class is associated with a class specific index (There exist a number of proprietary classification systems; patients within a class are relatively homogenous). 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.

• 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 induce a change in the associated indices.

• 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.

• 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 substantially mitigate the increase in readmission load (measured in bed-hours) faced by an ICU due to demand-driven discharges. This decrease can be as much as 30% 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.

As such, this study identifies a discharge procedure that allows us to utilize available ICU resources as effectively as possible. At a high level, our analysis suggests 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 patient outcomes, increased efficiency, and decreased costs.

1.1. Related Literature

The use of critical care is increasing, which is making already limited resources even more scarce (Halpern and Pastores 2010). In fact, it was shown that 90% of ICUs will not have the capacity to provide beds when needed (Green 2003). As such, it is the case that some patients may require premature discharges in order to accommodate new, more critical patients. In a recent econometric study (Kc and Terwiesch 2007), these types of patient discharges were shown to be a legitimate cause of patient readmissions 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.

There has been a significant body of research in the medical literature which has looked at the effects of patient readmissions. In Chrusch et al. (2009), high occupancy levels were shown to increase the rate of readmission and the risk of death. Unfortunately, readmitted patients typically have higher mortality rates and longer hospital lengths-of-stay (see Franklin and Jackson (1983), Chen et al. (1998), Chalfin (2005), Durbin and Kopel (1993) and related works).

When a new patient arrives to the ICU, either after experiencing some trauma or completing surgery, 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 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 (see for instance Bone et al. (1993) and a study by the American Thoracic Society (1997)) 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. (2010) 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 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 5, 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) considers 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 Green and Savin (2008), 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, he may be delayed before the physician is able to see him. 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.

In a related paper on ICU patient flow (Shmueli et al. 2003), the authors examine the affect of ICU admission strategies on the distribution of ICU bed occupancy. The authors assume it is possible for patients to wait for an ICU bed, regardless of their criticality. For the specific ICUs we consider, waiting is highly undesirable (thereby necessitating our modeling decisions that arriving patients be given a bed immediately), an interesting direction for future work would be to consider an intermediate scenario, where some patients may be delayed, whereas others must be given a bed immediately.
Finally, we note that from a technical perspective, the present paper bears a connection to recent work by us (Chan and Farias 2009), in that we develop a performance guarantee based on an analysis of one-step deviations from an optimal policy. That said, the present paper considers a class of models entirely distinct from the ‘depletion problems’ studied in Chan and Farias (2009) and succeeds in establishing relative approximation guarantees for a class of models left unaddressed by that past work. The properties we exploit in our analysis are new and it would be interesting to understand whether the techniques introduced here have application to the more natural cost-minimization variants of the queueing problems introduced in Chan and Farias (2009).

The rest of the paper proceeds as follows. In Section 2 we formally introduce the queueing model and patient dynamics which we study. In Section 3 we consider the performance of an index policy which selects patients to discharge in a greedy manner based on their expected costs in terms of medical outcomes and the burden possible readmissions may inflict upon the capacity-limited ICU. We explore a scenario where the proposed greedy policy (based on an information-rich model) is, in fact, optimal. Furthermore, in a more general setting, we show that 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 utilization. In Section 4 we provide numerical results which show that in practice this gap is likely to be much smaller—on the order of a couple percent. In Section 5 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 6 that the greedy policy outperforms a number of benchmarks of interest. We conclude in Section 7.

2. Model

We begin by proposing a stylized model of the patient flow dynamics in a hospital ICU and account for the fact that discharging a current ICU patient in order to accommodate a new one could result in an increased chance of the patient requiring readmission. This in turn would result in increased consumption of ICU resources down the road. At a high level, our model captures the fact that a newly admitted patient must receive ICU resources and that this requirement in turn could necessitate the discharge of an existing ICU patient. Such a discharged patient may require readmission to the ICU if his condition deteriorates. Since arriving patients cannot be queued or blocked, the model we consider is distinct from a typical queueing model. A natural goal is to find a patient discharge policy that maximizes ICU ‘throughput’ (see Section 2.1 for a rigorous definition). As opposed to doing so directly, we instead consider the simpler to understand and analyze task of minimizing the total expected workload incurred due to patient readmission, and relate the optimization of this objective in a precise way to the goal of throughput maximization.
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 $m \in \mathcal{M} = \{1, 2, \ldots, M\}$ 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 $1/\mu^0_m$. 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 $1/\mu^R_m$. Thus, such a demand-driven discharge of a patient of type $m$ results in an additional expected workload of $p_m/\mu^R_m$ 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 (see Section 3.5). 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 5); moreover 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 $\lambda$. We define $\hat{\rho} = \frac{\lambda}{\min_m \mu^0_m}$ as a measure of the utilization of the ICU: a higher $\hat{\rho}$ implies a more stressed ICU while a lower value implies more able bed resources. Notice that this measure does not rely on the relative arrival intensities of various patient types. We let $a_{t,m}$ denote the probability that a newly arriving patient at time $t$ is of type $m$. These probabilities are deterministic and known a priori to the optimal discharge policy; the policy we study will require neither knowledge of $\lambda$ 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 $x_{t,m} \in \{0, 1, \ldots, B\}$ 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 \( \mu_m^0 \) after any demand-driven discharge and/or admission occur, if required.

**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:

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

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 \( s_t \in S \) 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 \( A(s) \subseteq M \) 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 \( A(s) = \{m : x(s)_m > 0\} \). At all other states \( s \), \( A(s) = \{m : x(s)_m > 0\} \cup \{\emptyset\} \). Thus, an action \( A \in A(s) \) 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, \( A(s) \), 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- \((x(s)_m + y(s)_m - 1\{A=m\}, \mu_m^0)\) 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 \( \{1, 2, \ldots, M\} \cup \{\emptyset\} \); \( R_t = m \) with probability \( \lambda a_{t,m} \) for \( m \in \{1, 2, \ldots, M\} \) and \( R_t = \emptyset \) with the remaining probability. The vector denoting arrivals at the next state, \( Y_{t(s)+1} \) is then given by \( Y_{t(s)+1,m} = 1_{R_{t(s)+1} = m} \). Thus, \( s' = S(s, A) \) is defined as:

\[
x(s')_m = x(s)_m + y(s)_m - 1\{A=m\} - X_{t(s),m}, \\
y(s')_m = Y_{t(s)+1,m}, \\
t(s') = t(s) + 1.
\]

**Cost Function:** The cost incurred for taking action \( A \) is defined by a cost function \( C : S \times A \rightarrow \mathbb{R}_+ \). Such a cost function might capture a number of quality metrics. For instance, the cost function
might reflect the net decrease in quality-adjusted life years (QUALYS) as a result of a demand-driven discharge. Our development in the sequel extends to any such cost function.

Models describing the impact of a demand-driven discharge are unavailable to practitioners today. As such, this work will focus on a cost metric that is estimable from available data. In particular, we take $C(s, A) = \frac{p_A \mu_R}{\mu_A}$ for $A \in \{1, 2, \ldots, M\}$, and $C(s, \emptyset) = 0$. Recall that $p_A$ is the probability of readmission of patient class $A$ and $\mu^R_A$ is its expected service rate upon readmission so that the cost of a demand-driven patient discharge under this metric is the expected workload that patient will impede on the system due to potential readmission. Hence, the cost incurred by action $A$ is the expected readmission load due to demand-driven discharge of patient type $A$.

**Objective:** Let $\Pi$ denote the set of feasible discharge policies, $\pi$ which map the state space $S$ to the set of feasible actions $A$. Define the expected total cost-to-go under policy $\pi$ as:

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

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 $\pi^*$ a corresponding optimal policy, i.e. $\pi^*(s) \in \arg \min_{\pi \in \Pi} J^\pi(s)$.

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

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

for all $s \in S$ 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 $\pi^*$ may be found as the greedy minimizer with respect to $J^*$ in $\Pi$. The minimization takes into consideration the current state $s$, the distribution of future patient arrivals, as well as the impact of the current decision on future states. References to an optimal policy in subsequent sections will refer to precisely this policy. The size of $S$ 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.

In addition to the above objective, one may also consider the task of finding an average-cost optimal policy; i.e. the task of finding a stationary policy $\pi$ (a policy that satisfies $\pi(s) = \pi(s')$ for all $s, s'$ with $x(s) = x(s')$, and $y(s) = y(s')$), that solves

$$\kappa^*(s) = \min_{\pi} \kappa^\pi(s)$$
where $\kappa^\pi(s) = \limsup_{T \to \infty} \frac{1}{T} E \left[ \sum_{t'=t(s)}^{T-1} C(s_{t'}, \pi(s_{t'})) \right] | s_{t(s)} = s$. is the average-cost to go (i.e. the long run costs incurred due to demand-driven discharges) under policy $\pi$.

It is not difficult to see that the Markov chain on $\hat{S}$ (the projection of $S$ on its $x$ and $y$ coordinates) induced under any stationary policy $\pi$ is irreducible, so that in fact, the above problem is solved simultaneously for all $s$ by a common stationary policy $\pi^*$, and $\kappa^\pi(s) = \kappa^\pi$ for all $s \in S$ and a stationary policy $\pi$. Finally, the ergodic theorem for Markov chains implies (with some abuse of notation), that

$$\kappa^\pi = \sum_{s \in \hat{S}} \nu_\pi(s) C(s, \pi(s)),$$

where $\nu_\pi$ is the stationary distribution induced by $\pi$ on $\hat{S}$.

### 2.1. A Connection with Throughput

When costs associated with a demand-driven discharge are taken to be the expected excess load such a patient would bring to the system upon readmission, the objective just discussed is aligned with a notion of throughput maximization. We digress briefly to develop this connection. In particular, preserving the details of the model we have just presented, consider that upon discharge, a patient of a given class $m$ enters a ‘readmission queue’. Patients from the readmission queue can be accommodated in one of $B'$ beds (distinct from the $B$ beds serving first time admissions). Once allocated a bed, a readmitted patient of type $m$ occupies the bed for a geometrically distributed duration with mean length $1/\mu^R_m$ (with probability $p_m$) and requires no time with the remaining probability (i.e. is effectively not readmitted). Depending on the arrival rates of first time admissions and the policy used in selecting such patients when a demand-driven discharge is called for, this readmission queue may or may not be stable. If for a given profile of arrival rates of first-time admissions, there exists a demand-driven discharge policy that renders the readmission queue stable, we will refer to such a profile of arrival rates as ‘admissible’ and refer to the set of all admissible arrival rate profiles as the throughput region. Put simply, profiles of arrival rates for first-time admissions that lie outside the throughput region cannot be sustained without severe compromises to care, irrespective of the discharge policies used.

As it turns out, if in fact the profile of arrival rates for first-time admissions is admissible, the policy minimizing the long run criterion described earlier will guarantee that the readmission queue remains stable. We will demonstrate this fact in Appendix B. Of course, finding such a policy is difficult, and we will eventually settle on heuristics that ‘approximately’ minimize long run costs. For such heuristics, the arrival-rate profiles that can be stabilized are proportionately ‘smaller’. We make this fact precise in Appendix B.
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:

\[
\pi^g(s) \in \arg \min_{m \in A(s)} \frac{p_m}{\mu_m}. \tag{2}
\]

It is easy to see that the policy specified by (2) 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. In Section 3.5 we will comment on a natural analogue to the above policy for general cost metrics on the impact of a demand-driven discharge.

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. In the Appendix, Example A shows what can go wrong.

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’) where \(\hat{\rho} = \frac{\lambda}{\mu_{0}^{\min}}\) (here \(\mu_{0}^{\min} \triangleq \min_m \mu_0^m\)) is a measure of the utilization of the ICU defined in Section 2 a higher \(\hat{\rho}\) implies a more stressed ICU while a lower value implies more able bed resources. 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. The proof of this result can be found in the appendix. In particular we have the following theorem:
Theorem 1. (Greedy Optimality) Assume that for any two patient classes $i, j$, if $\frac{p_i}{\mu^0_i} \leq \frac{p_j}{\mu^0_j}$, then $\frac{1}{\mu^0_i} \geq \frac{1}{\mu^0_j}$. Then, we have that the greedy policy is optimal, i.e.

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

The above theorem considers problems for which patients with lower readmission loads also have higher 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 as discussed in Section 2. In particular, we will show that for any state $s \in \mathcal{S}$, $J^g(s) \leq (\hat{\rho} + 1)J^*(s)$, where $\hat{\rho} = \lambda \mu^0_{\min}$ is a utilization ratio defined in Section 2.

To show the desired bound, we begin with a few preliminary results for the optimal value function $J^*$. The proofs of these results can be found in the appendix. 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 $x(s) \geq x(s'), y(s) = y(s'), t(s) = t(s')$, we have:

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

In words, the above Lemma states that all else being equal, it is advantageous to start at a state with a fewer number of patients occupying the ICU. Now suppose in state $s$ we chose to take the greedy action as opposed to the optimal action (assuming of course that the two are distinct). It must be that the former leads to a higher cost state than does the optimal action. The following result places a bound on this cost increase. In particular, we have:

Lemma 2. (One Step Sub-optimality) For any state $s \in \mathcal{S}$ and $\alpha = \frac{\hat{\rho}}{\hat{\rho} + 1}$,

$$E[J^*(S(s, \pi^g(s)))] \leq \alpha C(s, \pi^*(s)) + E[J^*(S(s, \pi^*(s)))]$$
In words, Lemma 2 tells us that if we were to deviate from the optimal policy for a single epoch (say, in state \( s \)), the impact on long term costs is bounded by the quantity \( \alpha C(s, \pi^*(s)) \). We now use this bound on the cost of a single period deviation in an inductive proof to establish performance loss incurred in using the greedy policy; we show that the greedy heuristic is guaranteed to be within a factor of \( \hat{\rho} + 1 \) of optimal, where \( \hat{\rho} = \frac{\lambda}{\mu_{\min}} \) is the utilization ratio of the ICU defined in Section 2.

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

**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 \( \alpha = \frac{\hat{\rho}}{\hat{\rho} + 1} \), we have:

\[
J^*(s) = C(s, \pi^*(s)) + E[J^*(S(s, \pi^*(s)))] \\
\geq (1 - \alpha)C(s, \pi^*(s)) + E[J^*(S(s, \pi^g(s)))] \\
\geq (1 - \alpha)C(s, \pi^g(s)) + E[J^*(S(s, \pi^g(s)))] \\
\geq (1 - \alpha)C(s, \pi^g(s)) + E[(1 - \alpha)J^g(S(s, \pi^g(s)))] \\
= (1 - \alpha)J^g(s) \\
= \frac{1}{\hat{\rho} + 1} J^g(s) \tag{3}
\]

The first equality comes from the definition of the optimal policy. The first inequality comes from Lemma 2. 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.

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. Before we continue, we briefly discuss extensions to the model presented in Section 2 and how the presented results can be applied.

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 $h^m \in \{h^m_0, h^m_1, \ldots, h^m_{nm}\}$ 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 $h^m_0$. Assuming that a patient is in health state $h^m_n$ in some epoch, the patient departs with probability $\mu^0_m(h^m_n)$. If he does not depart, he evolves to health state $h^m_{n+1}$ with probability $\gamma^m_n$ and remains in state $h^m_n$ with probability $1 - \gamma^m_n$. Should a patient in health state $h^m_n$ be demand-driven discharged, the probability he requires readmission is $p_m(h^m_n)$ and upon readmission his expected length-of-stay is $1/\mu^R_m(h^m_n)$. 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 $\mu^0_m(h^m_n)$ 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 $x_{m,n}(s)$ denotes the number of class $m$ patients in health condition state $h^m_n$. We consider using the natural analogue to the greedy policy discussed thus far:

$$\tilde{\pi}^g(s) \in \arg \min_{(m,n):x_{m,n}(s)>0} \frac{p_m(h^m_n)}{\mu^R_m(h^m_n)}$$

Now, Lemma 2 can be established exactly as before for this new system, with the understanding that we will say $x(s) \geq x(s')$ iff $x_{m,n}(s) \geq x_{m,n}(s')$ for all $m,n$. Further, the analysis used in the proof of Lemma 2 also applies identically as in the case of that result to show that for $\alpha = \frac{\bar{a}}{\rho+1}$,

$$E[J^*(S(s, \tilde{\pi}^g(s)))] \leq \alpha C(s, \pi^*(s)) + E[J^*(S(s, \pi^*(s)))]$$
where we now define
\[
\hat{\rho} = \frac{\lambda}{\min_{m,n} \mu_m^0 (h_m^n)}.
\]
With these results, the proof of Theorem 2 applies verbatim to yield

**Theorem 3.** For all \( s \in S \), \( J^{x_0}(s) \leq (\hat{\rho} + 1)J^*(s) \).

### 3.4. Patient Diversions

Throughout our discussion we have assumed that all new patients must be given a bed immediately. In some cases, high occupancy levels in an ICU can lead to congestion in other areas of the hospitals, such as the Emergency Department (ED), because patients cannot be transferred across hospitals units. In Allon et al. (2009) and McConnell et al. (2005), it is shown that when ICU occupancy levels are high, ambulance diversions increase. Because of the inability to move patients from the ED to ICU, patients are blocked from the ED and ambulances must be diverted to other hospitals. In de Bruin et al. (2007), the authors examine the case of bed allocation given a maximum allowable number of patient diversions in the case of cardiac intensive care units. The authors identify scenarios where achieving the target number of patient diversions is possible, but do not consider how to make admission and discharge decisions. Ambulance diversion comes at a cost—for both the hospital and patient. The hospital loses the revenue generated for treatment (McConnell et al. 2006, Melnick et al. 2004, Merrill and Elixhauser 2005) while delays due to transportation time may result in worse outcomes for the diverted patient (Schull et al. 2004). On the other hand, diversions can sometimes alleviate over-crowding (Scheulen et al. 2001).

Typically, diverted ambulance patients are not the ones who require ICU care (Scheulen et al. 2001). However, within a hospital it may still be possible to block new ICU patients admissions, either by diverting them to another unit (i.e. a Transitional Care Unit or General Floor) within the same hospital or transferring them to an ICU in a different hospital (because of the integrated nature of the hospital system we study, such intra-hospital transfers do occur). Blocking new patients may reduce the number of demand-driven discharges. Note that these new patients are often being transferred from a different hospital unit (Emergency Department, Operation Room, General Ward, etc.) rather than being brought in by ambulances, which is the case of the extensive body of literature on ambulance diversions. Given the ability to divert patients, we consider how to incorporate patient diversions into our model and decision analysis. We extend our model to allow new ICU patients to be diverted to another hospital ICU or unit of lesser care. Hence, when an ICU is full the hospital administrator must decide whether to block the new patient or to make a demand-driven discharge of a current patient in order to admit the new patient.
To formalize the above decision making, we consider the following extension of our model: in a given state $s$, we permit an additional action corresponding to diversion which we denote by $D$; we let $C(s, D)$ denote the cost associated with a diversion in state $s$; as per our discussion above, this cost must capture the increased risks to the patient being diverted in state $s$ (i.e. the arriving patient in that state) as also potential revenue losses to the hospital. We then consider employing the following policy; for states $s \notin \mathcal{S}_{\text{full}}$, i.e. states where the ICU has available capacity, no action is necessary. Otherwise, we follow the following diversion/discharge policy:

$$\hat{\pi}(s) = \begin{cases} \pi^g(s), & \text{if } C(s, D) \geq C(s, \pi^g(s)); \\ D, & \text{otherwise.} \end{cases}$$

Now, Lemma 1 can be established exactly as before for this new system, and the analysis used in the proof of Lemma 2 also applies identically as in the case of that result to show that for $\alpha = \frac{\hat{\rho}}{\hat{\rho} + 1}$,

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

Given these properties, the proof of Theorem 2 applies verbatim to yield

**Theorem 4.** For all $s \in \mathcal{S}$, $J^*(s) \leq (\hat{\rho} + 1)J^*(s)$.

### 3.5. General Cost Metrics

One may argue that the expected excess load upon readmission due to a demand-driven discharge does not entirely capture the impact of such a discharge. For instance, one may worry about the impact of mortality (see Section 6.1), or more generally, the impact of such a discharge on a long-term health indicator such as quality life-years. Unfortunately, as things stand, there are no predictive models available that measure the impact of a demand-driven discharge along any dimension; as far as we know the expected excess readmission load we estimate here is the first such (crude) predictive model of its kind.

As more sophisticated models become available, the heuristic presented here has a natural analogue. In particular, let us assume a cost metric $h : \{1, \ldots, M\} \to \mathbb{R}_+$, that assigns a cost to a demand-driven discharge contingent on the patient type and consider the goal of minimizing expected total costs incurred over some horizon under this metric; of course, $C(s, \emptyset) = 0$. We then consider an index rule of the following type:

$$\pi^g(s) \in \arg \min_{m \in A(s)} h(m).$$

(4)

It is not difficult to see that the performance results of this Section extend *mutatis mutandis* to this new criterion. In particular, the statements of Lemmas 1 and 2 and consequently Theorem 2 hold verbatim; notice that those proofs did not rely on the actual definition of $C(s, A)$ beyond the fact that $C(s, A) \geq C(s, \emptyset)$ for all $A \in \mathcal{A}(s)$.
4. Comparison to Optimal

This section is devoted to examining the performance loss of the greedy policy via numerical studies. We compare the greedy and optimal policies for a set of smaller problems for which the optimal policy is actually computable. In the following section, we examine larger problem instances calibrated to empirical data and compare the performance of the greedy policy to a number of benchmark policies.

In Section 3.2, we have shown that the greedy performance is an \((\hat{\rho} + 1)\)-approximation algorithm to optimal. 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 with mean 25 hours, 2\%, and 125 hours, respectively. For each fixed set of parameters \(a_{i,t}, \mu_i^0, p_i, \) and \(\mu_i^R\), 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 sample paths.

Figure 1 shows the ratio of the greedy performance to the optimal performance \((J^g(s)/J^*(s))\) for a range of different arrival rates. As from Section 2, the probability of a patient arrival is
given by $\lambda$ 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 is substantially 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 differences are so small they can essentially be ignored due to possible numerical errors. The greedy policy does not require arrival rate information and is much simpler to compute than optimal. These simulation results suggest that using the greedy policy results in little performance loss while significantly reducing the computational complexity. In fact, while the complexity of the greedy policy grows linearly in the time horizon, $T$, and logarithmically in the number of patient types ($\log M$), the complexity of the optimal policy grows exponentially in a number of problem parameters despite only resulting in slightly higher performance. The simplicity and good performance of the greedy policy, which simply prioritizes different patient types, makes it desirable for real-world implementation.

5. 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 ($1/\mu^0_m$), the readmission probability ($p_m$), and the readmission length-of-stay ($1/\mu^R_m$).

**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 transferred from an out-of-system hospital since any lab results obtained prior to admission to an in-system hospital is not recorded in the system-wide Electronic Medical Records. 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 factors 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 influence patient discharge. 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 Kc 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 \((S, (L_1, F_1), (L_2, F_2), \ldots, (L_k, F_k))\) 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.
5.1. Estimation

Our estimator for $\mu^0_m$, the nominal length-of-stay for patient type $m$, is simply the empirical average

$$\hat{\mu}^0_m \triangleq \mu(\text{LOS}^0_{\text{low}})_m = \frac{\sum_i L_i 1\{F_i \neq \text{full}\} 1\{L_i^\text{a} > 0\} 1\{S^i \in m\}}{\sum_i 1\{F_i \neq \text{full}\} 1\{S^i \in m\}}.$$ 

where $\{F_i \neq \text{full}\}$ is the event that the ICU was full upon discharge of patient $i$ from his first ICU discharge and $\{S^i \in m\}$ is the event that the severity score of patient $i$ defines him as class $m$. Similarly $\sigma(\text{LOS}^0_{\text{low}})_m$ 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, $P(R|\text{Low})_m$. These readmitted patients relapse due to numerous medical reasons unrelated to being discharged; the discharge is likely to be a natural departure as there is no need to discharge patients in order to accommodate new ones when the ICU occupancy level is low and there are available beds. Thus,

$$P(R|\text{Low})_m = \frac{\sum_i 1\{F_i \neq \text{full}\} 1\{L_i^\text{a} > 0\} 1\{S^i \in m\}}{\sum_i 1\{F_i \neq \text{full}\} 1\{S^i \in m\}}.$$ 

where $\{L_i^\text{a} > 0\}$ denotes the event that patient $i$ was readmitted. 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:

$$\mu(\text{LOS}^R_{\text{low}})_m = \frac{\sum_i L_i^\text{b} 1\{F_i \neq \text{full}\} 1\{L_i^\text{a} > 0\} 1\{S^i \in m\} 1\{F_i \neq \text{full}\}}{\sum_i 1\{F_i \neq \text{full}\} 1\{L_i^\text{a} > 0\} 1\{S^i \in m\}}.$$ 

where $\{F_i \neq \text{full}\}$ the event that patient $i$ was discharged from a low occupancy ICU upon his second ICU discharge. Notice that $\mu(\text{LOS}^R_{\text{low}})_m$ is an estimate of patient length-of-stay upon readmission when the readmission is due to medical factors unrelated to demand-driven discharge. Table I states the values of the estimates for our data set including information about the relevant number of data points where relevant.

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, $P(R|\text{Full})_m$. We estimate this probability according to:

$$P(R|\text{Full})_m = \frac{\sum_i 1\{F_i = \text{full}\} 1\{L_i^\text{a} > 0\} 1\{S^i \in m\}}{\sum_i 1\{F_i = \text{full}\} 1\{S^i \in m\}}.$$ 

We have seen that patients who are not discharged in order to accommodate new patients may require readmission (Table I); 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
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 ($\mu_{\text{LOS}_{\text{low}}}^0$), readmission probability ($P(R|\text{Low})$) and readmission length-of-stay ($\sigma_{\text{LOS}_{\text{low}}}^R$) when discharged from a ‘low’ occupancy ICU. Length-of-stay is given in hours.

| Patient Type | # data points | $\mu(\text{LOS}_{\text{low}}^0)$ (hours) | $\sigma(\text{LOS}_{\text{low}}^0)$ | $P(R|\text{Low})$ | # 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                            |

5.2. 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 $P(R|\text{Full})_m - P(R|\text{Low})_m$. This is justified by the underlying assumption that for any severity score (i.e. health condition), a patient discharged from 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

$$
\mu(\text{LOS}^R_m) = \frac{\sum_i L_i^2 1\{F_i^1 = \text{full}\} 1\{L_i^2 > 0\} 1\{S_i \in m\} 1\{F_i^2 \neq \text{full}\}}{\sum_i 1\{F_i^1 = \text{full}\} 1\{L_i^2 > 0\} 1\{S_i \in m\} 1\{F_i^2 \neq \text{full}\}}.
$$

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

A full ICU is more likely to require readmission than otherwise and since severity scores and ICU occupancy levels upon discharge/departure were independent in our data set. 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 $P(\text{R}|\text{Low})_m/P(\text{R}|\text{Full})_m$, the remaining fraction of patients, $[P(\text{R}|\text{Full})_m - P(\text{R}|\text{Low})_m]/P(\text{R}|\text{Full})_m$ belong to the later category. This implies the following structural relationship

$$\frac{P(\text{R}|\text{Low})_m}{P(\text{R}|\text{Full})_m} \mu(\text{LOS}_{\text{low}}^R)_m + \frac{P(\text{R}|\text{Full})_m - P(\text{R}|\text{Low})_m}{P(\text{R}|\text{Full})_m} \frac{1}{\mu_\text{m}} = \mu(\text{LOS}_{\text{full}}^R)_m$$

which may be used to estimate $\mu_\text{m}^R$. Alternatively, this relationship allows us to estimate $\frac{\mu_\text{m}}{\mu_\text{m}^R}$ directly as $P(\text{R}|\text{Full})_m \mu(\text{LOS}_{\text{full}}^R)_m - P(\text{R}|\text{Low})_m \mu(\text{LOS}_{\text{low}}^R)_m$.

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.

Table 3  Estimated Model
Notice in Table 3 that patient type 8 yields a negative estimate of $\frac{p_m}{\mu_R}$ 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.

6. 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 vein:

- **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, $(1/\mu_0^m)$, 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. (2010)] 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 $p_m/\mu_R$ 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. While these decisions may in reality occur on a continuous basis, patient transfers are not instantaneous and the granularity of 6 minutes per hour is fine enough to emulate an actual ICU. Discharge policy simulations
are over 100 sample paths each. We use the parameters estimated in Table 3 for nominal length-of-stay, $\mu_0$, probability of readmission, $p_m$, and expected readmission load, $p_m/\mu^R_m$. We vary the probability of an arrival, $\lambda$ between 0 and 0.021 (i.e. between 0 and 5 arrivals on average every 24 hours). Arrival rate .021 corresponds to a turnover of 1/2 the beds in the ICU each day which is about the highest load seen in the ICU. It remains to specify the traffic mix across patient classes. Below we consider detailed simulation results for four 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 40% in some instances. The next best policy is the index policy based on the remaining LOS, i.e. the LOS index policy. However, because there are some patient types with moderate LOS, but high readmission load (see for instance patient type 7), this policy still results in higher readmission load. As expected, the random policy performs very poorly. 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 in order to accommodate the high influx of new patients. Consequently, the expected readmission load increases significantly.

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 $\lambda = 0.021$ the greedy policy incurs readmission load that is 16.7 hours lower than the next best policy (the LOS-index policy) which corresponds to the loss of a single ICU bed (in a 10 bed ICU) for a little more than half a day per week. The savings relative to the $p_m$-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 1,
its 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 would 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. While this is true, we can see in Figure 3 that the greedy policy still outperforms these two policies. The savings relative to the next best policy corresponds to 1.5 hours over one week at a net patient arrival rate of $\lambda = 0.021$. 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:** We now 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, its 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, and by a significant margin. The savings relative to the next best policy correspond to 49.2 hours over one week at a net patient arrival rate of $\lambda = 0.021$ (or 1 ICU bed out of 10 for 2 days over one week).
Empirically Proportional traffic: Finally we set the probability of arrival of each patient type according to the proportions seen in the empirical data. Therefore, patient type 9 is the most likely to arrive whereas patient types 3 and 7 are the least likely. We can see in Figure 5 that the greedy policy continues to outperform our benchmark policies. The savings relative to the next best policy correspond to 23.7 hours over one week at a net patient arrival rate of $\lambda = 0.021$ (or 1 ICU bed out of 10 for 1 day per week).

6.1. Impact on Mortality Rates

The cost metric inspiring our index policy is the expected load a demand-driven discharged patient might bring to the ICU upon potentially being readmitted. As discussed in Section 3.5, other cost metrics are just as easily introduced, provided one has calibrated a model quantifying the impact a demand-driven discharge might have on the metric of interest. One question that is natural to ask is whether demand-driven discharges impact patient mortality, and the extent of this effect. For the data set used in our simulation study, we calculate the mortality rate for each patient class as the fraction of patients who die at some point during their hospital stay following their first ICU
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.

discharge. As before we delineate patients who were discharged during a period of low occupancy (≤ 75%) versus one of high occupancy (> 75%).

| Patient | P(Death|Low) | P(Death|Full) |
|---------|----------|------------|
| 1       | 0        | 0          |
| 2       | 0        | 0          |
| 5       | 0        | 0          |
| 7       | .158     | .156       |
| 9       | .253     | .247       |

Table 4  In-hospital mortality rates for each patient class.

We notice that it is difficult to discern the impact of a demand-driven discharge on mortality. This is not particularly surprising: while there exist studies which suggest that demand-driven discharges increase mortality rates [Chrusch et al. (2009)], there are others which find that mortality risks are not predicted by occupancy levels [Iwashyna et al. (2000)].

We next consider two sets of simulations to understand the impact the demand-driven discharge policy employed might have on patient mortality. In particular, we compare the average number
Figure 5  Performance of greedy policy compared to benchmarks for various arrival rates and distribution across patient types according to the proportions seen in the empirical data.

of patient deaths within a week under each of the policies examined thus far, assuming that the proportion of arriving patients of each type correspond to the proportions observed in our data set. We consider two sets of simulations under distinct assumptions on mortality rates. In one simulation, we use the data in Table 4 and set the probability of death for a demand-driven and naturally discharged patient as \( P(\text{Death|Full}) \) and \( P(\text{Death|Low}) \), respectively. As expected, the policies are essentially indistinguishable given that the mortality rate is unaffected by whether or not the patient is demand-driven discharged; see Figure 6.

We next consider a scenario where the probability of death upon being demand-driven discharged is substantially increased. In particular, we consider two sets of experiments wherein we set the probability of death for a demand-driven discharged patient respectively 10% and 100% higher than the nominal probability of death for that patient class given in Table 4. Even after having significantly inflated the effect of demand-driven discharges on mortality, we find that the greedy index policy essentially incurs the lowest mortality rates of the policies considered here; see Figure 7.

6.2. Patient Diversions

Section 3.4 considered the scenario where patients can be diverted to other hospitals upon arrival to a full ICU rather than necessitating a demand-driven discharge. There is an inherent tradeoff
between blocking new patients and readmission load due to demand-driven discharges: increasing the number of blocked patients should reduce the number of demand-driven discharges and subsequently, reduce readmission load. We now consider this tradeoff via a numerical study.

We assume there are 10 beds in the ICU and, on average, a new patient arrives every 5 hours. A patient’s class is uniformly distributed across the 5 patient types described in Table 3. According to the heuristic proposed in Section 3.4, patients are diverted if the cost of blocking is less than the cost of a demand-driven discharge. We use this heuristic with the greedy policy as well as each of the benchmark policies. Hence, we make a demand-driven discharge if the readmission load of the patient selected by the policy in question (Greedy, LOS index, $p_m$ index, or Random) is smaller than the cost of blocking.

Because we do not have access to data regarding the costs of blocking patients, we assume the cost of blocking a patient is independent of patient type. We vary this cost to examine the tradeoff between blocking and readmission load. Figure 8 shows this tradeoff for the Greedy policy and the three benchmark policies. Not surprisingly, the Greedy policy achieves the best tradeoff while Random is the worst.
Figure 7  Number of patient deaths in one week under the greedy policy compared to benchmarks for various arrival rates. The distribution across patient types follows the proportions seen in the empirical data. Mortality rates due to demand-driven discharges are artificially inflated above natural discharge mortality rates by 10% and 100% respectively.

Figure 8  Tradeoff between blocking patients and readmission load of greedy policy compared to benchmarks
6.3. 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 $\mu^R_m$ 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 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 meaningful impact on the readmission load faced by an ICU. These gains can be summarized as being equivalent to an approximately 2 – 5% increase in effective ICU capacity.

3. The greedy index policy is robust to errors in estimation. Thus, the aforementioned increase in throughput comes at a relatively modest requirement to data collection requirements. In particular, the parameters $p_m$ and $\mu^R_m$ can be estimated from parameters available in existing historical data.

7. 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 reductions in readmitted patient load.
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 substantial reductions in readmitted patient load 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. It is remarkable that our model demonstrates benefits despite (from a clinical standpoint) being relatively simple—for example, it does not include diagnostic or physiologic data available at the time that a patient was discharged.

This work suggests several future potential research directions, including:

1. 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.

2. It would be interesting to understand the impact of a demand-driven discharge on other quantities of interest, particularly metrics measuring quality of life impact.

3. Theoretically, we have shown that our index policy is optimal in certain regimes and guaranteed to incur readmission loads of no greater that 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


**Appendix**

**A. 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 $\epsilon > 0$:

- For $i = 1$: $\mu_0^1 = 1/2, \ p_1 = p, \ \mu_R^1 = p$
- For $i = 2$: $\mu_0^2 = 1, \ p_2 = p, \ \mu_R^2 = \frac{p}{1-\epsilon}$

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, $x_{0,1} = 1$ and $x_{0,2} = 1$. Also, a new patient of type 1 arrives at $t = 0$ and $t = 1$, i.e. $y_{0,1} = y_{1,1} = 1$ while $y_{0,2} = y_{1,2} = 0$.

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 $\epsilon \to 0$ we see that $J^*(s_0) \leq \frac{4}{5}J^g(s_0)$ here.
B. A Connection with Throughput

Here we make precise the connection with throughput maximization described in Section 2.1. Consider an ICU with $C$ beds. We consider the following setup:

1. $B$ beds are reserved for first-time arrivals with $C - B \triangleq B'$ beds reserved for readmissions. Any reference to ‘state’ will be understood to correspond to the occupants of these $B$ beds and we will consequently employ the notation in Section 2.

2. The readmission queue is served according to a first-in-first-out discipline.

3. In the event that $B$ beds are occupied by first-time visitors, a new arrival will prompt a ‘demand-driven’ discharge according to a stationary policy $\pi$ that monitors the state of the $B$-beds reserved for first-time arrivals.

Note that readmitted patients cannot be demand-driven discharged. The rationale for this is natural: Readmitted patients are typically much worse off and have higher mortality rates and longer lengths-of-stay. This is well established in the medical literature (see [Chen et al. (1998), Durbin and Kopel (1993), Snow et al. (1985)] among others). As such, subjecting such patients to a demand-driven discharge is likely to be highly undesirable from a practitioners perspective. In addition, the policy that prioritizes patients should a demand-driven discharge be required may only consider the state of the $B$ beds reserved for first-time arrivals; one may dispense with this restriction, but doing so is beyond our scope here.

Given a vector $\lambda \in [0,1]^M$ defined so that $\lambda a_{t,m} \triangleq \lambda_m$ for all $m$ (assuming time homogenous rates), we will refer to a policy $\pi$ as stabilizing for $\lambda$ if, under this policy the readmission queue is stable. More precisely, we require the sequence of waiting times $\{W_n\}$ experienced by patients in the readmission queue (a waiting time is defined in the usual sense as the time between entry into the readmission queue and the time before service begins), has a sub-sequence that converges to a random variable $W$ that is a.e. finite.

Now, let us denote by the sequence $T_n$ the interarrival time between the $n$th and $(n+1)$st entry to the readmission queue, and by $S_n$, the service time required by the $n$th patient. Assume moreover that no demand-driven discharges occur in the absence of a need for one, i.e. $\pi(s) = \emptyset$ if $s \notin \{(x,y) : \sum_m x(s)_m + y(s)_m = B + 1; \sum_m y(s)_m = 1\} \triangleq \hat{S}_{\text{full}}$ (Recall again, that $s$ here corresponds to the state of the $B$ beds reserved for first-time admissions). Then, $T_n$ is simply the time between the $n$th and $(n+1)$st visit to a state in the set $\hat{S}_{\text{full}}$ while $S_n$ is a Geometric ($\mu_{\pi(s_n)}$) random variable with probability $p_{\pi(s_n)}$ (where $s_n$ corresponds to the state of the $B$ beds for first-time arrivals upon the $n$th discharge) and 0 with the remaining probability. Now, if $s_0 \sim \nu_\pi$, then it is not hard to see that $\{T_n, S_n\}$ is a stationary process. The process is also ergodic; a consequence
of the ergodicity of the Markov chain induced by $\pi$. A classical result of Loynes (Theorem 8 of [Loynes (1963)]) then establishes that the readmission queue is stable if $E[T_0] > E[S_0]/(C-B)$, and unstable if $E[T_0] < E[S_0]/(C-B)$. Now, elementary arguments (see [Durrett (1996)]) can be used to show that $E[T_0] = 1/\sum_{s \in \hat{\mathcal{S}}_{\text{full}}} \nu_{\pi}(s)$ and $E[S_0] = \sum_{s \in \mathcal{S}_{\text{full}}} \nu_{\pi}(s)C(s, \pi(s))/\sum_{s \in \mathcal{S}_{\text{full}}} \nu_{\pi}(s)$. In other words, we have that the readmission queue is stable if

$$\kappa^{\pi} < C - B,$$

and unstable if $\kappa^{\pi} > C - B$, so that minimizing $\kappa^{\pi}$ maximizes throughput which motivates the problem that is the focus of our study.

In addition, the following theorem shows that heuristics for the problem of minimizing long run readmission costs incur a proportionate ‘dilation’ of the set of arrival rate profiles that will result in stable readmission queues. In particular, let $\lambda$ be a vector of arrival rates that is in the interior of the throughput region for our model. By this we understand that there exists a demand-driven discharge policy $\pi^{\lambda}_\alpha$ under which the readmission queue is stable when the arrival rate vector is $\lambda$, and moreover there exists an $\epsilon > 0$ such that the arrival rate vector $\lambda(1 + \epsilon)$ can also be stabilized. Let us denote by $\pi^{\pi^{\lambda}_\alpha \lambda}$ a policy minimizing $\kappa^{\pi}$ for the arrival rate vector $\alpha \lambda$ where $\alpha \in (0, 1]$. Finally, let $\hat{\pi}_{\alpha \lambda}$ be a possibly sub-optimal demand-driven discharge policy for the arrival rate $\alpha \lambda$ satisfying $\kappa^{\hat{\pi}_{\alpha \lambda}}/\kappa^{\pi^{\pi^{\lambda}_\alpha \lambda}} \leq 1/\alpha$. We have:

**Theorem 5.** Assuming an arrival rate vector $\alpha \lambda$, the readmission queue is stable under the demand-driven discharge policy $\hat{\pi}_{\alpha \lambda}$.

**Proof:** Let us denote by $\pi^{\pi^{\lambda}_\alpha \lambda}$ (respectively $\pi^{\pi^{\lambda}_\lambda}$) a policy minimizing $\kappa^{\pi}$ in a system with arrival rate vector $\alpha \lambda$ (respectively $\lambda$). Now consider the following sub-optimal policy for an arrival rate $\alpha \lambda$: we simulate arrivals of ‘fictitious’ patients, so that the net stream of patients (both actual and fictitious) has arrival rate $\lambda$. To this system we apply policy $\pi^{\pi^{\lambda}_\lambda}$. Now by construction, a discharge under this policy will correspond to the discharge of an actual patient with probability $\alpha$; with the remaining probability, the discharge will be one of a fictitious patient and incur no costs. It thus follows that this sub-optimal policy incurs a cost of precisely $\alpha \kappa^{\pi^{\lambda}_\lambda}$. Moreover, since it is sub-optimal for the arrival rate vector $\alpha \lambda$, it must be that

$$\kappa^{\pi^{\lambda}_\alpha \lambda} \leq \alpha \kappa^{\pi^{\lambda}_\lambda}.$$  

It follows that

$$\kappa^{\hat{\pi}_{\alpha \lambda}} \leq (1/\alpha)\kappa^{\pi^{\lambda}_\alpha \lambda} \leq \kappa^{\pi^{\lambda}_\lambda}.$$
But given the fact that \( \lambda \) was in the interior of the stability region, it must be (by our earlier argument that showed \( \kappa^{\pi^*} \alpha \leq \alpha \kappa^{\pi^*} \)) that \( \kappa^{\pi^*} < C - B^* \), so that \( \kappa^{\pi^*} < C - B^* \), from which the claim follows.

We have demonstrated a stationary policy \( \pi^g \) satisfying, for a given arrival rate vector \( \lambda \), \( \kappa^{\pi^g} / \kappa^{\pi^*} \leq 1/(1 + \hat{\rho}) \) where \( \hat{\rho} \) was a measure of utilization. It follows that should the readmission queue be \textit{unstable} under \( \pi^g \), then it will remain unstable for any arrival rate vector that strictly dominates \( (1 + \hat{\rho}) \lambda \) under \textit{any} stationary discharge policy. In other words, the use of the \( \pi^g \) policy will correspond to a dilation of the throughput region by a factor corresponding to the approximation guarantee we have established.

C. Miscellaneous Technical Proofs

\textbf{Proof of Theorem 1} [1] 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 \( \pi^* \), \( \pi^*(s) \notin \text{arg min}_{m : x(s)_m > 0 \frac{p_m}{\mu_i}} \), 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 \( i = \pi^g(s) \in \text{arg min}_{m : x(s)_m > 0 \frac{p_m}{\mu_i}} \). Thus, by assumption we have that

\[ J^*(s) = C(s, j) + E [J^*(S(s, j))] < C(s, i) + E [J^*(S(s, i))]. \] (C1)

Now, let \( s_i = S(s, j) \), and \( s_j = S(s, i) \). We may assume that \( x(s)_k = x(s)_j \forall k \neq i, j \). 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) \( x(s)_i - x(s)_j = 1 \) and \( x(s)_j - x(s)_i = 0 \), (ii) \( x(s)_i - x(s)_j = 0 \) and \( x(s)_j - x(s)_i = 1 \) or (iii) \( x(s)_i - x(s)_j = 1 \) and \( x(s)_j - x(s)_i = 1 \). 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 $\mu_j^0$. If patient $j$ departs, patient $i$ will depart in the same time step with probability $\mu_i^0/\mu_j^0$; 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 $\mu_i^0$ and $\mu_j^0$ 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 $\mu_i^0/\mu_j^0$, and with the remaining probability assume a ‘virtual’ additional type $i$ patient that will complete service in subsequent periods with probability $\mu_i^0$. 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 $\pi'$. 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, $J^*(s_j) = \bar{C} + \bar{p}_i C(s, j) = J^*(s_i) + \bar{p}_i[C(s, j) - C(s, i)]$. 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 (C1) (since $C(s, i) \neq C(s, j)$) and yields our result. \hfill \Box

**Proof of Lemma $\blacksquare$** 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-($z(s'_t)_m, \mu^0_m$) 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-$(z(s_t)_m - z(s'_t)_m, \mu^0_m)$ 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

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

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. \qed
Proof of Lemma \[\text{2}.\] Without loss, we assume \(\pi^*(s) \neq \pi^\theta(s)\) (else, there is nothing to prove). By definition, we must have \(x(s)_{\pi^*(s)}, x(s)_{\pi^\theta(s)} > 0\). Let \(\tilde{S}(s, \pi^*(s))\) be the next state obtained if one discharged \textit{both} \(\pi^*(s)\) and \(\pi^\theta(s)\) in state \(s\). In particular, we define \(\tilde{S}(s, \pi^*(s)) \triangleq \tilde{s}\) according to

\[
x(\tilde{s})_{\pi^*(s)} = x(s)_{\pi^*(s)} + y(s)_{\pi^*(s)} - 1 - X_{t(s), \pi^*(s)},
\]

\[
x(\tilde{s})_{\pi^\theta(s)} = x(s)_{\pi^\theta(s)} + y(s)_{\pi^\theta(s)} - 1 - X_{t(s), \pi^\theta(s)},
\]

\[
x(\tilde{s})_m = x(s)_m + y(s)_m - X_{t(s), m}, \quad m \neq \pi^*(s), \pi^\theta(s)
\]

\[
y(\tilde{s})_m = Y_{t(s) + 1, m},
\]

\[
t(\tilde{s}) = t(s) + 1,
\]

where analogous to our earlier description of \(S(s, a)\), we define \(X_{t(s), \pi^*(s)}\) (resp. \(X_{t(s), \pi^\theta(s)}\)) as a Binomial \((x(s)_{\pi^*(s)} + y(s)_{\pi^*(s)} - 1, \mu^0_{\pi^*(s)})\) (resp. Binomial \((x(s)_{\pi^\theta(s)} + y(s)_{\pi^\theta(s)} - 1, \mu^0_{\pi^\theta(s)})\)) random variable. For \(m \neq \pi^*(s), \pi^\theta(s)\), we define \(X_{t(s), m}\) as a Binomial \((x(s)_m + y(s)_m, \mu^0_m)\) random variable. \(Y_{t(s) + 1, m}\) is defined as before for all \(m\). Now, by construction, \(x(\tilde{S}(s, \pi^*(s))) \leq x(S(s, \pi^*(s)))\), while \(y(\tilde{S}(s, \pi^*(s))) = y(S(s, \pi^*(s)))\), so that by Lemma \[\text{1}\] we have that \(E[J^*(\tilde{S}(s, \pi^*(s)))] \leq E[J^*(S(s, \pi^*(s)))].\)

Now, let us consider the following sub-optimal policy \(\pi'\) for the system in which the greedy action is taken at state \(s\). Define \(\tau = \min\{T > t > t(s) : \sum_m Y_{t,m} = 1\}\); i.e. \(\tau\) is the first time after the current time step \(t(s)\) that an arrival occurs (or infinite if no arrival occurs prior to time \(T\)). Then on the event that \(x(s,_{\pi^*(s)}) = x(s)_{\pi^*(s)} + y(t,s)_{\pi^*(s)} - 1, \pi'\) simply takes the optimal action for \(t \geq \tau\) (so that, in fact \(\pi'\) coincides with \(\pi^*\) on this event). On the event that \(x(s,_{\pi^*(s)}) = x(s)_{\pi^*(s)} + y(t,s)_{\pi^*(s)}, \pi'_{(s,_{\tau})} = \pi^*(s)\), and \(\pi'\) takes actions according to the optimal policy \(\pi^*\) for \(t > \tau\). The probability that an eviction occurs under \(\pi'\) at \(\tau\) is simply the probability that no patient of type \(\pi^*(s)\) has departed prior to the next arrival; an event whose probability is at most \(\lambda/(\lambda + \mu^0_{\pi^*(s)})\). Observe moreover that we may couple the systems starting at state \(S(s, \pi^\theta(s))\) and \(\tilde{S}(s, \pi^*(s))\) so that under the \(\pi'\) policy in the former system and the optimal policy in the latter, both state processes agree on \(t > \tau\), and moreover, no eviction will be required at times \(t \leq \tau\) in the latter system. It follows that

\[
E[J^\pi'(S(s, \pi^\theta(s)))] \leq \frac{\lambda}{\lambda + \mu^0_{\pi^*(s)}}C(s, \pi^*(s)) + E[J^\pi^*(\tilde{S}(s, \pi^\theta(s))].
\]

Since \(E[J^\pi^*(S(s, \pi^\theta(s))) \geq E[J^\pi'(S(s, \pi^\theta(s)))\] and as established earlier, \(E[J^\pi^*(\tilde{S}(s, \pi^\theta(s)))] \leq E[J^\pi^*(S(s, \pi^\theta(s)))\), the result follows. \(\square\)