Optimizing Intensive Care Unit Discharge Decisions with Patient Readmissions

Citation: Chan, Carri W., Vivek F. Farias, Nicholas Bambos, and Gabriel J. Escobar. "Optimizing Intensive Care Unit Discharge Decisions with Patient Readmissions." Operations Research 60, no. 6 (December 2012): 1323–1341.

As Published: http://dx.doi.org/10.1287/opre.1120.1105

Publisher: Institute for Operations Research and the Management Sciences (INFORMS)

Persistent URL: http://hdl.handle.net/1721.1/87734

Version: Original manuscript: author's manuscript prior to formal peer review

Columbia Business School

Columbia Business School Research Paper Series

“Optimizing ICU Discharge Decisions with Patient Readmissions”

Carri W. Chan

Vivek F. Farias

Nicholas Bambos

Gabriel J. Escobar

Optimizing ICU Discharge Decisions 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: August 30, 2011

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.

We study the impact of several different ICU discharge strategies on patient mortality and total readmission load. We focus on discharge rules that prioritize patients based on some measure of criticality assuming the availability of a model of readmission risk. We use empirical data from over 5000 actual ICU patient flows to calibrate our model. 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 while at the same time maintaining, or even improving upon, mortality rates. We explicitly provide a discharge policy that accomplishes this. In addition to our empirical work, we conduct a rigorous performance analysis for the family of discharge policies we consider. We show that our policy is optimal in certain regimes, and is otherwise guaranteed to incur readmission related costs 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. It is natural to conjecture that demand-driven discharges might be associated with *costs*; namely:

- **Patient Health Related Costs:** Patients subject to a demand-driven discharge could potentially face additional risks of physiological deterioration. Such deterioration might ultimately require readmission. 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)).
- **System Related Costs:** Readmitted patients impose an additional load on capacity-limited ICU resources. Ultimately this hampers access to the ICU for *other* patients

Thus motivated, the present work 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 better patient outcomes and *simultaneously* increase efficiencies in the use of scarce ICU resources. More formally, associating a demand-driven discharge with some cost which depends on the physiological characteristics of the patient discharged, our goal is to 'optimally' discharge patients so as minimize total expected costs associated with demand-driven discharges over time. One example of such a cost may be the increase in mortality risk due to a demand-driven discharge. As a second example, one might consider the increase in expected readmission load associated with the increased likelihood of readmission due to a demand-driven discharge. We will eventually estimate and test several such cost metrics.

Our analysis will consider a stylized model of an actual ICU where the number of ICU beds is fixed[1]. Patients arrive to the ICU at random times; patients are categorized into a finite number

[1] 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.

of classes based on their physiological characteristics upon admission. There exist a number of proprietary classification systems based on a patients physiological characteristics. 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[2]. The demand-driven discharge of a patient will incur a cost which depends on that patient's class; this cost is modeled to reflect the impact of the demand-driven discharge on the patient as well as the system as described above. Our goal will be to minimize the expected costs incurred due to demand-driven discharges over some finite horizon. This is a difficult problem, and our analysis of this stylized model will suggest simple policies for which we will develop performance guarantees. More interestingly, we will conduct a detailed simulation study based on real data to examine our recommendations.

1.1. Our Contributions

We make the following key contributions:

- **Interpretability:** We show that a *myopic* policy is a potentially good approximation to an optimal policy. This corresponds to an index policy wherein every patient class is associated with a class specific index. The index for a given class can be computed from historical patient flow data in a robust fashion. Depending on the cost metric under consideration, we will demonstrate that these indices can serve as natural measures for patient criticality that have both clinical as well as operational merit. The index policy then has an appealing clinical interpretation: when a patient must be discharged in order to accommodate new patients, one simply discharges an existing patient of the lowest possible criticality index.

- **Robustness:** 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 quantities relevant to specific classes of patients that are typically far simpler to estimate from data. For the data set under consideration, relative changes of estimated parameters greater than 50% were typically required to induce a change in the associated indices.

- **Performance Guarantees and Operational Relevance:** We demonstrate via a theoretical analysis that our index policy is, for a certain class of problems, optimal and in general incurs total expected cost 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 identify a cost metric – the increase

[2] We later consider an extension of our model which includes the additional option of blocking new patients.

in expected readmission load due to a demand-driven discharge – that in addition to enjoying a clinical interpretation as a measure of criticality, can be shown to capture a notion of throughput optimality.

- **Empirical Validation:** Most importantly, we calibrate our model to empirical data from over 5000 patient flows at a large privately owned partnership of hospitals and identify parameters for patient dynamics. We consider a variety of cost metrics, including several natural metrics motivated by existing clinical literature and modifications of these cost metrics such as the operationally relevant metric alluded to above. We measure the impact of these discharge policies along two dimensions. First, to understand impact at the individual patient level, we measure mortality rates under the various policies. Second, to understand system level impact we measure the readmission load incurred under the various policies. In doing so, we identify a policy that, in addition to fitting within the ethos of ordering patients by a measure of criticality, has substantive benefits over other, perhaps more 'obvious' policies: *Under modest assumptions on patient traffic, it incurs a* 30% *reduction in readmission load at no cost to mortality rate.*

As such, this study provides a framework for the design of demand-driven discharge policies and in doing so identifies a policy that allows us to utilize available ICU resources as effectively as possible while not sacrificing the quality of patient outcomes. 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.

1.2. 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 2011), 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 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 and ultimately compromising the quality of care for all patients involved. 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, relative to recent work by (Chan and Farias 2009), we note that the present paper considers a class of models entirely distinct from the 'depletion problems' studied there 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. Section 2 formally introduces the queueing model and patient dynamics we study. In Section 3, we analyze the performance of an index policy which selects patients to discharge in a greedy manner based on their expected costs incurred due to demand-driven discharges. 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 system utilization. In Section 4, we discuss various measures of criticality which constitute clinically relevant cost metrics. These measures include an important refinement to a criticality measure that has received some attention in the critical care literature. 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 our primary proposal 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 is undesirable for the discharged patient and comes at a ‘cost’. 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 suffer physiologic deterioration due to the demand-driven discharge. Since arriving patients cannot be queued or blocked, the model we consider is distinct from a typical queueing model. Presuming a measure of cost associated with a demand-driven discharged patient, a natural goal is to find a patient discharge policy that minimizes this cost.

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 and identical costs associated with a demand-driven discharge. Specifically, we assume that the initial length-of-stay for a patient of class m is a geometric random variable with mean $1/\mu_m^0$. If such a patient is discharged prior to completing treatment due to the arrival of a more acute patient, a cost, $\phi_m \geq 0$, is incurred. While the patient length-of-stay distribution is assumed to be memoryless for the purposes of analysis, our empirical study assumes log-normal distributions for length-of-stay that are fit to the empirical data (see Section 5). Finally, in Section 3.3, we discuss an extension to our model which is able to capture a patient’s evolution and changing condition during his ICU stay by using a ‘phase’-type length-of-stay distribution.

At most one new patient can arrive in each time-slot and an arrival occurs with probability λ. We define $\hat{\rho} = \frac{\lambda}{\min_m \mu_m^0}$ 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 λ nor the probabilities $a_{t,m}$.

We assume that the ICU has B beds. If all B beds are full and a new patient arrives, then a patient must be discharged prior to completing service in order to accommodate the newly arrived patient. We let $x_{t,m} \in \{0, 1 \ldots, B\}$ 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 tth 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 tth time-slot with probability μ_m^0 after any demand-driven discharge and/or admission occurs.

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

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

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 $s_t \in \mathcal{S}$ denote the state in the tth epoch. Note that because the $\{a_{t,m}\}$ process is assumed to be deterministic and given a-priori, the current time slot t completely specifies the arrival probabilities for each patient class.

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

Dynamics: Let $s' = S(s, A)$ denote the random next state encountered upon employing action A (demand-driven discharge of patient type A) in state s. A random number, $X_{t(s),m}$, of class m patients will complete treatment and depart naturally, where $X_{t(s),m}$ is a Binomial-$(x(s)_m + y(s)_m -$

$\mathbf{1}_{\{A=m\}}, \mu_m^0)$ random variable. Let R_t be independent random variables, defined for each t, indicating the type of an arriving patient at the start of the tth epoch. R_t takes values in $\{1,2,\ldots,M\} \cup \{0\}$; $R_t = m$ with probability $\lambda a_{t,m}$ for $m \in \{1,2,\ldots,M\}$ and $R_t = 0$ with the remaining probability. The vector denoting arrivals at the next state, $Y_{t(s)+1}$ is then given by $Y_{t(s)+1,m} = \mathbf{1}_{\{R_{t(s)+1}=m\}}$. Thus, $s' = S(s,A)$ is defined as:

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

Cost Function: The cost incurred for taking action A is defined by a cost function $C : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}_+$. 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 discussion is able to capture *any* such cost function. We take $C(s,A) = \phi_A$ for $A \in \{1,2,\ldots,M\}$, and $C(s,0) = 0$. In Section 4, we discuss clinically relevant cost metrics.

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

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

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

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

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

for all $s \in \mathcal{S}$ with $t(s) \leq T-1$. J^* may then be found as the solution to the Bellman equation $\mathcal{H}J = J$, with the boundary condition $J(s') = 0$ for all s' with $t(s') = T$. The optimal policy π^* may be found as the greedy minimizer with respect to J^* in (1). The 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 $\mathcal{S}$ precludes this straightforward dynamic programming approach. Moreover, 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 π (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 $\kappa^{\pi}(s)=\limsup_{T\to\infty}\frac{1}{T}E\left[\sum_{t'=t(s)}^{T-1}C(s_{t'},\pi(s_{t'}))\middle|s_{t(s)}=s\right]$ is the average-cost to go (i.e. the long run costs incurred due to demand-driven discharges) under policy π.

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

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

where ν_{π} is the stationary distribution induced by π on $\hat{\mathcal{S}}$.

3. A Priority Based Policy

This section introduces an index policy for the dynamic optimization problem proposed. Under such a policy, the patient selected for a demand-driven discharge is simply chosen from a patient class that would incur the minimal cost. In particular, such a policy states that the patient (class) $\pi^g(s)$ chosen for discharge satisfies:

$$\pi^g(s)\in\arg\min_{A\in\mathcal{A}(s)}C(s,A)=\arg\min_{m\in\mathcal{A}(s)}\phi_m. \tag{2}$$

It is easy to see that the policy specified by (2) has a natural implementation as an 'index' policy. It is interesting to note that implementing such a policy requires data about particular patient classes, but does not require the estimation of arrival rates of the various classes. This latter information is highly dynamic and difficult to estimate.

Since the policy we have proposed ignores the effect of future arrivals and the expected length-of-stay of the current occupants, it is natural to expect such a policy to be sub-optimal. In the appendix, Example A shows what can go wrong.

In light of the sub-optimality of our proposed priority based 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 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 with $\phi_i \le \phi_j$ we also have $1/\mu^0_i \ge 1/\mu^0_j$. 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 cost also have higher nominal lengths-of-stay. In this case, since eliminating a low cost 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 $s \in \mathcal{S}$, $J^g(s) \le (\hat{\rho}+1)J^*(s)$, where $\hat{\rho}=\frac{\lambda}{\mu^0_{\min}}$ 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 that having lower occupancy levels. In particular:

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

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

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 $\hat{\rho} = \frac{\lambda}{\mu^0_{\min}}$ is the utilization ratio of the ICU defined in Section 2.

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

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

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

$$\begin{aligned}
J^*(s) &= C(s,\pi^*(s)) + E[J^*(S(s,\pi^*(s)))] \\
&\geq (1-\alpha)C(s,\pi^*(s)) + E[J^*(S(s,\pi^g(s)))] \\
&\geq (1-\alpha)C(s,\pi^g(s)) + E[J^*(S(s,\pi^g(s)))] \\
&\geq (1-\alpha)C(s,\pi^g(s)) + E[(1-\alpha)J^g(S(s,\pi^g(s)))] \\
&= (1-\alpha)J^g(s) \\
&= \frac{1}{\hat{\rho}+1}J^g(s) \qquad (3)
\end{aligned}$$

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

The state space now needs to be expanded to incorporate the different health states each patient class can achieve. To do this, we can redefine $x(s)$ to be a 2-dimensional array where $x_{m,n}(s)$

denotes the number of class m patients in health condition state h_n^m. We consider using the natural analogue to the greedy policy discussed thus far:

$$\tilde{\pi}^g(s) \in \underset{(m,n):x_{m,n}(s)>0}{\arg\min} \phi_m(h_n^m)$$

Now, Lemma 1 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 $\alpha = \frac{\tilde{\rho}}{\tilde{\rho}+1}$,

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

where we now define

$$\tilde{\rho} = \frac{\lambda}{\min_{m,n} \mu_m^0(h_n^m)}.$$

With these results, the proof of Theorem 2 applies verbatim to yield

Theorem 3. *For all $s \in \mathcal{S}$, $J^{\tilde{\pi}^g}(s) \leq (\tilde{\rho}+1)J^*(s)$.*

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 $s \notin \hat{\mathcal{S}}_{\text{full}}$, 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 $s \in \mathcal{S}$, $J^{\hat{\pi}}(s) \leq (\hat{\rho}+1)J^*(s)$.*

3.5. 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 and cost of demand-driven discharge of each patient type which we vary uniformly at random with mean 25 hours and 2.5 units of cost, respectively. For each fixed set of parameters–$a_{i,t}$, μ_i^0, and ϕ_i–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.

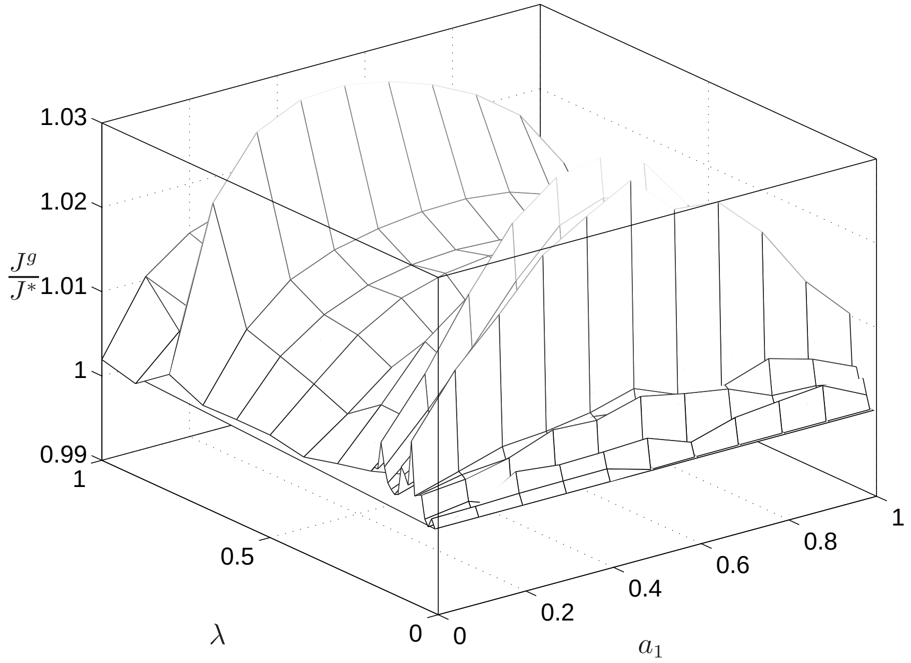

Figure 1 Performance of greedy policy compared to optimal for varying arrival rates.

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 λ while the probability an arrival is of patient type 1 is given by a_1. Values above 1 show the loss in performance due to using the greedy policy. We can see that the greedy policy performs within 3% of optimal, which 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.

4. Clinical Relevance

Our exposition thus far has treated the problem of prioritizing patients for demand-driven discharges as a purely *operational* problem. In a nutshell, we have shown that if one desires to minimize some long run cost metric impacted by demand-driven discharge decisions, then a priority rule that is 'greedy' with respect to the cost metric serves as a reasonable and operationally viable approximation to an optimal policy.

This section considers *clinical* issues relevant to the problem at hand. In particular, the clinical viability of a discharge policy is of paramount importance. In particular, what remains to be specified are clinically relevant cost metrics and priority rules which capture factors physicians would like to account for in making discharge decisions. Certainly, the general consensus of the medical community is that patients should be discharged in order of 'least critical first' (see, for instance, Swenson (1992)). However, what determines criticality is left wide open to interpretation and is highly dependent on the experience and training of an individual physician. In fact, disagreements on which patient should be discharged arise frequently and in an effort to building a process around this critical decision, many hospitals are adopting an intensivist-managed system that makes triage decisions for all patients in the ICU (Franklin et al. 1990, Task Force of the American College of Critical Care Medicine 1999). While such a process will remain necessarily subjective, there is a strong desire that the process be informed by quantitatively designed best-practice recommendations. In this sprit, we consider several policies that fall within the ethos of a priority rule based on measures of patient criticality that have been broached in the extant medical literature.

Mortality Risk: A natural measure of patient 'criticality' is mortality risk. In fact, the commonly used APACHE and SAPS severity scores are based on mortality predictions for ICU patients (Zimmerman et al. 2006, Moreno et al. 2005). While it is obvious that patients with high mortality risk are ‘critical’ and should not be demand-driven discharged, intensivists are likely to find this measure of criticality too crude to be of value in practical scenarios. To be more precise, one typically needs to be able to distinguish among patients all with relatively low mortality risk but variedly long and complex recoveries. In addition, a metric based solely on mortality risk will fail to capture a system-wide view of the ICU and in particular, the impact a discharge decision for a given patient might have on the ability to provide timely and quality care for other patients. Specifically, such a metric fails to account for the impact a discharge decision has on ICU *congestion* – congestion in the ICU can result in postponing surgeries, delaying admissions, and/or rerouting patients to other units–all of which are associated with worse outcomes (Metcalfe et al. 1997, Mitchell et al. 1995, Smith et al. 1995, Chalfin et al. 2007, Renaud et al. 2009, Rincon et al. 2010). As such, it is ethically important to consider factors related to congestion in making such decisions.

Readmission Risk: A potential refinement on using simply mortality risk as a measure of patient criticality is accounting for readmission risk. In fact, measures related to readmission risk have been gaining attention and credibility in the medical community motivated primarily by two factors: medical outcomes and payment structures. In terms of medical outcomes, readmitted patients have been shown to be worse off, with higher mortality and longer length-of-stay (Chen et al. 1998, Durbin and Kopel 1993, Rosenberg and Watts 2000). Recognizing the clinical risks associated with readmissions, many hospitals are adopting discharge strategies which account for patient readmissions (Franklin and Jackson 1983, Yoon et al. 2004). In terms of monetary incentives, readmissions can also increase costs by over 25% (Naylor et al. 2004). Acknowledging the detrimental impact of readmissions on patient outcomes and the extraordinarily high costs associated with the care of readmitted patients, the Patient Protection and Affordable Care Act (2010) requires Medicare to begin reducing readmissions in 2013. While physiology-based probabilistic models for assisting ICU physicians in making discharge decisions are not widely available, there has been recent interest in developing risk scores to assess readmission risks, similar to what the APACHE and SAPS scores do for mortality (Gajic et al. 2008). In this spirit, one may consider several concrete metrics:

A Crude Metric: As a concrete measure of readmission risk, one might consider the **likelihood of readmission**. One expects that such a measure will be fairly correlated with a measure of mortality risk. At the same time, such a measure will move towards addressing some of the pitfalls of using mortality risk alone. That said, such a measure remains somewhat coarse in two regards: First, it fails to account for the actual impact of the demand-driven discharge decision itself on readmission risk; since readmissions might arise due to a multitide of other factors, this is crucial. Second, it fails to account for the diversity in complications that might occur *upon* a readmission.

A Refinement ***(Our Proposed Policy):*** We consider a mild refinement to the above measure of readmission risk: we consider the *increase in readmission load, attributable to a demand-driven discharge*. Roughly speaking, we can think of this refinement as accounting not only for readmissions, but in addition, the typical length of stay upon such a readmission. More precisely, let p_m^N and $1/\mu_m^{R,N}$ be the probability of readmission and expected readmission LOS of patient class m given he is naturally discharged. Similarly, let p_m^D and $1/\mu_m^{R,D}$ be the probability of readmission and expected readmission LOS of patient class m given he is demand-driven discharged. By Chen et al. (1998), we expect to have $p_m^N < p_m^D$ and $\mu_m^{R,N} > \mu_m^{R,D}$. Then the *increase in readmission load attributable to the demand-driven discharge* is precisely:

$$\Delta\text{-Readmission Load} = \frac{p_m^{R,D}}{\mu_m^{R,D}} - \frac{p_m^{R,N}}{\mu_m^{R,N}}$$

We will in the subsequent sections consider a priority rule that measures patient criticality via the Δ-Readmission Load score. In addition to fitting in with the ethos of a priority rule that can be interpreted as a criticality measure, we see that this rule is consistent with assuming, in the notation of the previous Sections, a one period cost-function $C(s, A)$ that corresponds to the increase in readmission load due to the demand-driven discharge decision. In the appendix, we show that such a cost metric is also explicitly aligned with the desire to avoid a loss of throughput due to congestion effects.

Other Measures of Criticality: While we have outlined the two broad criticality measures one might consider in the medical community, yet other measures have been proposed in the operations research community. In particular, Dobson et al. (2010) considers prioritizing patients based on a patients expected length of *remaining* stay. Unfortunately, this is a fairly difficult quantity to estimate and as such models to predict this quantity are also unavailable. For completeness, we will also consider this measure in our empirical investigation.

5. Empirical Data

The goal of this section is to calibrate a model from real data that will permit us to compare the clinically relevant policies discussed in the preceding section. We analyze patient data from 7 different private hospitals for a total of $5,398$ patients who completed at least one ICU visit.

Patient Classes: Our first goal is to classify patients into a small number of groups, each of which is defined on the basis of physiological variables. There are may ways of doing this, and we chose a method that is aligned with the current process design philosophy of the hospital system from which the data for this study was obtained. In particular, we classified patients into 5 different classes based on ‘severity scores’ (see Escobar et al. (2008)) which are used to predict the likelihood of death. These 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.), etc. They are quite similar to the well studied APACHE and SAPS scoring systems (for instance, the c statistic for this score is in the 0.88 range) with the important addition that they incorporate additional physiological information obtained for patients in this particular hospital system within a short time prior to their being admitted to the hospital (that APACHE or SAPS scores would not assume available). Like scoring rules of this type, the severity scores we use to classify patients may be interpreted as a mortality risk figure. We quantize these severity scores into one of five different bins of equal size. Table 1 summarizes the severity scores for the 5 patient classes as well as the percentage of survivors. It is important to note that we only use these scores as a convenient and clinically interpretable way of classifying patients. We do not use the severity score of a patient for the purposes of predicting mortality, length of stay, probability of readmission and so-forth; rather, we directly estimate all of these factors from data.

Patient Class	Range for predicted mortality	# data points	% survivors
1	[0,.0048)	1089	99.5%
2	[.0048,.0148)	1084	97.0%
3	[.0148,.039)	1097	94.7%
4	[.039,.1025)	1067	91.8%
5	[.1025,1)	1061	85.4%

Table 1 Patient Classes

ICU Occupancy Levels: Our data set indicates the utilization of the ICU upon 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 (2011) 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 death, the likelihood of their readmission and the lengths of any subsequent visits depend solely on their health condition as summarized by their severity score, 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 (Full versus Low), 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 $(S, D, (L_1, F_1), (L_2, F_2), \ldots, (L_k, F_k))$ where S is a severity score, D is an indicator of patient death during hospital stay, L_n is the patient length-of-stay on his nth visit to the ICU in the episode and F_n is an indicator for whether the ICU was full upon his nth discharge.

5.1. Estimation

We first estimate the probability of death for patients discharged from a low versus full ICU. We estimate the nominal probability of death, $\mathrm{P(D|Low)}_m$, using the fraction of patients who were discharged from a low occupancy ICU and died during the same hospital stay.

$$\mathrm{P(D|Low)}_m = \frac{\sum_i \mathbf{1}_{\{D^i=1\}}\mathbf{1}_{\{F_1^i=0\}}\mathbf{1}_{\{S^i\in m\}}}{\sum_i \mathbf{1}_{\{F_1^i=0\}}\mathbf{1}_{\{S^i\in m\}}}.$$

where $\{F_1^i = 0\}$ is the event that the ICU occupancy level was low 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, we can calculate the probability of death when discharged from a full ICU.

$$\mathrm{P(D|Full)}_m = \frac{\sum_i \mathbf{1}_{\{D^i=1\}}\mathbf{1}_{\{F_1^i=1\}}\mathbf{1}_{\{S^i\in m\}}}{\sum_i \mathbf{1}_{\{F_1^i=1\}}\mathbf{1}_{\{S^i\in m\}}}.$$

Table 2 summarizes the estimated probabilities of death for each patient class along with the 95% confidence interval for these estimates.

We notice that it is difficult to discern any substantial 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 (for example (Chrusch et al. 2009)), there are others which find that mortality risks are not predicted by occupancy levels (Iwashyna et al. 2000).

Our estimator for the nominal length-of-stay (LOS) for patient type m, is simply the empirical average

$$\mu(\mathrm{LOS}^0_{\mathrm{low}})_m = \frac{\sum_i L_1^i \mathbf{1}_{\{F_1^i=0\}}\mathbf{1}_{\{S^i\in m\}}\mathbf{1}_{\{D^i=0\}}}{\sum_i \mathbf{1}_{\{F_1^i=0\}}\mathbf{1}_{\{S^i\in m\}}\mathbf{1}_{\{D^i=0\}}}.$$

Patient Class	# data points	$\mathrm{P(D \mid Low)}$	[95% CI]	# data points	$\mathrm{P(D \mid Full)}$	[95% CI]
1	739	.005	[.000,.010]	350	.003	[.000,.009]
2	682	.022	[.011,.033]	402	.017	[.004,.030]
3	679	.059	[.041,.077]	418	.043	[.024,.062]
4	669	.079	[.059,.099]	398	.088	[.060,.116]
5	621	.167	[.138,.196]	440	.116	[.086,.146]

Table 2 Mortality: probability of death when patients naturally depart and when patients are demand-driven discharged.

where $\{F_1^i = 0\}$ is the event that the ICU occupancy level was low 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 $\sigma(\mathrm{LOS}^0_{\mathrm{low}})_m$ is an empirical standard deviation. Note that when calculating LOS, we exclude patients who died. This is common practice in the medical community because various factors, such as Do-not-resuscitate orders can skew LOS estimates for patients who die (Norton et al. 2007, Rapoport et al. 1996).

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, $\mathrm{P(R|Low)}_m$. 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,

$$\mathrm{P(R|Low)}_m = \frac{\sum_i \mathbf{1}_{\{L_2^i>0\}}\mathbf{1}_{\{F_1^i=0\}}\mathbf{1}_{\{S^i\in m\}}}{\sum_i \mathbf{1}_{\{F_1^i=0\}}\mathbf{1}_{\{S^i\in m\}}}.$$

where $\{L_2^i > 0\}$ 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:

$$\mu(\mathrm{LOS}^R_{\mathrm{low}})_m = \frac{\sum_i L_2^i \mathbf{1}_{\{F_1^i=0\}}\mathbf{1}_{\{L_2^i>0\}}\mathbf{1}_{\{S^i\in m\}}\mathbf{1}_{\{F_2^i=0\}}\mathbf{1}_{\{D^i=0\}}}{\sum_i \mathbf{1}_{\{F_1^i=0\}}\mathbf{1}_{\{L_2^i>0\}}\mathbf{1}_{\{S^i\in m\}}\mathbf{1}_{\{F_2^i=0\}}\mathbf{1}_{\{D^i=0\}}}.$$

where $\{F_2^i = 0\}$ denotes the event that patient i was discharged from a low occupancy ICU upon his second ICU discharge. Again, we exclude patient who died in this estimation. Notice that $\mu(\mathrm{LOS}^R_{\mathrm{low}})_m$ is an estimate of patient length-of-stay upon readmission when the readmission is due to medical factors unrelated to demand-driven discharge. Table 3 states the values of the estimates for our data set including information about the relevant number of data points.

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

Patient	# data	$\mu(\mathrm{LOS}^0_{\mathrm{low}})$	$\sigma(\mathrm{LOS}^0_{\mathrm{low}})$	P(R|Low)	[95% CI]	# data	$\mu(\mathrm{LOS}^{\mathrm{R}}_{\mathrm{low}})$	$\sigma(\mathrm{LOS}^{\mathrm{R}}_{\mathrm{low}})$
CLass	points	(hours)				points	(hours)	
1	735	45.7	134.2	.073	[.054,.092]	34	36.1	40.5
2	667	46.7	50.8	.095	[.073,.117]	46	66.0	118.1
3	639	59.7	98.4	.102	[.079,.125]	39	106.9	212.5
4	616	78.1	201.8	.115	[.091,.139]	45	110.5	289.3
5	517	89.6	116.7	.119	[.094,.115]	34	161.4	365.5

Table 3 Nominal patient parameters: operational parameters when patients naturally depart and are not discharged in order to accommodate new patients. Average initial length-of-stay ($\mathrm{LOS}^0_{\mathrm{low}}$), readmission probability P(R|Low) and readmission length-of-stay ($\mathrm{LOS}^{\mathrm{R}}_{\mathrm{low}}$) when discharged from a 'low' occupancy ICU. Length-of-stay is given in hours.

$$\mathrm{P}(\mathrm{R}|\mathrm{Full})_m = \frac{\sum_i \mathbf{1}_{\{F_1^i=1\}}\mathbf{1}_{\{L_2^i>0\}}\mathbf{1}_{\{S^i\in m\}}}{\sum_i \mathbf{1}_{\{F_1^i=1\}}\mathbf{1}_{\{S^i\in m\}}}.$$

We have seen that patients who are not discharged in order to accommodate new patients may require readmission (Table 3); we expect that patients who are discharged from a full ICU may require readmission for those same reasons *in addition* to complications which arise due to being demand-driven discharged. Therefore, we expect the probability of readmission when discharged from a full ICU to be higher than when discharged from a low ICU. We also estimate the expected length-of-stay of such readmitted patients according to

$$\mu(\mathrm{LOS}^R_{\mathrm{full}})_m = \frac{\sum_i L_2^i \mathbf{1}_{\{F_1^i=1\}}\mathbf{1}_{\{L_2^i>0\}}\mathbf{1}_{\{S^i\in m\}}\mathbf{1}_{\{F_2^i=0\}}\mathbf{1}_{\{D^i=0\}}}{\sum_i \mathbf{1}_{\{F_1^i=1\}}\mathbf{1}_{\{L_2^i>0\}}\mathbf{1}_{\{S^i\in m\}}\mathbf{1}_{\{F_2^i=0\}}\mathbf{1}_{\{D^i=0\}}}.$$

Table 4 states the values of these estimates for our data set including information about the relevant number of data points.

Patient	# data	$\mu(\mathrm{LOS}^0_{\mathrm{full}})$	$\sigma(\mathrm{LOS}^0_{\mathrm{full}})$	P(R|Full)	[95% CI]	# data	$\mu(\mathrm{LOS}^{\mathrm{R}}_{\mathrm{full}})$	$\sigma(\mathrm{LOS}^{\mathrm{R}}_{\mathrm{full}})$
Type	points	(hours)				points	(hours)	
1	349	54.3	138.9	.086	[.057,.115]	9	61.4	71.8
2	395	51.7	54.1	.109	[.079,.140]	16	112.0	200.2
3	400	59.4	79.3	.120	[.089,.151]	17	99.6	86.8
4	363	62.8	68.1	.136	[.102,.170]	17	175.7	375.1
5	389	92.7	138.2	.132	[.100,.164]	17	237.1	577.8

Table 4 Demand-driven discharged patient parameters: operational parameters when patients are discharged in order to accommodate new patients. Average initial length-of-stay ($\mathrm{LOS}^0_{\mathrm{full}}$), readmission probability P(R|Full) and readmission length-of-stay ($\mathrm{LOS}^{\mathrm{R}}_{\mathrm{full}}$) when discharged from a 'full' ICU. Length-of-stay is given in hours.

Contrasting the results in Tables 3 and 4 we see that patients discharged during times of heavy ICU utilization *are markedly more likely to be readmitted, all else being the same.* In the following section, we will use the estimates we have computed here to construct and simulate the clinically relevant policies discussed in the previous Section.

6. Performance Evaluation

The goal of this Section is to explicitly construct the clinically relevant policies discussed in Section 4 using the estimates of the previous Section. For each of the policies we construct, we will primarily be interested in characterizing two things:

Mortality: This is a first order measure of the clinical impact of any demand-driven discharge practice. Given our discussion in Section 4, one would hope that any of the clinically relevant discharge policies considered there results in effectively equivalent mortality rates. If this were not the case, this would be cause to question the clinical viability of the policies.

Measures of Access: *Assuming* that two given policies possess similar mortality rates, one may be concerned about finer grained measures of performance. An important issue raised in Section 4 – and indeed a focus of this paper and recent healthcare reform –was that of access. It is crucial that the demand-driven discharge policies employed ensure *equitable* and *maximal* access to ICU resources while of course, ensuring no sacrifice in terms of mortality rates. In fact, it is entirely within reason *that these two goals are aligned as opposed to being at odds with each other.*

We next specify each of the policies discussed qualitatively in Section 4:

Mortality Risk 'P(D)': Under this policy, if a demand-driven discharge is called for, one selects a patient from the class with the smallest probability of death, P(D), of the patients currently in the ICU. Table 2 calibrates these figures for patients in our data set. This translates to the order $1, 2, 3, 4, 5$.

Readmission Risk I 'P(R)': Under this policy, one selects a patient from the class with the smallest nominal probability of readmission, P(R), of the patients currently in the ICU. In particular, given the estimates from our data set reported in Table 3, this translates to the order $1, 2, 3, 4, 5$.

Readmission Risk II 'Δ-Load': This policy, which as discussed earlier, is a refinement of the readmission risk metric above, has been a focal point of our study. We can estimate the increase in readmission load for a given patient class, m, as the quantity

$$\mathrm{P(R|Full)}_m \mu(\mathrm{LOS}^R_{\mathrm{full}})_m - \mathrm{P(R|Low)}_m \mu(\mathrm{LOS}^R_{\mathrm{low}})_m.$$

Using the data from Tables 3 and 4, this translates to the priority order $3, 1, 2, 4, 5$.

Remaining Length-of-stay 'LOS': Under this policy, one selects a patient from that class with the smallest *remaining* length-of-stay. As such this is not a static index rule. In particular, one needs to compute, for a patient of class m that has been in the ICU for time t, the quantity $\mathbb{E}[\text{LOS}^0_{\text{low}}|\text{LOS}^0_{\text{low}} \geq t]$, and prioritize patients in increasing order of this quantity. In our simulations, we give this policy more power and assume the realization for ICU LOS is known as soon as a patient begins ICU care. This policy is analyzed in Dobson et al. (2010) albeit for a model that is agnostic to readmission loads.

Table 6 summarizes the first three policies. It is interesting to note that of the first three policies, all three policies choose to protect patients of types 4 and 5 from a demand-driven discharge. These are patients with relatively higher mortality risk, and as such this is a desirable feature. Interestingly, the Δ-Load policy differs from the first two in how it prioritizes the first three patient classes which have low mortality risk. This allows for the following interpretation of the Δ-Load policy – it ensures that patients with high mortality risk are the least likely to be subject to a demand-driven discharge while carefully prioritizing among patients with low mortality risk to account not only for the likelihood they would have to be readmitted as a consequence of the discharge, but also the extent of the care they might require if such a readmission were to occur.

Patient Type	Nominal P(D)	Nominal P(R)	Δ-Readmission Load (hours)
1	.005	.073	2.65
2	.022	.095	5.94
3	.059	.102	1.05
4	.079	.115	11.19
5	.167	.119	12.09

Table 5 Estimated Policies

We consider the following simulation setup: We assume 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 $1,000$ sample paths each. We use the parameters estimated in Table 6 for nominal length-of-stay, probability of death, probability of readmission, and change in expected readmission load. A patient's nominal length-of-stay is log-normally distributed. We vary the probability of an arrival, λ between 0 and 0.021 (i.e. between 0 and 5 arrivals on average every 24 hours). An arrival rate $\lambda = .021$ corresponds to 5 patient per day, i.e. a turnover of 1/2 the beds in the ICU each day which is about the highest load seen in the ICU. We use a uniform traffic mix across patient classes, which is consistent with the empirical data. We now report on the two issues we set out to examine, namely mortality and patient access.

6.1. Mortality Rates

We compare the number of deaths per week under the various discharge policies. We consider an arrival rate of 2.5 patients per day, which corresponds to the load an average hospital could expect to a 10 bed ICU. In column (a) of Table 6 we compare the number of deaths per week using the point estimates of P(D|Full) and P(D|Low) given in Table 2. We also consider the following robustness check using the confidence intervals computed for our class specific mortality rate estimates: we consider that the various probabilities (namely, P(D|Full)$_m$ and P(D|Low)$_m$) each take on one of their upper or lower confidence limits, and consider all the 2^{10} resulting parameter combinations. We conduct a separate simulation for each of these parameter combinations, and report for each discharge policy the lowest and highest mortality rates across parameter combinations. The results are summarized in Table 6. We can see that using both the point estimates, as well as under our robustness check, all three policies are remarkably similar.

Policy	(a) # Deaths	(b) Min. # Deaths	(c) Max. # Deaths
Δ-Readmission Load	1.014	0.751	1.325
P(Death) & P(Readmission)	1.004	0.764	1.332
Shortest Remaining LOS	1.022	0.740	1.303

Table 6 Weekly Mortality Rate using (a) point estimates (b) the combination over the 95% confidence intervals with the lowest rate and (c) the combination over the 95% confidence intervals with the highest rate.

We next consider a further robustness check assuming that, in fact, the probability of death upon being demand-driven discharged is substantially increased (beyond the value estimated in the data) – we set the probability of death for a demand-driven discharged patient $10\%, 20\%, 30\%, 40\%$, and 50% higher than the estimated probability of death for that patient class given in Table 2. We compare the relative increase (decrease) in the number of deaths compared to the proposed Δ-Readmission Load policy. Table 7 summarizes these results. Again, the table reveals that the three policies continue to remain essentially identical across the range of perturbations with no single policy dominating.

From these experiments, we conclude that in as much as mortality rates are concerned all four policies are viable and result in essentially identical mortality rates. In spite of the fact that the

Inflation Factor	Shortest Remaining LOS	P(Death) & P(Readmission)
0	-0.9%	0.9%
10%	0.0%	0.5%
20%	0.8%	0.1%
30%	1.5%	-0.4%
40%	2.2%	-0.7%
50%	2.6%	-1.2%

Table 7 Percentage increase over Δ-Readmission Load policy in weekly mortality rate when artificially inflating P(Death|Full).

policies differ from each other, this reaffirms our earlier assertion that all four of the policies will protect patients with high mortality rates from a demand-driven discharge.

6.2. Patient Access

We measure access via the following proxy – since demand-driven discharges result in an increase in the expected critical care requirements for the discharged patient down the road, we measure the expected increase in these requirements, measured in hours of ICU care. In particular, we measure the expected increase in readmission load incurred due to demand-driven discharges under all four policies. Figure 2 shows the expected increased readmission load in hours for the four discharge policies. We can see that the proposed Δ-Load policy outperforms each of the benchmarks – in some cases by nearly 30%. The next best policy in this regard is the one based on (unadjusted) readmission and mortality risks, i.e. the P(R) & P(D) index policy. Thus, although the problem of minimizing readmission load due to required demand-driven discharges is a hard one, the proposed Δ-Load policy 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 is that for 5 patient arrivals per day, the Δ-Load policy incurs readmission load that is 13.5 hours lower than the next best policy (the P(D) & P(R) 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. Over the course of a year this corresponds to a free ICU bed for nearly a whole month. The savings relative to the LOS index policy are higher. Finally, in light of our study on mortality rates, these difference in performance do not come at the cost of increased mortality.

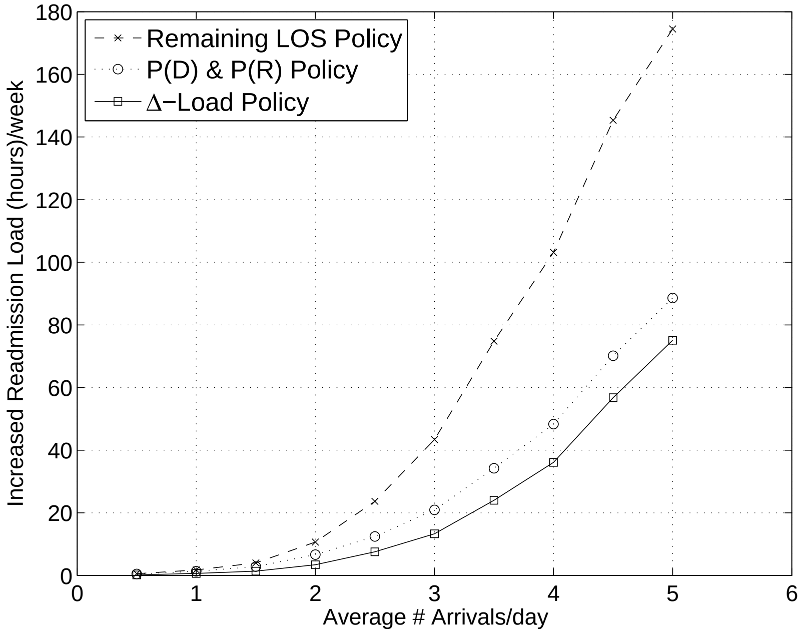

Figure 2 Performance of proposed index policy compared to benchmarks for various arrival rates and distribution across patient types according to the proportions seen in the empirical data.

In conclusion, we have observed the following:

Mortality: All four policies we considered at the outset in Section 4 result in essentially identical mortality rates. We have verified this fact across multiple 'robustness' checks. We attribute this to an attractive feature common to the first three policies, namely the fact that patients with high mortality risk are protected from a demand-driven discharge.

Access: In terms of access (or equivalently, increase in ICU load incurred due to demand-driven discharges) the policies are quite dissimilar. The Δ-Load policy (that has been a focal point in this paper) provides the greatest access. We attribute this to the fact that the policy carefully prioritizes among patients with low mortality risk.

As such, we believe that the Δ-Load policy might serve as a useful guide to intensivists prioritizing demand-driven discharge decisions among patients medically fit for discharge.

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 cost associated with a demand-driven patient discharge. We examine a number of clinically relevant cost metrics including mortality and readmission risks. We focus on a measure of readmission risk which incorporates the likelihood of readmission in addition to the complexity of the readmission: change in readmission load. 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 without sacrificing mortality 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

Allon, G., S. Deo, W. Lin. 2009. Impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Working Paper, Northwestern University, Kellogg Graduate School of Management*.

American Thoracic Society. 1997. Fair allocation of intensive care unit resources. American Journal Respiratory Critical Care Medicine.

Bone, R. C., N. E. McElwee, D. H. Eubanks, E. H. Gluck. 1993. Analysis of indications for intensive care unit admission. clinical efficacy assessment project: American college of physicians. *CHEST* **104** 1806–1811.

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

Chalfin, D. B., S. Trzeciak, A. Likourezos, B. M. Baumann, R. P. Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* **35** 1477–1483.

Chan, C. W., V. F. Farias. 2009. Stochastic depletion problems: Effective myopic policies for a class of dynamic optimization problems. *Mathematics of Operations Research* **34**(2) 333–350.

Chen, L. M., C. M. Martin, S. P. Keenan, W. J. Sibbald. 1998. Patients readmitted to the intensive care unit during the same hospitalization: clinical features and outcomes. *Critical Care Medicine* **26** 1834–1841.

Chrusch, C. A., K. P. Olafson, P. M. McMillan, D. E. Roberts, P. R. Gray. 2009. High occupancy increases the risk of early death or readmission after transfer from intensive care. *Critical Care Medicine* **37** 2753–2758.

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

Dobson, G., H.-H. Lee, E. Pinker. 2010. A Model of ICU Bumping. *Operations Research* **58** 1564–1576.

Durbin, C.G., R.F. Kopel. 1993. A case-control study of patients readmitted to the intensive care unit. *Critical Care Medicine* **21** 1547–1553.

Durrett, R. 1996. *Probability: Theory and Examples*. Duxbury Press.

Escobar, G. J., J. D. Greene, P. Scheirer, M. N. Gardner, D. Draper, P. Kipnis. 2008. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Medical Care* **46** 232–239.

Franklin, C., D. Jackson. 1983. Discharge decision-making in a medical ICU: Characteristics of unexpected readmissions. *Critical Care Medicine* **11** 61–66.

Franklin, C., E. C. Rackow, B. Mamdani, G. Burke, M. H. Weil. 1990. Triage considerations in medical intensive care. *Arch Intern Med* **150** 1455–1459.

Gajic, O., M. Malinchoc, T. B. Comfere, M. R¿ Harris, A. Achouiti, M. Yilmaz, M. J. Schultz, R. D. Hubmayr, B. Afessa, J. C. Farmer. 2008. The Stability and Workload Index for transfer score predicts unplanned intensive care unit patient readmission: Initial development and validation. *Crit Care Med* **36** 676–682.

Green, L. V. 2003. How many hospital beds? *Inquiry* **39** 400–412.

Green, L. V. 2006. *Queueing Analysis in Healthcare*, chap. Patient Flow: Reducing Delay in Healthcare Delivery. Springer, New York, N.Y.

Green, L. V., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Operations Research* **56** 1526–1538.

Green, L. V., S. Savin, B. Wang. 2003. Managing patient service in a diagnostic medical facility. *Operations Research* **54** 11–25.

Green, L. V., J. Soares, J. F. Giglio, R. A. Green. 2006. Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13** 61–68.

Halpern, N. A., S. M. Pastores. 2010. Critical care medicine in the united states 2000-2005: An analysis of bed numbers, occupancy rates, payer mix, and costs. *Critical Care Medicine* **38** 65–71.

Huang, X. A. 1995. A planning model for requirement of emergency beds. *Journal of Mathematics Applied in Medicine Biology* **12** 345–353.

Iwashyna, T. J., A. A. Kramer, J. M. Kahn. 2000. Intensive care unit occupancy and patient outcomes. *Critical Care Medicine* **37** 1545–1557.

Kc, D., C. Terwiesch. 2011. An econometric analysis of patient flows in the cardiac ICU. *MSOM* to appear.

Kwak, N., C. Lee. 1997. A linear programming model for human resource allocation in a health-care organization. *Journal of Medical Systems* **21** 129–140.

Loynes, R.M. 1963. The stability of a queue with non-independent interarrival and service times. *Proceedings of the Cambridge Philisophical Society* **58** 497–530.

McConnell, K. J., C. F. Richards, M. Daya, S. L. Bernell, C. C. Weathers, R. A. Lowe. 2005. Effect of increased icu capacity on emergency department length of stay and ambulance diversion. *Annals of Emergency Medicine* **45** 471–478.

McConnell, K. J., C. F. Richards, M. Daya, C. C. Weathers, R. A. Lowe. 2006. Ambulance diversion and lost hospital revenues. *Annals of Emergency Medicine* **48** 702–710.

Melnick, G. A., A. C. Nawathe, A. Bamezai, L. Green. 2004. Emergency department capacity and access in california 1990-2001: An economic analysis. *Health Affairs* **23**.

Merrill, C. T., A. Elixhauser. 2005. Hospitalization in the United States, 2002: HCUP fact book no. 6. Rockville, MD. *Agency for Healthcare Research and Quality* .

Metcalfe, M. A., A. Sloggett, K. McPherson. 1997. Mortality among appropriately referred patients refused admission to intensive-care units. *Lancet* **350** 7–11.

Mitchell, I., M. Grounds, D. Bennett. 1995. Intensive care in the ailing UK health care system. *Lancet* **345** 652.

Moreno, R.P., P. G. Metnitz, E. Almeida, B. Jordan, P. Bauer, R.A. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, J.R. Le Gall. 2005. SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* **31** 1345–1355.

Murray, M., M. Davies, B. Boushon. 2007. Panel size: how many patients can one doctor manage? *Family Practice Management* **14** 44–51.

Naylor, M. D., D. A. Brooten, R. L. Campbell, G. Maislin, K. M. McCauley, J. S. Schwartz. 2004. Transitional Care of Older Adults Hospitalized with Heart Failure: A Randomized, Controlled Trial. *Journal of the American Geriatrics Society* **52**(5) 675–684.

Norton, S.A., L.A. Hogan, R.G. Holloway, H. Temkin-Greener, M.J. Buckley, T.E. Quill. 2007. Proactive palliative care in the medical intensive care unit: effects on length of stay for selected high-risk patients. *Crit Care Med* **35** 1530–1535.

Patient Protection and Affordable Care Act. 2010. Hospital readmissions reduction program. Sec. 3025.

Rapoport, J., D. Teres, S. Lemeshow. 1996. Resource use implications of do not resuscitate orders for intensive care unit patients. *Am J Respir Crit Care Med* **153** 185–190.

Renaud, B., A. Santin, E. Coma, N. Camus, D. Van Pelt, J. Hayon, M. Gurgui, E. Roupie, J. Hervé, M.J. Fine, C. Brun-Buisson, J. Labarère. 2009. Association between timing of intensive care unit admission and outcomes for emergency department patients with community-acquired pneumonia. *Critical Care Medicine* **37**(11) 2867–2874.

Rincon, F., S.A. Mayer, J. Rivolta, J. Stillman, B. Boden-Albala, M.S V. Elkind, R. Marshall, J.Y. Chong. 2010. Impact of Delayed Transfer of Critically Ill Stroke Patients from the Emergency Department to the Neuro-ICU. *Neurocritical Care* **13** 75–81.

Rosenberg, A. L., C. Watts. 2000. Patients readmitted to ICUs: a systematic review of risk factors and outcomes. *Chest* **118** 492–502.

Scheulen, J. J., G. Li, G. D. Kelen. 2001. Impact of ambulance diversion policies in urban, suburban, and rural areas of central Maryland. *Academic Emergency Medicine* **8** 1553–2712.

Schull, M. J., M. Vermuelen, G. Slaughter, L. Morrison, P. Daly. 2004. Emergency department crowding and thrombolysis delays in acute myocardial infarction. *Annals of Emergency Medicine* **44** 577–585.

Shmueli, A., C. L. Sprung, E. H. Kaplan. 2003. Optimizing admissions to an intensive care unit. *Health Care Management Science* **6**(3) 131–136.

Smirnov, N. 1939. Estimating the deviation between the empirical distribution functions of two independent samples. *Moscow University Mathematics Bulletin* **2** 3–16.

Smith, G. B., B. L. Taylor, P. J. McQuillan, E. Nials. 1995. Rationing intensive care. Intensive care provision varies widely in Britain. *BMJ* **310** 1412–1413.

Snow, N., K.T. Bergin, T.P Horrigan. 1985. Readmission of patients to the surgical intensive care unit: Patient profiles and possibilities for prevention. *Critical Care Medicine* **13** 961–985.

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

Task Force of the American College of Critical Care Medicine, Society of Critical Care Medicine. 1999. Guidelines for intensive care unit admission, discharge, and triage. *Crit Care Med* **27** 633–638.

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

Yoon, K. B., S. O. Koh, D. W. Han, O. C. Kang. 2004. Discharge decision-making by intensivists on readmission to the intensive care unit. *Yonsei Med J* **45** 193–198.

Zimmerman, J. E., A. A. Kramer, D.S. McNair, F. M. Malila. 2006. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med* **34** 1297–1310.

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 $\epsilon > 0$:

$$\begin{array}{lll} \text{For } i=1: & \mu_1^0 = 1/2, & \phi_1 = 1 \\ \text{For } i=2: & \mu_2^0 = 1, & \phi_2 = 1-\epsilon \end{array}$$

Therefore, patient type 1 has nominal expected length-of-stay of 2 and cost of 1. Similarly, patient type 2 has nominal expected length-of-stay of 1 and cost 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 cost 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 when the cost metric of interest is the Δ-Readmission Load associated with a demand-driven discharge. We consider a stylized model of the ICU which accounts for patient readmissions. Patients who are naturally discharged require ICU readmission with probability 0. Patients who are demand-driven discharged are readmitted to the ICU with probability p_m and have readmission LOS which is exponentially distributed with mean $1/\mu_m^R$. Hence, the cost associated with a demand-driven discharge of patient type m, is the Δ-Readmission Load:

$$C(s,m)=\frac{p_m}{\mu_m^R}$$

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 π 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 π as *stabilizing* for λ 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 nth and $(n+1)$st entry to the readmission queue, and by S_n, the service time required by the nth patient. Assume moreover that no demand-driven discharges occur in the absence of a need for one, i.e. $\pi(s) = 0$ if $s \notin \{(x,y) : \sum_m x(s)_m + y(s)_m = B+1, \sum_m y(s)_m = 1\} \triangleq \hat{\mathcal{S}}_{\text{full}}$ (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 nth and $(n+1)$st visit to a state in the set $\hat{\mathcal{S}}_{\text{full}}$ while S_n is a Geometric $(\mu^R_{\pi(s_n)})$ 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 nth 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 π. 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 κ^π 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 λ 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 π^*_λ under which the readmission queue is stable when the arrival rate vector is λ, and moreover there exists an $\epsilon > 0$ such that the arrival rate vector $\lambda(1+\epsilon)$ can also be stabilized. Let us denote by $\pi^*_{\alpha\lambda}$ a policy minimizing κ^π 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 $\kappa^{\hat{\pi}_{\alpha\lambda}}/\kappa^{\pi^*_{\alpha\lambda}} \leq 1/\alpha$. 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 $\pi^*_{\alpha\lambda}$ (respectively π^*_λ) a policy minimizing κ^π in a system with arrival rate vector $\alpha\lambda$ (respectively λ). 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 λ. To this system we apply policy π^*_λ. Now by construction, a discharge under this policy will correspond to the discharge of an actual patient with probability α; 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 $\alpha\kappa^{\pi^*_\lambda}$. Moreover, since it is sub-optimal for the arrival rate vector $\alpha\lambda$, it must be that

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

It follows that

$$\kappa^{\hat{\pi}_{\alpha\lambda}} \leq (1/\alpha)\kappa^{\pi^*_{\alpha\lambda}} \leq \kappa^{\pi^*_\lambda}.$$

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

We have demonstrated a stationary policy π^g satisfying, for a given arrival rate vector λ, $\kappa^{\pi^g}/\kappa^{\pi^*_\lambda} \leq 1/(1+\hat{\rho})$ where $\hat{\rho}$ was a measure of utilization. It follows that should the readmission queue be *unstable* under π^g, then it will remain unstable for any arrival rate vector that strictly dominates $(1+\hat{\rho})\lambda$ under *any* stationary discharge policy. In other words, the use of the π^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

Proof of Theorem 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 π^*, $\pi^*(s) \notin \arg\min_{m:x(s)_m>0} \phi_m$, i.e. the patient selected for the demand-driven discharge under any optimal policy is not among the set of patient types that minimizes one-period costs at state s. For notational convenience, we take $\pi^*(s) = j$, and $i = \pi^g(s) \in \arg\min_{m:x(s)_m>0} \phi_m$. Thus, by assumption we have that

$$J^*(s) = C(s,j) + E\left[J^*(S(s,j))\right] < C(s,i) + E\left[J^*(S(s,i))\right]. \tag{C1}$$

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

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

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

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

Cases (i), (ii), and (iii) together yield $E[J^*(S(s,i)) - J^*(S(s,j))] \leq C(s,j) - C(s,i)$ which contradicts (C1) (since $C(s,i) \neq C(s,j)$) and yields our result. □

Proof of Lemma 1: 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 tth epoch are coupled as follows: If $z(s_t)_m \geq z(s'_t)_m$, then the number of patients of type m that finish service and depart naturally from the s' system is given by the Binomial-$(z(s'_t)_m, \mu_m^0)$ random variable $X'_{t,m}$ while the number of patients of type m that finish service and depart naturally from the s system is given by $X'_{t,m} + Z_{t,m}$ where $Z_{t,m}$ is a Binomial-$(z(s_t)_m - z(s'_t)_m, \mu_m^0)$ random variable. A symmetric situation must hold if $z(s'_t)_m \geq z(s_t)_m$.

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

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

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

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

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

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

$$\begin{aligned}
x(\tilde{s})_{\pi^*(s)} &= x(s)_{\pi^*(s)} + y(s)_{\pi^*(s)} - 1 - X_{t(s),\pi^*(s)}, \\
x(\tilde{s})_{\pi^g(s)} &= x(s)_{\pi^g(s)} + y(s)_{\pi^g(s)} - 1 - X_{t(s),\pi^g(s)}, \\
x(\tilde{s})_m &= x(s)_m + y(s)_m - X_{t(s),m}, \quad m \neq \pi^*(s), \pi^g(s) \\
y(\tilde{s})_m &= Y_{t(s)+1,m}, \\
t(\tilde{s}) &= t(s) + 1,
\end{aligned}$$

where analogous to our earlier description of $S(s,a)$, we define $X_{t(s),\pi^*(s)}$ (resp. $X_{t(s),\pi^g(s)}$) as a Binomial $(x(s)_{\pi^*(s)} + y(s)_{\pi^*(s)} - 1, \mu^0_{\pi^8(s)})$ (resp. Binomial $(x(s)_{\pi^g(s)} + y(s)_{\pi^g(s)} - 1, \mu^0_{\pi^g(s)})$) random variable. For $m \neq \pi^*(s), \pi^g(s)$, 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 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 π' 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. τ 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 $x(s_\tau)_{\pi^*(s)} = x(s)_{\pi^*(s)} + y_{t(s),\pi^*(s)} - 1$, π' simply takes the optimal action for $t \geq \tau$ (so that, in fact π' coincides with π^* on this event). On the event that $x(s_\tau)_{\pi^*(s)} = x(x)_{\pi^*(s)} + y_{t(s),\pi^*(s)}$, $\pi'(s_\tau) = \pi^*(s)$, and π' takes actions according to the optimal policy π^* for $t > \tau$. The probability that an eviction occurs under π' at τ 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 $S(s,\pi^g(s))$ and $\tilde{S}(s,\pi^*(s))$ so that under the π' 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

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

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