Critical care in hospitals: When to introduce a Step Down Unit?

Mor Armony
Stern School of Business, New York University marmony@stern.nyu.edu

Carri W. Chan
Decision, Risk, and Operations, Columbia Business School cwchan@columbia.edu

Bo Zhu
Courant Institute of Mathematical Sciences, New York University zhubo@cims.nyu.edu

This version: February 22, 2014

Step Down Units (SDUs) provide an intermediate level of care between the Intensive Care Units (ICUs) and the general medical-surgical wards. Because SDUs are less richly staffed than ICUs, they are less costly to operate; however, they also are unable to provide the level of care required by the sickest patients. There is an ongoing debate in the medical community as to whether and how SDUs should be used. On one hand, an SDU alleviates ICU congestion by providing a safe environment for post-ICU patients before they are stable enough to be transferred to the general wards. On the other hand, an SDU can take capacity away from the already over-congested ICU. In this work, we propose a queueing model to capture the dynamics of patient flows through the ICU and SDU in order to determine how to size the ICU and SDU. We account for the fact that patients may abandon if they have to wait too long for a bed, while others may get bumped out of a bed if a new patient is more critical. Using fluid and diffusion analysis, we examine the tradeoff between reserving capacity in the ICU for the most critical patients versus gaining additional capacity achieved by allocating nurses to the SDUs due to the lower staffing requirement. Despite the complex patient flow dynamics, we leverage a state-space collapse result in our diffusion analysis to establish the optimal allocation of nurses to units. We find that under some circumstances the optimal size of the SDU is zero, while in other cases, having a sizable SDU may be beneficial. The insights from our work will be useful for hospital managers determining how to allocate nurses to the hospital units, which subsequently determines the size of each unit.

Key words: Healthcare, queueing, fluid analysis, diffusion analysis, state-space collapse

1. Introduction

Step Down Units (SDUs) provide an intermediate level of care between the Intensive Care Units (ICUs) and the general medical-surgical wards. These units, which are also commonly referred to as intermediate care units and transitional care units[1], are found in many, but not all, hospitals in developed nations. Typically, these units are staffed at a higher nurse to patient ratio than general medical-surgical wards but not as high as ICUs. ICUs care for the sickest patients and consume a disproportionate share of total health care costs (nearly $82 billion annually (Halpern and Pastores 2010), which amounts to 20-35% of total hospital costs

[1] In fact, there are over 10 different names these units can be named (Stacy 2011).

with ICU beds occupying only 5-10 percent of inpatient beds (Joint Commission Resources 2004)). Consequently, a voluminous literature in both the medical and operations communities exists that addresses the need to understand and improve how these units function (see, for example, Chalfin et al. (2007), Chan et al. (2012), Kc and Terwiesch (2012), Kim et al. (2012), Shmueli et al. (2003)). In contrast, very few studies address these issues with respect to SDUs, despite the fact that, in hospitals that have them, the SDU plays an important role in patient flow through the ICU.

The purpose of an SDU is to treat patients who are more severe than the typical ward patient, but who do not require as intense monitoring as the most critical ICU patients. In theory, having a unit that can both care for sicker patients and, at the same time, take pressure off the ICU should result in both better patient outcomes as well as increased efficiency. In practice, whether such promise is actually fulfilled is not known, as the effects of having an SDU are not fully understood. Some studies suggest that SDUs can improve outcomes of critical patients by increasing access to the ICU (Byrick et al. 1986, Zimmerman et al. 1995), while others suggest that all critical patients should be treated in the ICU rather than the SDU (Hanson et al. 1999, Simchen et al. 2004). Some studies have shown that an SDU can be cost effective (Harding 2009, Stacy 2011), while history suggests that there is not enough evidence to conclude this to be the case in general (Keenan et al. 1998). Our goal in this work is to better understand the role SDUs play in the treatment of critically ill patients.

Semi-critical patients who can be treated in the SDU can also be treated in the ICU without any impact on their quality of care. Conversely, due to the lower staffing requirements in the SDU, Critical patients who are treated in the SDU will not be able to receive the high level monitoring and care provided in the ICU, resulting in substantial degradation of their quality of care. Hence, not only do ICUs provide care for the sickest patients, they can also be considered ‘flexible servers’ in the sense that they can also treat moderately severe patients. However, largely due to the high nurse-to-patient ratio requirement, they are more costly to operate than SDUs. In California, an ICU is legally obligated to have at least one nurse for every two ICU patients; in practice, many hospitals operate with one nurse per patient. In contrast, SDUs can be staffed anywhere from one nurse per two to four patients. In particular, the SDU can accommodate more patients for the same number of nurses. This creates an interesting tradeoff between overall capacity gains (SDU) for all patient severities versus maintaining more capacity for the highest priority patients (ICU).

This work is motivated by a conversation with the chief intensivist at a large urban hospital. The hospital did not have SDU beds and was considering creating some by reducing capacity in the ICU. The main debate centered on how many SDU beds should be created without modifying the number of nursing staff on budget. The hospital did not want to increase the number of nursing staff on budget due to cost considerations–any physical changes would primarily have a one time occurrence (at the time of change),

but staffing costs would perpetuate long into the future. On the other hand, cutting the number of nursing staff on the payroll would hurt hospital morale and result in substantial backlash by hospital staff which would make it difficult to implement the new plan. The goal was to rotate the current ICU nurses between the ICU and new SDU, so that the main differentiation between the two units would be the nurse-to-patient ratio. Critical care nurses require additional training in order to have the skill set required to treat and monitor Critical patients. The decision to use critical-care nurses in the SDU was clinically strategic–management wanted to ensure that the nurses were capable of dealing with any complications which could arise in the unit. Other hospitals have also used critical-care nurses to staff the SDU (e.g. Eachempati et al. (2004)). The majority of the nurses in the SDU in Harding (2009) had at least 1 year of telemetry or critical care experience. While some hospitals (e.g. Aloe et al. (2009)) use medical-surgical nurses in their SDU, our primary focus will be on the hospitals which use critical care nurses in both the ICU and SDU.

We introduce a model where Critical patients arrive to the ICU and must wait for a bed if none is available. If the wait is too long, the patient will receive care elsewhere (in another unit/hospital) or, in the most extreme case, die due to the long wait–we refer to such events as patient 'abandonment'. A Critical patient who is admitted to the ICU will be treated until reaching a Semi-critical state where he can be treated in the SDU or stay in the ICU. Semi-critical patients can be bumped out of the ICU if a Critical patient requires a bed. Our objective is to determine the size of the SDU and ICU in order to minimize the costs associated with patient abandonment and bumping. We start by considering a budget neutral setting in which the total number of nurses for the two units as well as the required nurse-to-patient ratios are known. The limiting resource in our setting is nurses, rather than beds. Hence, our objective can be framed as allocating nursing staff between SDUs and ICUs. We will also consider an extension which relaxes the budget neutral constraint on nurses. There has been some work in operations management looking at staffing in healthcare (e.g. Green et al. (2006), de Véricourt and Jennings (2008), Yankovic and Green (2011), Green et al. (2013), Yom-Tov and Mandelbaum (2013)). Most of the prior work focuses on a single unit and has not considered the impact of the SDU. Due to the fact that hospital units (and particularly the ICUs) are often operated near or at capacity (Green 2003, Pronovost et al. 2004), we analyze our queueing system via fluid and diffusion approximations as they tend to be accurate for systems operated in heavy traffic.

Our main contributions can be summarized as follows:

- We provide justification for the highly varied use of SDUs observed in practice. In particular, we find there exist two operational regimes which depend on the relative costs between Critical patient abandonment (due to lack of available beds in the ICU) and Semi-critical patient bumping (off-placement to other, less desirable units). In one–the ICU driven (ID) regime–virtually all nurses are allocated to the ICU (so the SDU is very small or is of size zero), while in the other–the ICU and SDU driven (ISD) regime–a significant number of nurses are allocated to both units.

- We infer physicians' estimates for the relative costs of lack of access to care for Critical and Semi-Critical patients based on the analysis of a commonly used prioritization rule. Based on these estimates, we find easily verifiable sufficient conditions such that a hospital should operate in the ID regime. Thus, we provide a simple prescription for hospitals to verify whether it is optimal to not have an SDU.
- Our approach utilizes different methodological tools, including fluid and diffusion analysis, dynamic programming and simulation. Using these tools, we are able to garner insights about a complex system of ICU and SDU beds and Critical and Semi-critical patients. In particular, our analysis provides simple rules for the management of ICU and SDU beds. We find that the approximations generated from these approaches result in reasonably accurate sizing recommendations.

The rest of the paper proceeds as follows. We conclude this section with a brief overview of related literature. We present our model of the ICU and SDU in Section 2. In Section 3, we examine the nurse allocation problem using fluid analysis. We refine our analysis by considering the diffusion level in Section 4. In Section 5, we leverage existing medical literature to numerically explore the insights generated from our fluid and diffusion analysis. In Section 6, we consider a number of extensions to our original model. Finally, we conclude in Section 7.

1.1. Literature Review

While there exists an extensive body of literature in the medical community on ICUs–there are multiple journals, including *Critical Care* and *Intensive Care Medicine*, devoted to this topic–much less attention has been directed towards SDUs. One of the reasons for this is that patient severity scores tend to be limited to the ICU (e.g., the APACHE and SAPS scores (Moreno et al. 2005, Vincent and Moreno 2010)) or else to specific ailments (e.g., pneumonia (Fine et al. 1997, Charles et al. 2008)). These types of scores have been very useful in comparing the impact of different care interventions and pathways on heterogenous patients as they make risk-adjustment possible. Because of the lack of severity scores for SDU patients, risk-adjustment is not possible which limits the ability to examine the impact of SDUs on the outcomes of most patients. For instance, without the ability to do risk-adjustment for patients who are treated in the SDU versus no SDU, it is difficult to determine whether better outcomes are because of the SDU or because patients who were sent to the SDU happened to be less severe, with higher likelihood of good outcomes.

Despite these challenges, there has been some work looking at SDUs. The majority of this work has focused on the impact of SDUs on ICU care. Though there may not be a general consensus as to whether SDUs can be cost-effective for treating semi-critical patients (Keenan et al. 1998), there are a number of studies focused on either specific ailments or at individual institutions which suggest the presence of an SDU can benefit patients. For instance, having an SDU can reduce ICU LOS (Byrick et al. 1986); this is intuitive because patients do not have to reach as high a level of stability to be discharged from the ICU

to the SDU rather than the general medical/surgical floor. In a study of patients with Acute Myocardial Infarction (i.e. heart attacks), the presence of an SDU was shown to reduce cost by $1.5 million a year for the treatment of patients with moderate risk (Tosteson et al. 1996). It is also argued there that high risk patients should not be treated in the SDU .

In capturing the patient flow dynamics through an ICU and an SDU, we consider a modification to the commonly used N-model queueing system (see Figure 16 in Gans et al. (2003)). The N-model is a queueing model with two classes of customers and two server pools. One server pool is flexible, and can serve both customer classes. The other is dedicated to only one of those classes. The N-model arises in our case due to the fact that the ICU consists of flexible beds (servers), while the SDU does not. In our setting, once a Critical patient completes treatment (service) in the ICU, he may transition into a Semi-critical patient who can be treated in either the ICU or SDU. This patient flow dynamic introduces a feedback into our model, which is not captured by existing N-models. In various settings, a threshold priority policy for routing patients to the flexible servers (Bell and Williams 2001, Tezcan and Dai 2010, Ghamami and Ward 2012), and a generalized C-μ priority policy (Mandelbaum and Stolyar 2004, Dai and Tezcan 2008, Gurvich and Whitt 2009b) minimize costs for the N-model in heavy traffic asymptotic regimes. With the exception of Wallace and Whitt (2005) and Gurvich and Whitt (2010), in all of these works, prioritization and routing of customers is the primary concern. In contrast, in the hospital setting, routing is largely dictated by medical necessity, so we focus on the question of staffing and sizing of units while assuming that a prioritization and routing rule is given. We will also examine the implications of this commonly used rule.

There is a rich literature on flexibility in queueing systems (e.g. Green (1985), Hopp et al. (2004), Iravani et al. (2005), Ata and Van Mieghem (2009), Bassamboo et al. (2012), Tsitsiklis and Xu (2012)). An important aspect discussed in this literature is how to design the network topology (pairing, chaining, full flexibility, etc.). Another focus is quantifying how to split the resources between flexible and dedicated servers. For example, there has been a series of recent work which considers this question with respect to tandem systems (Andradottir et al. 2013, Zhang and Ayhan 2013, Kirkizlar et al. 2013). We find a second order effect in our system which falls under this second category as we determine how to allocate the nurses between the ICU (flexible) and the SDU (dedicated). While we also look at a tandem system, the flow patterns exhibit different dynamics, such as bumping, which arise in a hospital.

In developing an understanding of the hospital system, we utilize a number of analytic methods. To start, we examine the system using fluid analysis (e.g. Whitt (2006), Bassamboo and Randhawa (2010)), that uses law-of-large-number principles to evaluate cost terms that are of the order of the arrival rate. Next, we refine our analysis by using diffusion approximations as in Jagerman (1974), Garnett et al. (2002),

Mandelbaum and Zeltyn (2009), that leverage central-limit-theorem type results to evaluate fluctuations about the fluid limit that are of order square-root of the arrival rate. Through the diffusion analysis, we establish a state-space collapse result similar to Gurvich and Whitt (2009a), albeit for a very different queueing system. Using these methodologies, we are able to evaluate the average abandonment and bumping costs and optimize the size of the units to minimize these costs. In our asymptotic analysis we take formal fluid and diffusion limits of the nurse allocation problem and then analyze the corresponding fluid and diffusion optimization problems directly. Using simulations we demonstrate the efficacy of the asymptotic solutions for the original queueing system. This approach is similar to the one taken by Harrison and Zeevi (2004), Rubino and Ata (2009), Kostami and Ward (2009), Akan et al. (2012) and Ata et al. (2012).

2. Model

We consider a system with a fixed number of N nurses. These nurses are flexible in the sense that they can work in either the ICU or SDU. While not all hospitals use critical-care nurses to staff the SDU, many–such as that in Eachempati et al. (2004)–do. For safety reasons, a strict nurse-to-patient ratio must be maintained in each unit. Let r_I ($< r_S$) be the given number of patients each nurse can manage in the ICU (SDU). Our goal is to determine how to allocate nurses between the two units, which is analogous to determining the number of ICU and SDU beds, B_I and B_S. We consider *budget neutral* allocations of nurses, so that we must allocate up to N nurses on salary. No additional nurses can be hired. This means that

$$\frac{B_I}{r_I} + \frac{B_S}{r_S} \leq N \tag{1}$$

so that we allocate up to N nurses to the ICU and SDU while satisfying the nurse-to-patient ratios. We refer to any pair (B_I, B_S) of non-negative integers that satisfy (1) as a feasible bed (nurse) allocation. As critical-care is often a bottleneck in the hospital (Ryckman et al. 2009, Kc and Terwiesch 2012, Beck 2011), we will assume there is ample space in the general medical-surgical ward. This will also allow us to focus on the flow of critical and semi-critical patients.

Patients can be in one of two states: Critical or Semi-critical state. If a patient is in the Critical state, he *must* be treated in the ICU and his service time in the Critical state is exponentially distributed with rate μ_C. Upon completion of service, with probability p the patient becomes a Semi-critical patient; with probability $1-p$ he leaves the system, which can practically correspond to a number of different situations, such as the patient being transferred and treated in the ward, being discharged home, or dying. Semi-critical patients can be treated in the SDU or ICU. Regardless of the type of bed, the service time of a Semi-critical patient is exponentially distributed with rate μ_{SC}.

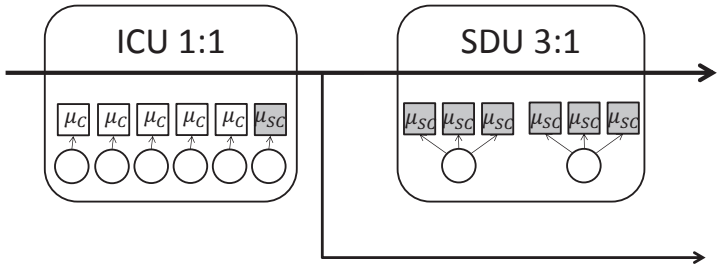


Figure 1 **System Model: Nurses are depicted as circles, patients are depicted at squares. Critical patients are served in the ICU. A Critical patient may become a Semi-critical patient upon finishing service in the ICU. Semi-critical patients are depicted in gray and are served in the SDU or ICU. One Semi-critical patient is currently being served in the ICU.**

See Figure 1 as an example of an allocation of nurses amongst the ICU and SDU. The nurse-to-patient ratio in the ICU is $r_I = 1$ and in the SDU it is $r_S = 3$. There are $N = 8$ nurses who are allocated to $B_I = 6$ ICU beds and $B_S = 6$ SDU beds.

New Critical patients arrive to the ICU according to a Poisson process with rate λ. If there is space in the ICU, the patient will begin service immediately. If there is no space in the ICU, he will wait in a queue. For instance, the patient could wait for ICU admission in the Emergency Department (ED). The abandonment rate from this queue is θ. An abandonment cost of w_C is incurred for each Critical patient who abandons from the system while waiting for an ICU bed. Abandonment can correspond to a patient being routed to another unit or hospital. Over the course of 2 years, 895 major surgeries which required ICU care were canceled or postponed and 487 patients were transferred to other hospitals due to access block issues in Melbourne, Australia (Duke et al. 2009). In another hospital, of the 381 patients referred for ICU admission, 16% were refused admission and were placed in another unit (Shmueli and Sprung 2005). This rate of denied ICU referrals was as high as 26% in Metcalfe et al. (1997). While these rates can vary across different hospitals and certainly depend on bed availability and patient mix, such off-placements–which we refer to as 'abandonments'–do occur with some regularity. Note that, for tractability, we use patient abandonment to capture the undesirability of making Critical patients wait for ICU service. Other adverse events of patient wait, such as an increase in LOS (Chan et al. 2013a), could also be considered.

If there is a Semi-critical patient in the ICU, he can be bumped out by an incoming Critical patient. If there is space for him in the SDU, this bumping comes at no cost. However, if there is no space in the SDU, this patient will be bumped to the general ward and a bumping cost w_{SC} is incurred, which captures the undesirability of this event. By definition, Critical patients are more critical than Semi-critical patients, so we assume that $w_C > w_{SC}$. Our queueing model is depicted in Figure 2.

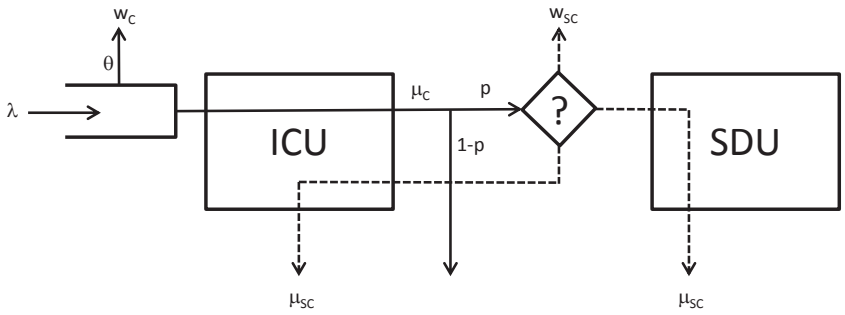


Figure 2 ICU-SDU queueing model: The '?' represents the assignment decision of a Semi-critical patient. Solid lines depict Critical patient flows while dotted lines depict Semi-critical patient flows.

Our objective is to minimize the long time average abandonment and bumping costs. Let a_n be an indicator that equals 1 if the n^{th} arriving patient, with arrival time t_n, *abandons* from the queue while waiting for ICU treatment. Similarly, let b_n be an indicator that equals 1 if the n^{th} arriving patient is *bumped* to the general ward. Note that a patient cannot be bumped if he departs the system without becoming a Semi-critical patient (either by abandoning or leaving after completing ICU service). Our objective is thus to determine the allocation of nurses to specify the number of ICU and SDU beds in order to minimize the following cost function:

$$\limsup_{T\to\infty} \frac{1}{T}\left[\sum_{n=0}^{\infty}\left[w_C a_n + w_{SC} b_n\right] 1_{\{t_n \le T\}}\right] \tag{2}$$

One could also consider including waiting costs for Critical patients in the problem formulation. By Little's law, in steady-state the expected waiting time is proportional to the expected queue length, which is in turn proportional to the abandonment rate. Thus, including linear waiting costs would simply amount to re-calibrating the parameter w_C.

In this work, we examine a stylized model of the ICU and SDU. Byrick et al. (1986) found that having an SDU can reduce ICU LOS–this reduction is captured by our service requirements of Critical and Semi-critical patients. With an SDU, the mean LOS of a patient in the ICU will be $1/\mu_C$ plus some additional time depending on if there is space in the ICU to treat him while in the Semi-critical state. However, without an SDU, more Semi-critical patients will be treated in the ICU, thus increasing overall ICU LOS. While there are some elements our model does not capture, such as external arrivals to the SDU, readmissions, or treatment of Critical patients in the SDU, it does capture the essence of the tradeoff between increasing capacity for all patient severities versus maximizing capacity for the most vulnerable patients. For tractability, we focus on the patient flows described in this section and find that, in doing so, we can gain many insights into the role of the SDU. In Section 6, we will consider some extensions to this initial model.

2.1. Cost parameters

It is reasonable to assume the optimal policy will depend on the per-patient cost of abandonment, w_C, and bumping w_{SC}. Our formulation allows for *any* quality metric–it could capture clinical costs such as the net decrease in quality-adjusted life years (QALYs) or financial costs such as loss in revenue due to not treating a patient in the ICU. We now discuss a number of clinically relevant costs, which hospitals are likely to consider when making decisions surrounding ICUs and SDUs.

Mortality Risk: A natural cost metric is mortality. Specifically, there is some risk of death associated with each patient, even if the patient follows the 'desired' care pathway. However, if a Critical patient is unable to get ICU care–and abandons the queue–or a Semi-critical patient is bumped out of the SDU, then it is reasonable to consider how this may impact the patient's nominal mortality risk. In such a case, w_C and w_{SC} could capture the *increase in mortality risk* due to abandonment or bumping, respectively. Then, solving the optimization problem in (2) would correspond to selecting the ICU and SDU sizes which would minimize the mortality rate of Critical and Semi-critical patients. In some practical settings, this cost metric may be too crude to be of value as access to care is typically granted for patients whose mortality risk would be significantly increased. Thus, we also consider other clinical measures of interest.

Readmission Risk: Another measure the medical community has focused on is patient readmissions, and more specifically, the probability of readmission. This cost metric has clear clinical implications as readmitted patients tend to be worse off (Durbin and Kopel 1993). It also has operational implications as readmitted patients will utilize ICU and SDU beds, which could have been used for new patients.

Readmission Load: Related to the measure of readmission risk is readmission load: the expected number of readmission bed hours due to abandonment or bumping. This measure was considered in Chan et al. (2012) and accounts for both the risk of readmission as well as the complexity of the readmission as measured by the typical length of stay upon readmission. In Section 6 we will examine a model which explicitly incorporates patient readmissions. Moreover, we will make a precise connection between minimizing readmission load and maximizing throughput.

For each of these clinical measures, the cost parameters w_C and w_{SC} would correspond to the increase in mortality risk, readmission risk or readmission load due to abandonment or bumping. Then, solving the optimization problem in (2) would correspond to selecting the ICU and SDU sizes which would minimize the number of corresponding adverse patient outcomes. While a hospital administrator may wish to focus on one clinical outcome, one could also consider a weighted sum and/or other potential cost measures.

2.1.1. Estimating cost parameters Thus far, there has been limited work quantifying the aforementioned cost metrics in the SDU setting. However, there has been recent work doing so in the ICU setting. For instance, Kc and Terwiesch (2012) estimates the costs of bumping patients from the cardiac ICU in

terms of readmission risk, while Kim et al. (2012) estimates the impact of rerouting patients waiting for ICU admission. Similar methodologies can be used as more data surrounding SDUs becomes available.

2.1.2. Inferring relative costs While providing quantitative estimates of the costs of abandonment and bumping will likely be possible in the future, there currently do not exist such measures. However, a general consensus amongst physicians we consulted with is that Critical patients are more unstable than Semi-critical patients. Hence, it is typically more costly for a Critical patient to abandon (i.e. he will get neither ICU or SDU level care) than to bump a Semi-critical patient. We leverage this observation to try to extract the relative cost ratio: w_C/w_{SC}.

Consider the optimization problem of choosing a priority and routing rule so as to minimize the expected average abandonment and bumping costs. Our goal is to derive properties of w_C/w_{SC} given the assumption that the optimal control policy is to give strict priority to Critical patients. Hence, if there are ever any Critical patients waiting to be admitted to the ICU and there are Semi-critical patients in the ICU, the Semi-critical patients will be sent to the SDU or be bumped to the general ward if it is full. We find the following relationship for the weights:

Proposition 1*If it is optimal to always give priority of ICU beds to Critical patients, i.e. one should bump Semi-critical patients from the ICU if a new Critical patient arrives, and the ICU is full, then the per-patient costs of abandonment,* w_C*, and bumping,* w_{SC}*, satisfy the following inequality:*

$$\frac{w_C}{w_{SC}} \geq \frac{\theta - \mu_C(1-p) + \mu_{SC}}{\max\{\mu_C, 2\theta - \mu_C\}} \tag{3}$$

All proofs can be found in the appendix. In what follows, we will assume that strict priority is given to Critical patients, so that a Semi-critical patient will be bumped out of the ICU if a new Critical patient needs the bed. We will examine further what this relationship implies about the optimal allocation of nurses.

2.2. Motivating our asymptotic approach

In theory, one could calculate the steady-state distribution of abandonment and bumping given a fixed allocation of nurses to the ICU and SDU. Then, an exhaustive search would reveal the allocation that obtains the lowest cost. Unfortunately, the numerical approach provides little intuition for the general model as to the impact of various system parameters on the optimal solution. Exact analysis is also extremely difficult because while Critical patients follow an $M/M/B_I + M$ queueing model, the number of Semi-critical patients strongly depends on the number of Critical patients in a non-trivial way. The result is a 2-dimensional Markov chain with no known closed-form expression for the steady-state distribution. Hence, our goal is to develop such an understanding by considering different operational regimes of our ICU and SDU hospital system. The asymptotic regime we consider is one with many nurses. While the

average hospital has 15-40 ICU beds[2] (8-40 ICU nurses), we will see via numeric examples in Section 5 the asymptotic analysis can be quite accurate even with moderate number of nurses. In particular, we consider a sequence of systems indexed by the number of nurses N, with N and λ growing to ∞, while the rest of the parameters do not change. Our first-order analysis relies on fluid scaling which considers terms of the order of N. Our second-order analysis relies on diffusion scaling, in which we consider fluctuations of the order of $\sqrt{N}$.

3. Fluid Analysis

We begin our analysis via a fluid modeling approach. Because ICUs and SDUs are so expensive to operate, hospital administrators do not want to have many empty beds in these units. As a consequence, these units are often operated at or above capacity. With that in mind, we consider a system that is heavily loaded, even if all of the available nurses are optimally allocated between the ICU and the SDU. We let $1/\mu_T = \left(\frac{1}{\mu_C} + \frac{p}{\mu_{SC}}\right)$ be the mean amount of time a new patient should be treated while in the Critical and Semi-Critical states. Let B_I and B_S be arbitrary bed allocations to the ICU and SDU, respectively. Then the traffic intensity associated with Critical patients may be defined as

$$\rho_C(B_I, B_S) := \frac{\lambda}{B_I \mu_C},$$

and the overall traffic intensity associated with all patients (in Critical and Semi-Critical state) is

$$\rho_T(B_I, B_S) := \frac{\lambda}{(B_I + B_S)\mu_T}.$$

Our heavy traffic assumption is such that for all bed allocations at least one of these two traffic intensities is greater than 1. More formally, we say that the system is in heavy traffic if:

$$\min_{\left\{B_I \geq 0, B_S \geq 0, \frac{B_I}{r_I} + \frac{B_S}{r_S} \leq N\right\}} \left[\max\left\{\rho_C(B_I, B_S), \rho_T(B_I, B_S)\right\}\right] > 1, \tag{4}$$

which, after some algebra, may be shown to be equivalent to assuming that

$$\frac{\lambda(r_I \mu_C p + r_S \mu_{SC})}{N r_I r_S \mu_C \mu_{SC}} > 1. \tag{5}$$

Our asymptotic approach is to consider a sequence of systems indexed by the number of nurses N, in which both N and λ grow without bound, while the rest of the system parameters remain fixed. For notational compactness, we omit the indexing of λ by N. The following proposition justifies our definition of heavy traffic.

[2] In California the number of certified medical/surgical ICU beds ranges from 2 to 110, with an average size of 20 beds (State of California Office of Statewide Health Planning & Development 2010-2011).

Proposition 2 1. *If* $\limsup_{N\to\infty} \frac{\lambda(r_I\mu_C p + r_S\mu_{SC})}{N r_I r_S \mu_C \mu_{SC}} \leq 1$, *then there exists a feasible bed allocation such that the total cost of abandonment plus bumping is* $o(N)$[3].

2. *Otherwise, if* $\liminf_{N\to\infty} \frac{\lambda(r_I\mu_C p + r_S\mu_{SC})}{N r_I r_S \mu_C \mu_{SC}} > 1$, *then for any feasible bed allocation the total cost rate of abandonment plus bumping is at least* $\mathcal{O}(N)$[4].

Under the heavy traffic assumption, we wish to examine the optimal allocation of nurses given the abandonment and bumping cost parameters. To do this, we turn to fluid analysis. The fluid analysis is based on scaling the arrival rate and the number of beds and nurses by $1/N$ and ignoring quantities that are $o(N)$. This way, we can focus on the main drivers of nurse allocation. We begin by defining our fluid scaling. Let

$$\bar{\lambda} := \frac{\lambda}{N}, \quad b_i := \frac{B_i}{N}, \quad i = I, S,$$

and note that

$$\frac{b_I}{r_I} + \frac{b_S}{r_S} \leq 1. \tag{6}$$

Assume that the system is in heavy traffic. Then, given a nurse allocation, and assuming that Critical patients are given priority over Semi-critical patients, the fluid scaled abandonment rate is equal to $\left(\bar{\lambda} - b_I\mu_C\right)^+$. In particular, under optimal allocation, it is clear that $b_I\mu_C \leq \bar{\lambda}$. This is because further increasing the number of nurses allocated to the ICU increases the bumping costs without affecting the abandonment cost.

Now, since we assume that Critical patients are given priority over Semi-critical patients, in this regime, the ICU is always full with Critical patients. There will not be any Semi-critical patients in the ICU. Thus the fluid-scaled abandonment rate from the ICU is equal to the scaled ICU arrival rate minus its service rate: $\bar{\lambda} - b_I\mu_C$. Similarly, the fluid-scaled bumping rate from the SDU is equal to the scaled SDU arrival rate minus its service rate: $(b_I\mu_C p - b_S\mu_{SC})^+$, where we take the positive part, because patients can only be treated if they are available. Combining these two expressions together gives us the average abandonment and bumping cost.

Recognizing that constraint (6) is satisfied as an equality under the optimal allocation, we can specify our fluid objective in terms of b_I. Our goal is thus to determine, $0 \leq b_I \leq \left(r_I \wedge \frac{\bar{\lambda}}{\mu_C}\right)$ (where $\wedge$ is the minimum function) and $0 \leq b_S \leq r_S$, the allocation of nurses to ICU and SDU beds, respectively, such that we minimize the cost function:

$$\min_{0\leq b_I \leq \left(r_I \wedge \frac{\bar{\lambda}}{\mu_C}\right)} \left\{ w_C\left(\bar{\lambda} - b_I\mu_C\right) + w_{SC}\left(b_I\mu_C p - r_S\left(1 - \frac{b_I}{r_I}\right)\mu_{SC}\right)^+ \right\} \tag{7}$$

We can solve the preceding optimization problem to determine how to allocate nurses between the ICU and SDU. We find that the optimal policy is highly dependent on the relationship between abandonment and bumping costs. More formally, we have:

[3] $f(x) := o(x)$ if $f(x)/x \to 0$ as $x \to \infty$.

[4] $f(x) := \mathcal{O}(x)$ if $f(x)/x \to c > 0$ as $x \to \infty$.

Proposition 3*In the fluid model, under heavy traffic, the optimal allocation of nurses can be split into two cases, depending on the relative cost between abandonment and bumping. The cost minimizing allocation of nurses to ICU beds is given by:*

$$b_I^* = \begin{cases} r_I \wedge \frac{\bar{\lambda}}{\mu_C}, & \text{if } \frac{w_C}{w_{SC}} > \frac{r_I p \mu_C + r_S \mu_{SC}}{r_I \mu_C}, \textbf{ ID regime} \\ \frac{r_I r_S \mu_{SC}}{r_I \mu_C p + r_S \mu_{SC}}, & \text{if } \frac{w_C}{w_{SC}} \leq \frac{r_I \mu_C p + r_S \mu_{SC}}{r_I \mu_C}, \textbf{ ISD regime} \end{cases} \quad \text{and } b_S^* = r_S \left(1 - \frac{b_I^*}{r_I}\right)$$

Our proposed nurse allocation to ICU and SDU beds, respectively, based on fluid analysis is thus[5]

$$B_I^* = b_I^* N, \quad B_S^* = b_S^* N.$$

The proof of Proposition 3 is trivial and, hence, omitted. Note that one must verify that the value of b_I^* under the second scenario does not exceed $\bar{\lambda}/\mu_C$, which is true due to the heavy traffic condition. We have two regimes of interest. When the per-patient cost of abandonment is very large, we see the optimal policy is to allocate as many nurses to the ICU as needed in order to satisfy all Critical patients demand. If there are not enough nurses to meet all of this demand (i.e. $r_I \mu_C < \bar{\lambda}$), then all nurses should be allocated to the ICU. We call this regime the ICU-Driven (ID) regime[6]. On the other hand, when the per-patient costs of abandonment and bumping are close, then the optimal policy is to allocate some nurses to the SDU and allow for some (or more) Critical patients to abandon. We call this regime the ICU- and SDU- Driven (ISD) regime. Also note that the larger the capacity gained by transferring a nurse from the ICU to the SDU (increasing $\frac{r_S \mu_{SC}}{r_I \mu_C}$), the more likely the ISD regime is to be optimal. Additionally, if many Critical patients become Semi-critical (large p) the SDU becomes more beneficial.

To understand the intuition behind Proposition 3 consider the following marginal analysis. Suppose that $N \wedge \frac{\lambda}{\mu_C r_I}$ nurses are initially allocated to the ICU. Let us examine the impact of moving one nurse to the SDU or, equivalently, removing r_I beds from the ICU, while adding r_S beds to the SDU. In particular, and since $r_S > r_I$, the overall number of beds increases. This will result in an increase in average abandonment cost of Critical patients of approximately $w_C r_I \mu_C$. Additionally, there will be a decrease in average bumping cost of Semi-critical patients of approximately $w_{SC}(r_I \mu_C p + r_S \mu_{SC})$. Thus, transferring this nurse from the ICU to the SDU is worthwhile if and only if the total cost increase is negative, i.e. if and only if $\frac{w_C}{w_{SC}} \leq \frac{r_I \mu_C p + r_S \mu_{SC}}{r_I \mu_C}$. If this condition holds, it will be worthwhile to transfer nurses from the ICU to the SDU until the average bumping cost becomes 0, i.e. when $b_I \mu_C p = b_S \mu_{SC} = r_S \left(1 - \frac{b_I}{r_I}\right) \mu_{SC}$. This is equivalent to the following condition: $b_I = \frac{r_I r_S \mu_{SC}}{r_I \mu_C p + r_S \mu_{SC}}$, which is precisely the result of Proposition 3.

[5] From here on we ignore the integrality constraints and assume that those are obtained by rounding up or down the resulting proposed solution.

[6] Note that the optimality of allocating all nurses to the ICU in the ID regime is related to the literature on the optimality of imbalance in queueing systems (e.g. Green and Guha (1995)).

In further interpreting the results of Proposition 3, we have that in the ISD regime, the SDU size is selected such that the SDU is *critically* loaded: $\lambda_{SDU} \approx B_I^* \mu_C p \approx B_S^* \mu_{SC}$. At the same time, in the same regime, and under our heavy traffic assumption the ICU is strictly overloaded (due to Proposition 2). This is surprising because we assume that an abandonment from the ICU is more costly than bumping an SDU patient. Yet, this allocation results in having abandonment rate which is of order N and bumping rate which is of order $o(N)$. It seems that in the ISD regime, the capacity gains of allocating nurses to the SDU are more substantial than the gain of keeping the nurses in the ICU to serve the high priority (Critical) patients. In the ID regime, the needs of the Critical patients dominate. In practice, we see that some hospitals have SDUs while others do not. This raises the question as to whether different hospitals view the relative costs between abandonment and bumping differently as well as how different the system parameters are.

3.1. Implications of Relative Costs

The results of our fluid analysis suggest the optimal nurse allocation depends on the cost parameters w_C and w_{SC}. As stated earlier, it is currently difficult to concretely quantify w_C and w_{SC}. Though this will change as more data becomes available, it would be useful to explore what parameter regime (and subsequently, what allocation of nurses) hospitals should operate in. According to Proposition 3, it is sufficient to know the ratio w_C/w_{SC}. We leverage the result of Proposition 1 to see if this implies something about this ratio and thereby the optimal allocation of nurses in the overloaded regime.

Assuming that it is always optimal to give priority to Critical patients, we consider under what conditions this also implies that the system should operate in the ID regime. We see that when:

$$\frac{\theta - \mu_C(1-p) + \mu_{SC}}{\max\{\mu_C, 2\theta - \mu_C\}} > \frac{r_I \mu_C p + r_S \mu_{SC}}{r_I \mu_C}$$

it is optimal to allocate either enough nurses to the ICU in order to treat all Critical patients; if that is not possible, *all* nurses should be allocated to the ICU. This condition can be simplified to:

$$\begin{aligned} 1 + \frac{\theta - \mu_C}{\mu_{SC}} &\geq \frac{r_S}{r_I}, \text{ if } \mu_C \geq \theta \\ \frac{\mu_C[(\theta - \mu_C)(1-2p) + \mu_{SC}]}{\mu_{SC}(2\theta - \mu_C)} &\geq \frac{r_S}{r_I}, \text{ if } \mu_C < \theta \end{aligned} \tag{8}$$

Note that in practice, we expect that the abandonment rate will be substantially higher than the service rate of Critical patients so that $\theta > \mu_C$. This is because service times in the ICU are on the order of days, and critically ill patients cannot afford to wait that long for a bed. If the wait is that substantial, the patient is likely to be transferred to another unit or hospital (hence, 'abandons'). Additionally, the condition for $\mu_c \geq \theta$ can never be satisfied because, by assumption, $r_s \geq r_I$.

4. Diffusion Analysis

In this section, we consider refining our analysis from Section 3 by examining the impact of reallocating a small number of nurses to either the ICU or SDU. Our starting point is the analysis of the fluid approximation in Section 3. Under the ID regime it is optimal to have as big of an ICU as necessary/possible, while in the ISD regime, it is optimal to have an SDU which is comparable in size to the ICU. In this section, we consider how the reallocation of nurses of the order $\sqrt{N}$ may help. We find that in some cases, this reallocation can be quite helpful.

The fluid analysis finds the optimal allocation of nurses to the ICU and SDU up to an order of $o(N)$. However, the fluid analysis excludes lower ordered terms and so it might still be beneficial to reallocate a small number of nurses, say of order $\sqrt{N}$ to the SDU or ICU. We will use *diffusion* analysis to examine these two regimes.

4.1. Diffusion Analysis in the ID regime

Recall that in the ID regime, the fluid solution allocates all nurses possible/required to the ICU so that the abandonment cost is $o(N)$ (if possible), or negligible in fluid scale. We now explore the benefits of reallocating a *small* number of nurses, of order $\mathcal{O}(\sqrt{N})$ between the ICU and the SDU. In this section we assume that

$$\frac{w_C}{w_{SC}} > \frac{r_I\mu_C p + r_S\mu_{SC}}{r_I\mu_C},$$

and therefore, on a fluid level, it is optimal to operate the system in the ID regime. Additionally, as before, we assume that Critical patients get priority for ICU beds.

While ICUs are often operated at or above capacity, it is undesirable to continuously be unable to satisfy the demand for ICU care. Patients who require ICU care are the most critical patients in the hospital, so it may be desirable to operate the ICU under critical load, rather than in overload. In particular, suppose that the number of nurses is large enough to satisfy

$$Nr_I \geq \frac{\lambda}{\mu_C} + o(N). \tag{9}$$

That is, the number of beds allocated to the ICU is $B_I^* = \lambda/\mu_C + o(N)$, and the ICU is critically loaded with respect to the Critical patients.

We now postulate the following refinement of the above nurse allocation scheme:

$$B_I = \frac{\lambda}{\mu_C} + \beta\sqrt{\frac{\lambda}{\mu_C}} + o(\sqrt{N}), \quad B_S = \frac{r_S}{r_I}\left(Nr_I - \frac{\lambda}{\mu_C} - \beta\sqrt{\frac{\lambda}{\mu_C}}\right) + o(\sqrt{N}), \tag{10}$$

where β is only restricted by the non-negativity constraints on B_I and B_S[7]. In particular, the ICU is critically loaded when focusing on Critical patients, and works under the QED regime (Halfin and Whitt 1981,

[7] Recall that we ignore the integrality constraints.

Garnett et al. 2002) with respect to the same patients. At the same time, due to our heavy traffic condition and by Proposition 2, the SDU is overloaded.

It is not clear that in this operating regime, the ICU and SDU will always be full with Critical and Semi-critical patients, respectively, as is the case under the fluid scaling. Because the ICU may not be full of Critical patients, the dynamics of our queueing system and, specifically, the flow of the Semi-critical patients is more complex. Before we can determine the optimal allocation of nurses, we must first understand more precisely when and to what extent Semi-critical patients will be treated in the ICU.

4.1.1. State-Space Collapse In order to develop an understanding of the patient flow dynamics, one can examine the two-dimensional process with state $(Q + Z_C, Z_{SC})$ (where Q denotes the queue length and Z_C (Z_{SC}) denotes the number of Critical (Semi-critical) patients occupying a bed). This process is clearly a Markov process under the strict priority of Critical patients over Semi-critical patients in the ICU; however, the dynamics of this process are intricate. While the dynamics of the Critical patients follow that of a fairly standard $M/M/B_I + M$ model, the dynamics of the Semi-critical patients cannot be analyzed separately from the Critical patients; the dynamics of the Critical patients determine precisely how many beds are available in the ICU to treat Semi-critical patients.

Given our goal is to gain some insights as to how to allocate the nurses between the two units in this case, it is important to be able to characterize the patient flows through the ICU and SDU. Despite the challenges which arise with the two-dimensional Markovian model, we are able to show that this two-dimensional process may be accurately approximated by a one-dimensional process. Let

$$\hat{Z}_C^N := \frac{1}{\sqrt{\lambda}}\left(Z_C^N - B_I^N\right), \quad \hat{Z}_{SC}^N := \frac{1}{\sqrt{\lambda}}\left(Z_{SC}^N - B_S^N\right),$$

describe the diffusion scaled number of patients occupying a bed within each of the two states, respectively. Also, let $\Rightarrow$ represent weak convergence. Then we have:

Theorem 1*(State-Space Collapse)* *In the ID regime and under the nurse allocation of (10) we have a state-space collapse. More formally, assuming that at time 0, $\hat{Z}_C^N(0) + \hat{Z}_{SC}^N(0) \Rightarrow 0$, as $N \to \infty$, then*

$$\hat{Z}_C^N + \hat{Z}_{SC}^N \Rightarrow 0, \ \text{ as } N \to \infty,$$

where the convergence is in D the space of all RCLL (Right Continuous with Left Limits) functions with values in $\mathbb{R}$, equipped with the Skorohod J_1 metric (see Whitt (2002)).

According to Theorem 1, in the diffusion scale, all beds are always full. In particular, it is sufficient to know the value of the one dimensional process $X_C^N := Q^N + Z_C^N$ in order to figure out the value of the two

dimensional process (X_C^N, Z_{SC}^N) (up to order $o(\sqrt{N})$). For example, if there is no queue ($Q^N = 0$), then we know that any ICU bed which is not occupied by a Critical patient will be used to treat a Semi-critical patient. Hence the term 'State-space collapse'. Specifically, the dynamics of our system can be summarized as follows:

1. The ICU is operated in the QED-regime with respect to Critical patients, so the number of Critical patients can be approximated by the diffusion analysis of the Erlang-A model of Garnett et al. (2002) with B_I servers.

2. The SDU is always full. If there are fewer than B_I Critical patients in the system, then Semi-critical patients fill the remaining ICU beds.

The intuition behind this theorem is as follows: The SDU is overloaded. In particular the rate at which it is losing patients due to lack of space is of order N. At the same time the ICU is in the QED regime with respect to Critical patients. In particular, the number of ICU beds that are not occupied by Critical patients is at most of order $\sqrt{N}$. As soon as some of these beds are empty they almost instantaneously become occupied by Semi-critical patients. Hence all beds are always full. Figure 3 shows a sample path of a simulation which illustrates the state-space collapse result. The figure shows that the total number of patients who are occupying beds deviates very little from the total number of 44 beds.

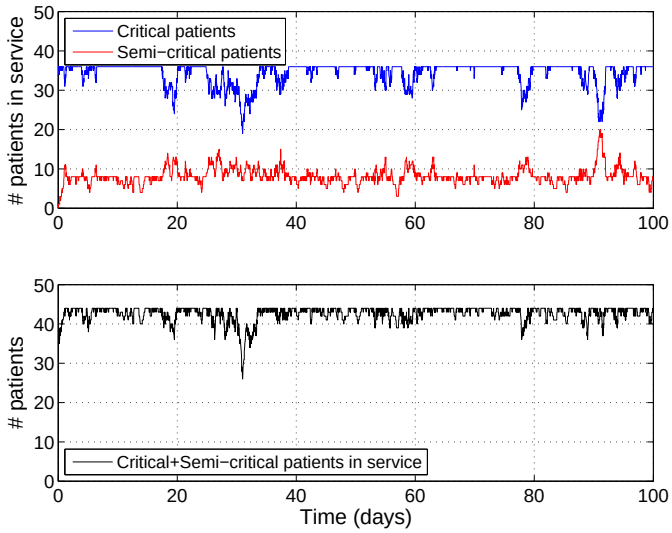


Figure 3 **A sample path demonstrating the State Space Collapse. The number of Critical and Semi-critical patients being treated in a bed are depicted. Note that if both the ICU and SDU are full (there are 44 total combined ICU and SDU beds), there may be a queue of Critical patients. The patients in the queue are not depicted in this figure. ($\lambda = 20$, $B_I = 36$, $B_S = 8$, $p = 0.5$, $\mu_C = \mu_{SC} = 0.5$, $\theta = 1$)**

4.1.2. Diffusion cost function We now leverage our results from above to examine the nurse allocation problem. Our aim is to derive expressions for the cost function using a diffusion approximation. Given the state-space collapse result that applies to the process $(Q + Z_C, Z_{SC})$, it is reasonable to expect that a similar state-space collapse applies in steady-state as well. Establishing this requires a formal justification of a limit interchange argument as in Theorem 9.10 of Ethier and Kurtz (1985). To avoid a lengthy and rather technical mathematical argumentation here we simply postulate that the same state-space collapse holds in steady-state as well.

We begin by stating a result that follows directly from results in Garnett et al. (2002) and Browne and Whitt (1995). Note, one could also consider using an alternative approximation, such as that in Baron and Milner (2009). Let $\hat{Q}^N := \frac{Q^N}{\sqrt{\lambda}}$ and $\hat{I}^N = \frac{I^N}{\sqrt{\lambda}}$ be the scaled queue length and "idleness" processes, where I^N is the number of ICU beds not occupied by Critical patients. Note that due to Theorem 1, I^N is also approximately equal to the number of Semi-critical patients who are being treated in the ICU. With a slight abuse of notation we also let $\hat{Q}^N$ and $\hat{I}^N$ represent these quantities in steady-state.

Theorem 2*(Erlang-A in Steady-State) In the ID regime, and under the nurse allocation in (10), we have that* $(\hat{Q}^N, \hat{I}^N) \Rightarrow (\hat{Q}, \hat{I})$*, as* $N \to \infty$*, with*

$$E[\hat{Q}] = \left(1 + \frac{h(\beta\sqrt{\mu_C/\theta})}{\sqrt{\mu_C/\theta} \cdot h(-\beta)}\right)^{-1} \cdot \left(-\frac{\sqrt{\mu_C}\beta}{\theta} + \sqrt{\frac{1}{\theta}} \cdot h\left(\beta\sqrt{\frac{\mu_C}{\theta}}\right)\right)$$

and

$$E[\hat{I}] = \frac{1}{\sqrt{\mu_C}}\left(1 - \left(1 + \frac{h(\beta\sqrt{\mu_C/\theta})}{\sqrt{\mu_C/\theta} \cdot h(-\beta)}\right)^{-1}\right) \cdot (\beta + h(-\beta))\,,$$

where $h(x) = \frac{\phi(x)}{1-\Phi(x)}$ *is the hazard rate function of the Standard Normal distribution.*

Note that Theorem 2 states the weak convergence of the steady-state random variable $(\hat{Q}^N, \hat{I}^N)$ but does not argue that convergence in expectation applies as well. This requires an additional technical argument which we omit. We simply postulate the convergence in expectation applies as well.

We now come up with diffusion approximations for the abandonment and bumping rates. Let Ab and Bm denote the steady-state abandonment and bumping rate, respectively. Then $Ab = \theta E[Q]$ and may be approximated by $Ab \approx \sqrt{\lambda}\theta E[\hat{Q}^N]$. The expression for the bumping rate is more involved. The starting point is that the bumping rate is equal to the Semi-critical arrival rate minus its total service rate. The arrival rate may be expressed as: $E[Z_C]\mu_C p$. Similarly, and assuming that the SSC result of Theorem 1 holds in steady-state, the departure rate may be expressed as: $B_S\mu_{SC} + E[I]\mu_{SC} + o(\sqrt{N})$. Putting all of the above

together we see that, under the ID regime and the nurse allocation (10), the cost function (centered by $w_{SC}\left(\lambda p+\frac{r_S}{r_I}\left(Nr_I-\frac{\lambda}{\mu_C}\right)\mu_{SC}\right)$ and scaled by $1/\sqrt{\lambda}$) may be approximated by:

$$C(\beta):=w_C\theta E[\hat{Q}]+w_{SC}\left[\beta\sqrt{\mu_C}p+\frac{r_S\beta\mu_{SC}}{r_I\sqrt{\mu_C}}-(\mu_{SC}+\mu_C p)E[\hat{I}]\right], \tag{11}$$

where the expressions for $E[\hat{Q}]$ and $E[\hat{I}]$ are explicitly given in Theorem 2.

Let $\beta^*:=\arg\min_\beta C(\beta)$, where we choose the supremum on β if there are multiple values of β that minimize the cost $C(\beta)$. Then our proposed solution in the ID regime is:

$$B_I^*:=\frac{\lambda}{\mu_C}+\beta^*\sqrt{\frac{\lambda}{\mu_C}},\quad B_S^*=\frac{r_S}{r_I}\left(Nr_I-\frac{\lambda}{\mu_C}-\beta^*\sqrt{\frac{\lambda}{\mu_C}}\right).$$

Note that we have not imposed upper and lower bounds on β^*. In particular, it is plausible that β^* is so small (including $\beta^*=-\infty$), that B_I^* is in fact smaller than what is proposed by the ISD regime, even though, by assumption, the system operates in the ID regime. To remedy this, we set a lower bound on B_I^* and an upper bound on B_S^* that are dictated by the fluid solution. In doing so, the allocation of nurses is given by:

$$B_I^*:=\max\left\{\frac{\lambda}{\mu_C}+\beta^*\sqrt{\frac{\lambda}{\mu_C}},\ \frac{r_Ir_S\mu_{SC}}{r_I\mu_C p+\mu_{SC}r_S}N\right\},$$

and

$$B_S^*=\min\left\{\frac{r_S}{r_I}\left(Nr_I-\frac{\lambda}{\mu_C}-\beta^*\sqrt{\frac{\lambda}{\mu_C}}\right),\ \frac{r_Ir_S\mu_C p}{r_I\mu_C p+r_S\mu_{SC}}N\right\}.$$

In the ID regime, the ICU is operated in QED with respect to the Critical patients. Hence, some Semi-critical patients will be treated in the ICU, so we can see that the reallocation of beds in this regime translates to balancing the tradeoff between flexibility (ICU beds) versus capacity (SDU beds). Note that this tradeoff only arises in this second order analysis of the ID regime.

4.2. Diffusion Analysis in the ISD regime

Recall that the fluid analysis identified two operating regimes for the system: the ID and ISD regimes. Now we take a closer look at the ISD regime. In particular, we focus on the case where

$$\frac{w_C}{w_{SC}}\le\frac{r_I\mu_C p+r_S\mu_{SC}}{r_I\mu_C}.$$

In this case, according to Proposition 3, we have that

$$B_I^*=b_I^*N+o(N),\quad b_I^*=\frac{r_Ir_S\mu_{SC}}{r_I\mu_C p+r_S\mu_{SC}},\text{ and }B_S^*=b_S^*N+o(N),\quad b_S^*=\frac{r_Ir_S\mu_C p}{r_I\mu_C p+r_S\mu_{SC}}.$$

In particular, we have that the ICU is overloaded and the SDU is critically loaded. Our aim here is to see whether an order of $\sqrt{N}$ refinement for the $o(N)$ terms above can lead to a lower cost. We further assume

that $\lambda = \mathcal{O}(N)$ so that the ICU operates in the efficiency-driven (ED) regime (Gans et al. 2003). Otherwise, the ICU will be "super" overloaded, and refinements of this order will not make a noticeable difference. Set

$$B_I = b_I^* N + o(N) = \gamma R_I + \delta\sqrt{R_I} + o(\sqrt{R_I}), \quad R_I := \frac{\lambda}{\mu_C}, \tag{12}$$

where $\gamma = \frac{N r_I r_S \mu_{SC} \mu_C}{\lambda(r_I \mu_C p + r_S \mu_{SC})}$ is less than 1 due to our heavy traffic condition. Also, let

$$B_S = b_S^* N + o(N) = R_S + \beta\sqrt{R_S} + o(\sqrt{R_S}), \quad R_S := \frac{B_I \mu_C p}{\mu_{SC}}, \tag{13}$$

where β and δ are only restricted by the non-negativity constraints on B_I and B_S. R_I is the offered load of the ICU, by definition. We argue that R_S is the offered load of the SDU. To see this, note that, since $\gamma < 1$, the ICU is indeed operated in the ED regime. In particular, all ICU beds are full with Critical patients all the time almost surely. Hence, the arrival rate into the SDU is equal to $B_I \mu_C p$, and the offered load is indeed equal to $\frac{B_I \mu_C p}{\mu_{SC}}$. Note that, as expected, the SDU is critically loaded, and operates in the QED regime. Finally, using the relation $\frac{B_I}{r_I} + \frac{B_S}{r_S} = N + o(\sqrt{N})$ we obtain that

$$\delta := \delta(\beta) = -\sqrt{\frac{N}{\lambda}} \frac{\beta r_I \mu_C \mu_{SC} \sqrt{\frac{r_I r_S p}{r_I \mu_C p + r_S \mu_{SC}}}}{r_I \mu_C p + r_S \mu_{SC}}.$$

Our goal at this point is to find a value for β that minimizes the expected abandonment and bumping costs. In particular, in this regime, we wish to minimize

$$w_C \lambda Pr\{Ab\} + w_{SC} B_I \mu_C p Pr\{Bm\}. \tag{14}$$

We are now ready to use asymptotic expressions for the probabilities of abandonment and bumping, respectively, that are available in the literature. Specifically, from Theorem 4.3 of Mandelbaum and Zeltyn (2009), we have that

$$Pr\{Ab\} = (1-\gamma) - \delta \Big/ \sqrt{\frac{\lambda}{\mu_C}} + o(1/\sqrt{\lambda}).$$

Additionally, from Jagerman (1974) we have that

$$Pr\{Bm\} = \frac{1}{\sqrt{B_S}} h(-\beta) + o(1/\sqrt{\lambda}).$$

Plugging the above into the cost expression (14), centering by $w_C \lambda(1-\gamma)$, scaling by $1/\sqrt{N}$, and letting $N \to \infty$, we obtain that the relevant cost function may be approximated by

$$C(\beta) = \mu_{SC} \sqrt{\frac{r_I r_S \mu_C p}{r_I \mu_C p + r_S \mu_{SC}}} \left(w_C \frac{\beta r_I \mu_C}{r_I \mu_C p + r_S \mu_{SC}} + w_{SC} h(-\beta) \right). \tag{15}$$

Let $\beta^* := \arg\min_\beta C(\beta)$, and let $\delta^* := \delta(\beta^*)$. Analogous to the ID regime, it is plausible that β^* is so small that the proposed B_I^* is larger than what is proposed by the ISD regime. We set an upper bound on B_I^* and

a lower bound on B_S^* that are dictated by the fluid solution. Then our proposed solution in the ISD regime is:

$$B_I^* = \min\left\{\gamma R_I + \delta^*\sqrt{R_I}, r_I N, \frac{\lambda}{\mu_C}\right\}, \quad R_I := \frac{\lambda}{\mu_C}, \tag{16}$$

and

$$B_S^* = \max\left\{R_S^* + \beta^*\sqrt{R_S^*}, \frac{r_S}{r_I}\left(Nr_I - \frac{\lambda}{\mu_C}\right)\right\}, \quad R_S^* := \frac{B_I^*\mu_C p}{\mu_{SC}}. \tag{17}$$

In the next section we examine the quality of our approximations in real world parameter regimes.

5. Numeric Results

We have utilized fluid and diffusion analysis to determine how to allocate nurses to ICU and SDU beds. We find that two operational regimes exist: the ID regime in which the SDU has very few beds, if any, and the ISD regime in which the SDU is comparable in size to the ICU. We now use numerical approaches to estimate the quality of our solutions in real hospital settings.

5.1. Empirical Data

To start, we must first calibrate the parameters of our model. To do this, we leverage the existing medical literature. Given the limited literature on SDUs, we identified two articles which specify the necessary parameters for our queueing model. The first article looks at the impact of adding an SDU for the cardiothoracic ICU at the University of Missouri Hospitals (Cady et al. 1995). They find that introducing the SDU reduces ICU length-of-stay (LOS) and then document the average patient flows through the ICU and SDU. The second article also considers the impact of introducing an SDU, but this time for the surgical ICU at New York-Presbyterian Hospital (Eachempati et al. 2004). The parameters from these articles are summarized in Table 5.1. Note that in the case of a surgical ICU, patients who wait for an ICU bed are likely waiting in the Post-Anesthesia Recovery Unit (PACU). As the PACU becomes more congested, surgeries may be delayed or canceled, which is highly undesirable. If the queue for the ICU gets very large, it is possible, though rare, for surgical patients to be treated in a medical ICU. This would be considered as 'abandoning' in our model. This is again, highly undesirable. Thus, it is likely in this case that $w_C \gg w_{SC}$. We let $\mu_{SC} = 1/ICULOS$ and $\mu_C = 1/SDULOS$. Note that this ignores Semi-Critical patients who may be treated in the ICU as well as censored observations due to abandonment and bumping.

Source	ICU LOS	SDU LOS	p	r_I	r_S
Cady et al. (1995)	2.5 days	1.2 days	0.65	1†	2-3
Eachempati et al. (2004)	4.8 days	2.3 days	0.8	2	4

Table 1 **Summary of ICU and SDU patient flow parameters. †The ICU nurse-to-patient ratio is not given in this article, so we assume it to be one-to-one.**

The last parameter of our model is the abandonment rate, θ. This parameter can be difficult to estimate in practice. While there may be records which capture the time an ICU bed was requested and if/when that request was recanted, hospitals may be reluctant to divulge denied patient admissions to the ICU due to potential legal liabilities. As such, we leverage the intuition generated by conversations with medical professionals and estimate that patients are willing to wait on average up to 1 day for an ICU bed (i.e. $\theta \geq 1$). If the time to admission is larger, they will either find treatment at another hospital/unit or die.

5.1.1. Hospital Operating Regime Given these hospital parameters, we can easily verify if the hospitals should be operating in the ID regime. Specifically, assuming that it is optimal to give priority to Critical patients, Section 3.1 provides the condition (8) which guarantees that it is optimal to allocate almost all nurses to the ICU. We examine whether this condition is satisfied by the hospitals in Table 5.1.

Assuming Critical patients have an average patience of 1 day (i.e. $\theta = 1$), we find that neither hospital satisfies the constraint. Thus, we cannot conclude that it is optimal to not have an SDU. This is somewhat comforting as both hospitals do have SDUs.

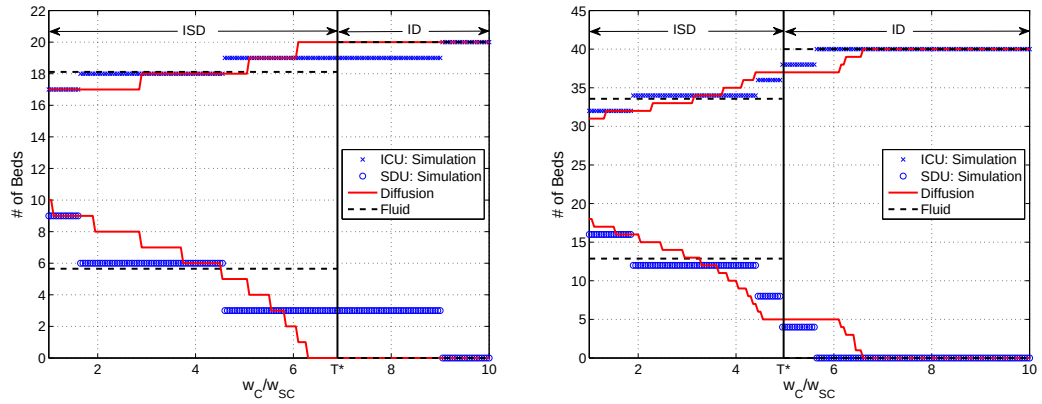
5.2. Simulation versus Theoretical Analysis

We now leverage the parameters from Table 5.1 to simulate patient flows through the ICU and SDU. Using an exhaustive search over simulations which examine the average costs incurred under every combination of nurse allocations, we find the optimal number of ICU and SDU beds. We then compare this to the allocation of nurses given by our fluid and diffusion analysis.

In considering the staffing level in the ICU, we expect that number of ICU beds to be non-decreasing in the ratio between the per-patient abandonment and bumping costs: w_C/w_{SC}. It turns out that because we have two different solution regimes (ID and ISD) at the diffusion level, it is possible the monotonicity is violated near the transition between these two regimes, i.e. when $w_C/w_{SC} = T^* := \frac{r_I \mu_C p + r_S \mu_{SC}}{r_I \mu_C}$. Indeed, we encounter this issue in our numeric analysis in some scenarios. For such scenarios, in order to translate our diffusion solution to maintain the desired monotonicity, at T^*, we assigned the number of ICU beds to be the average between the ID and ISD diffusion solutions. That is, let $B_I^*(\text{ID}, T^*)$ be the ID solution (minimizes (11)) and let $B_I^*(\text{ISD}, T^*)$ be the ISD solution (minimizes (15)) when $w_C/w_{SC} = T^*$. Then, our diffusion solution is $B_I^* = \frac{1}{2}[B_I^*(\text{ID}, T^*) + B_I^*(\text{ISD}, T^*)]$, which also serves as a lower (upper) bound for the number of ICU beds in the ID (ISD) regime.

In our simulations, we assume the arrival rate is such that the ICU is critically loaded in case all the nurses are allocated to the ICU. Specifically, $\lambda = N\mu_C r_I$. We also set $\theta = 1$. Figure 4 compares the simulated allocation to the derived allocation when there are 20 nurses to split amongst the ICU and SDU. As we can see in these figures, the solution determined by minimizing the cost in (11) and (15) is very close to

the solution determined by using exhaustive search over simulations. The fluid model is fairly accurate for many different weights, but can be quite coarse at times. Additionally, the accuracy of our approximation depends on the size of the system, with better results for larger systems. Because the nurse-to-patient ratios in Eachempati et al. (2004) require fewer nurses per patient than in Cady et al. (1995), the size of the units is twice as large for the Eachempati et al. (2004) parameters. As such the quality of the solution from the diffusion analysis is more accurate in Figure 4b than in Figure 4a. We also find that when we increase the number of nurses to allocate (for instance, to $N = 100$), the approximations become even more accurate.

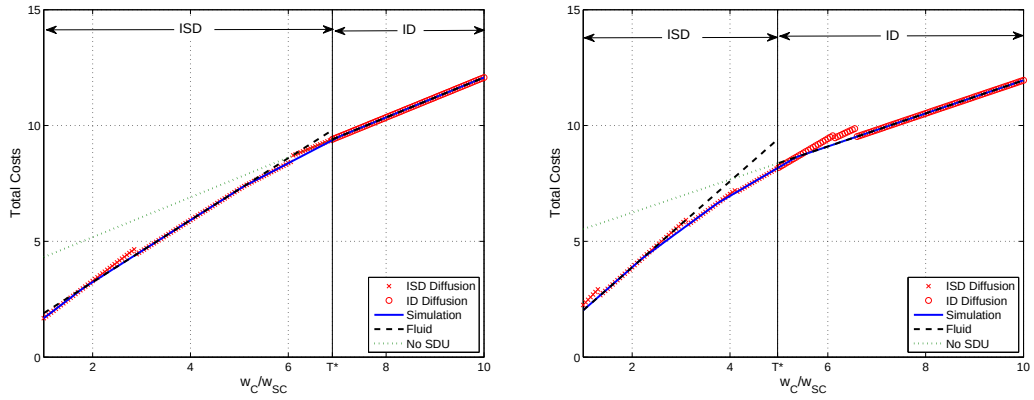


(a)Hospital Parameters from Cady et al. (1995) (b)Hospital Parameters from Eachempati et al. (2004)

Figure 4 Optimal allocation of nurses to beds via diffusion analysis and exhaustive search. $N = 20$ **nurses.**

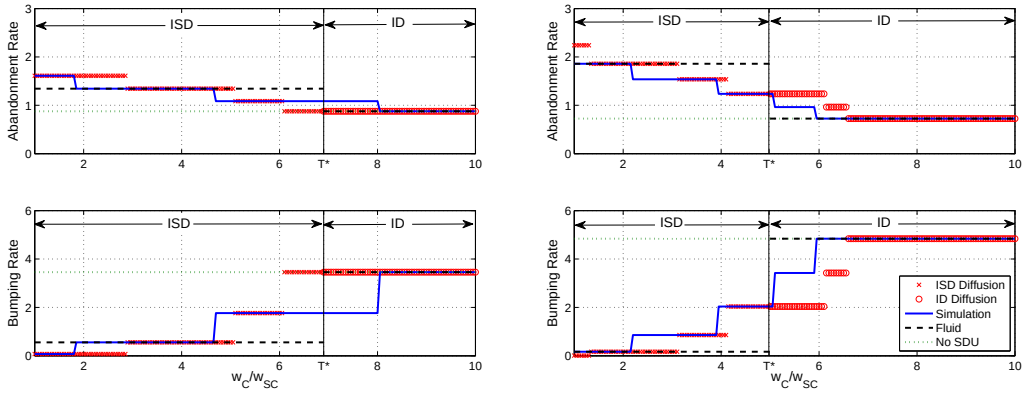
Though we see discrepancies in the number of beds in the ICU and SDU under the diffusion approximations, we find that the actual average cost incurred is quite close to optimal. Figure 5 compares the simulated costs (average abandonment and bumping costs) under the diffusion and fluid solutions to the minimum cost achieved via exhaustive search. Figure 6 compares the same when split by abandonment and bumping rates. We also provide a benchmark of not having any SDU. The cost differences under the diffusion solutions are always less than 6% and are typically within less than .1% of optimal. On the other hand, the fluid solution can incur more than 10% costs compared to optimal. Depending on the operating parameter regime, it may be sufficient to implement the fluid solution. In other instances, the diffusion solution can provide an important refinement to reduce costs.

We can also see that in the ID regime, it is certainly reasonable to put all nurses in the ICU. When the system is in the ISD regime, it is very important to consider introducing an SDU; not having an SDU can result in costs which are up to three times higher than that achieved via the optimal allocation.



(a)Hospital Parameters from Cady et al. (1995) (b)Hospital Parameters from Eachempati et al. (2004)

Figure 5 Average cost incurred under optimal allocation of nurses to beds via diffusion analysis and exhaustive search. $N = 20$ **nurses.**



(a)Hospital Parameters from Cady et al. (1995) (b)Hospital Parameters from Eachempati et al. (2004)

Figure 6 Average abandonment and bumping rates (in # patients per day) under optimal allocation of nurses to beds via diffusion analysis and exhaustive search. $N = 20$ **nurses.**

5.3. Sensitivity Analysis

Thus far, we have used simulation to examine the ICU and SDU sizing decision derived from our analysis. In doing so, we focused on a parameter regime where the ICU was overloaded and the system dynamics were estimated from 2 hospitals. We now consider how sensitive our results are to these parameter regimes.

5.3.1. System Parameters We start by considering how sensitive the nurse allocation decision is to changes to the system parameters, such as the abandonment rate, θ, the probability of becoming a Semi-critical patient, p, as well as the service rates of Critical and Semi-critical patients, μ_C and μ_{SC}. In order to focus on changes in these parameters, rather than the quantization effects of nurses, we consider the case where $N = 100$. We examine both the ID and ISD regimes. We perturb each of the system parameters

one-by-one and examine the change in the optimal allocation of nurses[8]. Since we are most concerned with the impact on patient outcomes, as measured by our weighted cost of abandonment and bumping rates, we focus on the increase in cost due to the mis-specification of system parameters. In particular, we optimize the allocation over the incorrect system parameters, but evaluate the cost with the true parameters which were originally summarized in Table 5.1.

Percentage perturbation	$w_C/w_{SC}=3$				$w_C/w_{SC}=10$			
	μ_C	μ_{SC}	θ	p	μ_C	μ_{SC}	θ	p
5%	0.34%	0.34%	0.34%	1.83%	0.85%	0.64%	0.64%	0.64%
10%	0.34%	1.83%	0.34%	1.83%	4.07%	0.64%	0.64%	0.64%
15%	1.83%	4.40%	0.34%	4.40%	6.72%	0.64%	0.64%	0.64%
20%	1.83%	4.40%	0.34%	6.69%	12.82%	0.64%	0.64%	0.64%
25%	4.00%	10.01%	0.34%	10.01%	12.82%	2.70%	0.64%	0.64%

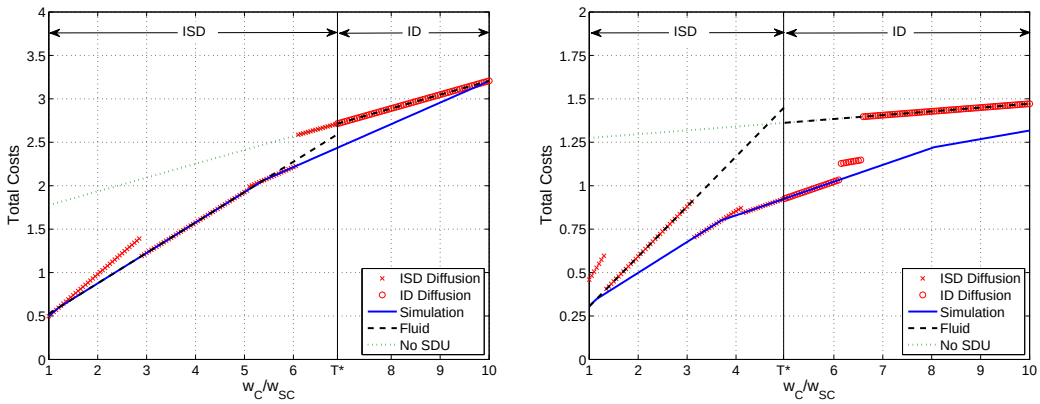
Table 2 **Percentage increase in cost when optimizing over perturbed system parameters. System parameters are given as in Cady et al. (1995)**

Table 2 summarizes the change in cost due to the mis-specification of these system parameters for the hospital in Cady et al. (1995). Similar results are seen with Eachempati et al. (2004). The system performance is very robust to mis-specifications of the abandonment rate, θ. For small perturbations ($\leq 10\%$, the system performance is reasonably robust–the nurse allocation is always within 4% of optimal. On the other hand, we can see that the system performance can be very sensitive to the service rate of Critical (μ_C) and Semi-Critical (μ_{SC}) patients, as well as the proportion of Critical patients who become Semi-Critical (p). Thus, it is important to provide reasonably accurate estimates of these parameters. We note that in the ISD regime, the performance is very sensitive to μ_C. This is because variation in μ_C can move the system away from the heavy traffic regime, substantially changing the optimal allocation.

5.3.2. Moderately Heavy Traffic Our fluid and diffusion analysis assumed a heavily loaded regime where the units were nearly always full. A number of hospitals strive for a target utilization of 85% and within New York, the average ICU occupancy level was 75% (Green 2003). We note that these utilization metrics are censored measures of the true demand due to adaptive techniques–such as abandonment and bumping–which can divert arrivals and reduce length-of-stay. Still, there may be cases where the ICU is not run in heavy traffic, so we also consider the quality of our analysis in a 'moderately' heavy traffic regime. The traffic load in this case is such that if all nurses are allocated to the ICU, the nominal load of patients is 85%, i.e. $\frac{\lambda(r_I\mu_C p+r_S\mu_{SC})}{Nr_I r_S\mu_C\mu_{SC}}=.85$. While the optimal allocation of nurses changes slightly in this case, we see in Figure 7 that the diffusion and fluid solutions still perform reasonably well in terms of costs in the ISD

[8] The optimal allocation is derived via exhaustive search.

regime. In the ID regime, the asymptotic approach does not work as well. This is because, in this moderate traffic, the ICU is very far from operating in heavy traffic and the quality of the approximations noticeably degrades.



(a)Hospital Parameters from Cady et al. (1995) (b)Hospital Parameters from Eachempati et al. (2004)

Figure 7 Average cost incurred under optimal allocation of nurses to beds via diffusion analysis and exhaustive search. $N = 20$ **nurses. Moderate traffic:** $\frac{\lambda(r_I\mu_C p + r_S\mu_{SC})}{N r_I r_S \mu_C \mu_{SC}} = .85$

6. Model Extensions

Thus far, the focus of this work has been on the model presented in Section 2. We now consider a number of extensions to our initial model which capture additional dynamics which can arise in various hospital settings. In particular, we explicitly consider readmissions, variants to the budget neutral nursing constraint, and time-varying arrivals.

6.1. Readmissions

We start by considering a stylized model which incorporates patient readmissions. In particular, when a patient leaves the system, there is some probability he will return to the hospital in the Critical state. The probability a patient will be readmitted depends on how the patient left the system. We let p_A and p_C denote the readmission probability for patients who abandoned as Critical patients or who completed ICU service and left the system, respectively. Similarly, p_B and p_{SC} denote the readmission probability if the patient is bumped as a Semi-critical patient or if the patient completed service as a Semi-critical patient in the ICU or SDU. As readmitted patients are typically worse off, we will assume that readmitted patients will not abandon and cannot be bumped. Thus, the expected length of stay of a readmitted patient, not including waiting time, is $E[LOS_R|Readmitted] = \frac{1}{\mu_C} + p\frac{1}{\mu_{SC}}$. Finally, the expected readmission load is then $p_R E[LOS_R|Readmitted]$, where p_R denotes the readmission risk of the patient and depends on how the patient departs the system.

In the appendix, we formally introduce this model with readmissions. Additionally, we establish the stability condition of such a system. Similar to Chan et al. (2012), we find that minimizing the expected readmission load corresponds to maximizing throughput.

Corollary 1*If the abandonment and bumping costs capture the* **increase in readmission load** *associated with these events, then the allocation of nurses which minimizes the average abandonment and bumping costs will also minimize the number of nurses necessary to stabilize the readmission queue.*

Now, we use simulation to compare the the quality of our nurse allocation derived from our original model when considering a model which incorporates readmissions. We assume the following readmission probabilities: $p_A = .10$, $p_B = .05$, and $p_C = p_{SC} = .02$. We assume the time to readmission is exponentially distributed with mean $1/\delta = 5$ days. We consider the nurse allocation for our original model in Section 2 which minimizes the readmission rate on the diffusion scale derived in Section 4. We then evaluate the performance of this solution via a simulation model that does have readmissions to the solution achieved via an exhaustive search for the model with readmissions. We consider the case with $N = 20$ nurses.

	Percentage of Patients Readmitted		
	Original Model (without Readmissions)	Exhaustive Search via simulation	All nurses in ICU (No SDU)
Eachempati et al. (2004)	18.6%	18.8%	20.5%
Cady et al. (1995)	31.0%	31.4%	36.5%

Table 3 **A system with readmissions: Comparison of readmission rates for solution which ignores readmissions (Original Model) to solution established via exhaustive search.**

Table 3 summarizes our simulation results for a system with readmissions. We can see that the number of readmissions achieved via our diffusion solution for a model *without* readmissions, but with cost appropriately defined as the increase in readmission risk due to abandonment and/or bumping, is very close to the minimum percentage of readmissions. As a benchmark, we see that when there is no SDU, the percentage of readmissions increases.

6.2. Relaxing the nursing constraint

Thus far, we have considered the ICU and SDU sizing decision under the assumption that the number of nurses must be held constant. This budget neutral constraint appears in a number of settings. However, it is conceivable that the joint ICU and SDU sizing decision may not have such a strict constraint on the number of nurses. For instance, a hospital may consider hiring M additional nurses and must determine whether to allocate them all to the ICU or SDU or split the nurses across both units. Alternatively, a hospital may not want to completely resize the units and may just want to consider 2 potential options.

Our analysis provides some insight into these other problem formulations. In particular, given an allocation a specific number of ICU and SDU beds, B_I and B_S, one can easily calculate the number of nurses N. Given the arrival rate λ at the hospital, one can use the analysis from Sections 3 and 4 to evaluate the operational parameter regime and assess the performance–in terms of abandonment and bumping rates–of such a configuration. That is, our results are also useful for *performance analysis*.

6.3. Time-varying arrivals

In practice, hospitals tend to have arrival rates that are highly time variable (Green et al. 2006, Armony et al. 2010), while the unit sizes remain fixed for a while. Accounting for this time variation when determining staffing levels in the Emergency Department (ED) can lead to much better provision of care (Green et al. 2006, Yom-Tov and Mandelbaum 2013). As many ICU patients originate from the ED, the time-varying arrival rates to the hospital translate to time-varying arrival rates to the ICU. However, unlike the ED, the service times in the ICU are very long ($\sim$2-4 days as seen in Table 5.1) whereas the variation in arrival rates is on the order of hours. This difference in time scale suggests that it is not essential to capture time variation when establishing staffing levels in the ICU. For more discussion of this see Yom-Tov and Mandelbaum (2013) as well as Section 5.2 and Figure 13 in Chan et al. (2013b).

7. Conclusion

Within the medical community, there is a lot of uncertainty on how to manage and size SDUs. In this work, we consider the optimal allocation of nurses for the inpatient units used to treat the hospitals most critical patients: the ICU and SDU. In doing so, we provide insight into when and how the SDU can be useful in managing patient flow.

Hospital units are often congested, so we analyze our hospital system under a many-beds heavy-traffic asymptotic regime and consider how to optimally tradeoff flexibility and capacity given abandonment and bumping costs. Via our fluid analysis, we identify two parameter regimes–the ICU-Driven and ICU-and-SDU Driven regimes–which dictate the optimality of allocating a very small (including zero) or a substantial number of nurses to the SDU. We also leverage a state-space collapse result to evaluate and optimize the staffing allocation in the diffusion scale. Numerically, we find that our analysis in these asymptotic regimes can be quite accurate.

One of the biggest challenges in translating from a model into practice is calibrating the model and determining the appropriate cost parameters. Using a Dynamic Programming framework, we translate medical intuition provided via conversations with clinicians into concrete conditions on the relationship between our model's cost parameters. In combining this with our queueing analysis, we establish an easily verifiable condition in Section 3.1 which is sufficient for the optimality of the ID regime. Thus, hospital managers

can examine their patient population and relative staffing requirement in the ICU versus SDU to determine whether it is optimal to close (or never open) an SDU.

In practice, there is high variation across hospitals as to whether it has an SDU and if so, how large the unit is in comparison to the ICU. On the surface, this variation could be attributed to the fact there is limited concensus in the medical community as to the management of SDUs. However, our analysis provides justification for this variation. The optimal size of an SDU is highly dependent on patient mix (including differences in service times and the likelihood of becoming a Semi-critical patient following ICU care), staffing requirements in the ICU versus SDU, as well as the relative cost of Critical patient abandonment versus bumping a Semi-critical patient to an even lower level of care, such as the general ward. Because these factors are likely to vary substantially across different hospitals and geographic areas, it is reasonable–and highly desirable–that hospitals utilize and size SDUs in a heterogenous manner. One size does not fit all.

This work suggests several potential directions for future research. For instance, if a new hospital were being built, it would be useful to consider the staffing decision without the budget neutral constraint. In such a setting, a third tradeoff arises: staffing costs versus abandonment and bumping costs. Another direction would be to consider other patient flows through the SDU. In this work, we only consider SDU patients who originate in the ICU; however, in practice, there may be patients treated in the SDU who never visit the ICU and/or patients from the SDU who visit the ICU following their SDU stay. One could also consider different priority rules, so that in some cases a Critical patient will have to wait (and potentially abandon), even if there is a Semi-critical patient in the ICU which could be bumped. Another option is to allow Critical patients to balk when and if the queue gets too long. Incorporating such dynamics would alter patient flows and require new analysis. Finally, in this paper we have focused on sizing the ICU and SDU, while ignoring the size of the general wards. This is because the ICU is often considered the hospital bottleneck. An interesting direction for future research is to explicitly model the size and dynamics of the general ward along with the other two units.

Despite some of these limitations of our model, our work provides an important first step into addressing the substantial debate in the medical community as to if and how SDUs should be used. The prevailing sentiment amongst SDU supporters is that they are a cost effective way to provide care to Semi-critical patients. This is true in some cases (ISD regime). However, in the ID regime, we see that the need of the high priority patients outweighs the additional capacity generated by moving nurses to the SDU. Still, even in this regime, a *small* SDU can be beneficial in serving as a buffer between the ICU and the hospital wards. The insights from our work will be useful for hospital managers deciding how to staff their hospital units.

References

Akan, M., B. Ata, T. Olsen. 2012. Congestion-based leadtime quotation for heterogeneous customers with convex-concave delay costs: Optimality of a cost-balancing policy based on convex hull functions. *Operations Research, forthcoming* .

Aloe, K., L. Raffaniello, M. Ryan, L. Williams. 2009. Creation of an Intermediate Respiratory Care Unit to Decrease Intensive Care Utilization. *Journal of Nursing Administration* **39**(11) 494–498.

Andradottir, S., H. Ayhan, H. Eser Kirkizlar. 2013. Flexible servers in tandem lines with setups. *Working paper, Georgia Institute of Technology* .

Armony, M., S. Israelit, A. Mandelbaum, Y.N. Marmor, Y. Tseytlin, G.B. Yom-tov. 2010. Patient flow in hospitals: A data-based queueing-science perspective. *Working Paper, Stern School of Business* .

Ata, B., B.L. Killaly, T.L Olsen, R.P. Parker. 2012. On hospice operations under medicare reimbursement policies. *Management Science, forthcoming* .

Ata, B., J. A. Van Mieghem. 2009. The Value of Partial Resource Pooling: Should a Service Network Be Integrated or Product-Focused? *Management Science* **55**(1) 115–131.

Baron, O., J. Milner. 2009. Staffing to maximize profit for call centers with alternate service-level agreements. *Operations Research* **57**(3) 685–700.

Bassamboo, A., R. S. Randhawa. 2010. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations Research* **58** 1398–1413.

Bassamboo, A., R.S. Randhawa, J.A. Van Miegham. 2012. A Little Flexibility is All You Need: On the Asymptotic Value of Flexible Capacity in Parallel Queuing Systems. *Operations Research* **60**(6) 1423–1435.

Beck, M. 2011. Critical (Re)thinking: How ICUs are getting a much-needed makeover. *Wall Street Journal, March 28* .

Bell, S. L., R. J. Williams. 2001. Dynamic Scheduling of a System with Two Parallel Servers in Heavy Traffic with Resource Pooling: Asymptotic Optimality of a Threshold Policy. *Annals of Applied Probability* **11**(3) 608–649.

Bertsekas, D. P. 2005. *Dynamic Programming and Optimal Control*. Athena Scientific.

Browne, S., W. Whitt. 1995. Piecewise-linear diffusion processes. Dshalalow, ed., *Advances in queueing: Theory, methods, and open problems*. CRC Press, Boca Raton, FL, 463–480.

Byrick, R.J., J.D. Power, J.O. Ycas, K.A. Brown. 1986. Impact of an intermediate care area on ICU utilization after cardiac surgery. *Critical care medicine* **14**(10) 869.

Cady, N., M. Mattes, S. Burton. 1995. Reducing Intensive Care Unit Length of Stay: A Stepdown Unit for First-Day Heart Surgery Patients. *Journal of Nursing Administration* **25**(12) 29–35.

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, N. Bambos, G. Escobar. 2012. Optimizing ICU Discharge Decisions with Patient Readmissions. *Operations Research* **60** 1323–1341.

Chan, C. W., V. F. Farias, G. Escobar. 2013a. The Impact of Delays on Service Times in the Intensive Care Unit. *Working paper, Columbia Business School* .

Chan, C.W., G. Yom-Tov, G. Escobar. 2013b. When to use Speedup: An Examination of Service Systems with Returns. *Operations Research (to appear)* .

Charles, PG, R Wolfe, M Whitby, MJ Fine, AJ Fuller, R Stirling, AA Wright, JA Ramirez, KJ Christiansen, GW Waterer, RJ Pierce, JG Armstrong, TM Korman, P Holmes, DS Obrosky, P Peyrani, B Johnson, M Hooy. 2008. Smart-cop: a tool for predicting the need for intensive respiratory or vasopressor support in community-acquired pneumonia. *Clin Infec Dis* **47**(3) 375–384.

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.

Dai, J. G., T. Tezcan. 2008. Optimal Control of Parallel Server Systems with Many Servers in Heavy Traffic. *Queueing Systems* **59** 95–134.

de Véricourt, F., O.B. Jennings. 2008. Dimensioning large-scale membership services. *Operations Research* **56**(1) 173–187.

Duke, G.J., M.D. Buist, D. Pilcher, C.D. Scheinkestel, J.D. Santamaria, G.A. Gutteridge, P.J. Cranswick, D. Ernest, C. French, J.A. Botha. 2009. Interventions to circumvent intensive care access block: a retrospective 2-year study across metropolitan Melbourne. *Med J Aust* **190** 375–378.

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.

Eachempati, S. R., L. J. Hydo, P. S. Barie. 2004. The effect of an intermediate care unit on the demographics and outcomes of a surgical intensive care unit population. *Archives of Surgery* **139**(3) 315–319.

Ethier, S.N., T.G. Kurtz. 1985. *Markov processes, characterization and convergence*. John Wiley & Sons.

Fine, MJ, TE Auble, DM Yealy, BH Hanusa, LA Weissfeld, DE Singer, CM Coley, TJ Marrie, WN Kapoor. 1997. A prediction rule to identify low-risk patients with community-acquired pneumonia. *NEJM* **336**(4) 243–250.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone Call Centers: Tutorial, Review, and Research Prospects. *MSOM* 79–141.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4** 208–227.

Ghamami, S., A. R. Ward. 2012. Dynamic Scheduling of a Two-Server Parallel Server System with Complete Resource Pooling and Reneging in Heavy Traffic: Asymptotic Optimality of a Two-Threshold Policy. *Mathematics of operations research (to appear)* .

Green, L. 1985. A Queueing System with General-Use and Limited-Use Servers. *Operations Research* **33** 168–182.

Green, L., D. Guha. 1995. On the efficiency of imbalance in multi-facility multi-server service systems. *Management Science* **41**(1) 179–187.

Green, L. V. 2003. How many hospital beds? *Inquiry* **39**(4) 400–412.

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.

Green, L.V., S. Savin, N. Savva. 2013. 'Nursevendor Problem': Personnel Staffing in the Presence of Endogenous Absenteeism. *Management Science* **59**(10) 2237–2256.

Gurvich, I., W. Whitt. 2009a. Queue-and-Idleness-Ratio Controls in Many-Server Service Systems. *Math of OR* **34** 363–396.

Gurvich, I, W Whitt. 2009b. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing and Service Operations Management* **11**(2) 237–253.

Gurvich, I, W Whitt. 2010. Service-level differentiation in many-server service systems via queue-ratio routing. *Operations Research* **58**(2) 316–328.

Halfin, S., W. Whitt. 1981. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research* **29** 567–588.

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. *Crit Care Med* **38** 65–71.

Hanson, C. W., C. S. Deutschman, H. L. Anderson, P. M. Reilly, E. C. Behringer, C. W. Schwab, J. Price. 1999. Effects of an organized critical care service on outcomes and resource utilization: A cohort study. *Critical Care Medicine* **27**(2) 270–274.

Harding, A. D. 2009. What Can An Intermediate Care Unit Do For You? *Journal of Nursing Administration* **39**(1) 4–7.

Harrison, J.M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the halfin and whitt heavy traffic regime. *Operations Research* **52** 243–257.

Hopp, W.J., E. Tekin, M.P. Van Oyen. 2004. Benefits of skill chaining in serial production lines with cross-trained workers. *Managemet Science* **50**(1) 83–98.

Iravani, S.M.R., M.P. Van Oyen, K.T. Sims. 2005. Structural flexibility: A new perspective on the design of manufacturing and service operations. *Management Science* **51**(2) 151–166.

Jagerman, D. L. 1974. Some properties of the erlang loss function. *Bell Systems Tech. J.* **53** 525–551.

Joint Commission Resources. 2004. *Improving Care in the ICU*. Joint Commission on Accredidation of Healthcare Organizations.

Kc, D., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* **14**(1) 50–65.

Keenan, S. P., W. J. Sibbald, K. J. Inman, D. Massel. 1998. A Systematic Review of the Cost-Effectiveness of Non-cardiac Transitional Care Units. *Chest* **113** 172–177.

Kim, S-H, C. W. Chan, M. Olivares, G. Escobar. 2012. Managing Inpatient Units: An Empirical Study of Capacity Allocation and its Implication on Service Outcomes. *Working Paper, Columbia Business School* .

Kirkizlar, H. Eser, S. Andradottir, H. Ayhan. 2013. Flexible servers in understaffed tandem lines. *POMS, to appear* .

Kostami, V., A.R. Ward. 2009. Managing service systems with an offline waiting option and customer abandonment. *Manufacturing & Service Operations Management* **11**(4) 644–656.

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.

Mandelbaum, A, A Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized c μ -rule. *Operations Research* **52**(6) 836–855.

Mandelbaum, A., S. Zeltyn. 2009. Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research* **57** 1189–1205.

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

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**(10) 1345–1355.

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

Reiman, M.I. 1984. Some diffusion approximations with state space collapse. F. Baccelli, G. Fayolle, eds., *Modelling and Performance Evaluation Methodology*. Springer-Verlag, 209–240.

Rubino, M., B. Ata. 2009. Dynamic control of a make-to-order parallel-server system with cancellations. *Operations Research* **57**(1) 94–108.

Ryckman, F.C., P.A. Yelton, A.M Anneken, P.E. Kiessling, P.J. Schoettker, U.R. Kotagal. 2009. Redesigning intensive care unit flow using variability management to improve access and safety. *Joint Commission journal on quality and patient safety / Joint Commission Resources* **35** 535–43.

Shmueli, A., C.L. Sprung. 2005. Assessing the in-hospital survival benefits of intensive care. *International Journal of Technology Assessment in Health Care* **21**(1) 66–72.

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

Simchen, E., C.L. Sprung, N. Galai, Y. Zitser-Gurevich, Y. Bar-Lavi, G. Gurman, M. Klein, A. Lev, L. Levi, F. Zveibil, et al. 2004. Survival of critically ill patients hospitalized in and out of intensive care units under paucity of intensive care unit beds. *Critical care medicine* **32**(8) 1654.

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.

Stacy, K. M. 2011. Progressive Care Units: Different but the Same. *Critical Care Nurse* **31**(3) 77–83.

State of California Office of Statewide Health Planning & Development. 2010-2011. Annual Financial Data. URL `http://www.oshpd.ca.gov/HID/Products/Hospitals/AnnFinanData/CmplteDataSet/index.asp`.

Tezcan, T., J.G. Dai. 2010. Dynamic Control of N-Systems with Many Servers: Asymptotic Optimality of a Static Priority Policy in Heavy Traffic. *Operations Research* **58** 94–110.

Tosteson, A., L. Goldman, I. S. Udvarhelyi, T. H. Lee. 1996. Cost-effectiveness of a coronary care unit versus an intermediate care unit for emergency department patients with chest pain. *Circulation* **94**(2) 143–150.

Tsitsiklis, J.N., K. Xu. 2012. On the power of (even a little) resource pooling. *Stochastic Systems* **2** 1–66.

Vincent, J. L., R. Moreno. 2010. Clinical review: scoring systems in the critically ill. *Crit Care* **14**(2) 207.

Wallace, R.B., W. Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manufacturing and Service Operations Management* **7** 276–294.

Whitt, W. 2002. *Stochastic-Process limits: An Introduction to Stochastic Process Limits and their applications to Queues*. Spring-Verlag, New York.

Whitt, W. 2006. Fluid Models for Multiserver Queues with Abandonments. *Operations Research* **54** 37–54.

Yankovic, N., L. Green. 2011. Identifying Good Nursing Levels: A Queuing Approach. *Operations Research* **59** 942–955.

Yom-Tov, G., A. Mandelbaum. 2013. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management (to appear)* .

Zhang, Bo, H. Ayhan. 2013. Optimal admission control for tandem queues with loss. *IEEE Transactions on Automatic Control, to appear* .

Zimmerman, J.E., D.P. Wagner, W.A. Knaus, J.F. Williams, D. Kolakowski, E.A. Draper. 1995. The use of risk predictions to identify candidates for intermediate care units. *Chest* **108**(2) 490.

Appendix

A. Miscellaneous Proofs

PROOF OF PROPOSITION 1:

We start with a continuous time infinite horizon, average cost formulation. Let $S(t) = (Q(t), Z_C(t), Z_{SC}(t))$ be the state of the system at time t. $Q \geq 0$ is the number of Critical patients waiting for service; $Z_C \in [0, B_I]$ is the number of

Critical patients in service; and, $Z_{SC} \in [0, B_S + B_I - Z_C]$ is the number of Semi-critical patients in service (in either the SDU or ICU). The objective is to determine policy $u(t) \in \{\mathsf{BUMP}, \mathsf{WAIT}\}$:

$$\lim_{n\to\infty} \frac{1}{E[t_n]} E\left[\int_0^{t_n} g(S(t), u(t))dt\right]$$

where t_n is the time of the n^{th} transition and $g(s,u)$ is the cost rate which is given as:

$$g(s,u) = w_C 1_{\{\text{abandonment of Critical patient from queue}\}} + w_{SC} 1_{\{u=\mathsf{BUMP}\}}$$

Using the uniformization technique (Bertsekas 2005), we transform our continuous time problem into a discrete time equivalent model. In order to ensure finite transition times, we introduce $Q_{\max} < \infty$ as the maximum number of patients waiting in the queue. If a new ICU state patient arrives and there are already $Q_{\max}$ patients in the queue, it will be lost at no cost. Thus, the maximum possible transition rate is $v = \lambda + Q_{\max}\theta + B_I \max\{\mu_C, \mu_{SC}\} + B_S\mu_{SC}$. We will see our results are independent of $Q_{\max}$. We can write the Bellman equation for this optimization problem as:

$$\gamma + h(S) = \min_{u(t)\in\{\mathsf{BUMP},\mathsf{WAIT}\}} \left[g(S,u) + \sum_{S'} p_{SS'}(u) h(S') \right]$$

where γ is the optimal average cost per stage, h is the differential cost, and $\{p\}$ is the transition matrix.

If there is room for a new Critical patient, the decision is straight-forward: admit the new patient. The challenge arises when a new Critical patient arrives and all ICU beds are currently occupied. If at least one of the beds is occupied by a Semi-critical patient, the decision is whether to admit the Critical patient and BUMP the Semi-critical patient out in the process or to wait and see if an ICU bed becomes available later, risking the potential of the Critical patient abandoning from the queue. Note that because bumping a patient when there is no state-transition is suboptimal, we do not explicitly include it as an option here.

- If $Q = 0$

  —If $Z_C = B_I$ or $Z_C + Z_{SC} = B_I + B_S$

$$\begin{aligned}\gamma + h(0, Z_C, Z_{SC}) = \frac{1}{v}[&\lambda h(1, Z_C, Z_{SC}) + Z_C\mu_C[ph(0, Z_C - 1, Z_{SC} + 1) + (1-p)h(0, Z_C - 1, Z_{SC})] \\ &+ Z_{SC}\mu_{SC}h(Q, Z_C, Z_{SC} - 1) + (v - \lambda - Z_C\mu_C - Z_{SC}\mu_{SC})h(0, Z_C, Z_{SC})]\end{aligned}$$

  —If $Z_C < B_I$ and $Z_C + Z_{SC} < B_I + B_S$

$$\begin{aligned}\gamma + h(0, Z_C, Z_{SC}) = \frac{1}{v}[&\lambda h(0, Z_C + 1, Z_{SC}) + Z_C\mu_C[ph(0, Z_C - 1, Z_{SC} + 1) + (1-p)h(0, Z_C - 1, Z_{SC})] \\ &+ Z_{SC}\mu_{SC}h(Q, Z_C, Z_{SC} - 1) + (v - \lambda - Z_C\mu_C - Z_{SC}\mu_{SC})h(0, Z_C, Z_{SC})]\end{aligned}$$

- If $Q > 0$

  —If $Z_C = B_I$

$$\begin{aligned}\gamma + h(Q, Z_C, Z_{SC}) = \frac{1}{v}[&\theta Q[w_C + h(Q-1, Z_C, Z_{SC})] + \lambda h(Q_{\max} \wedge Q + 1, Z_C, Z_{SC}) \\ &+ Z_C\mu_C[ph(Q, Z_C - 1, Z_{SC} + 1) + (1-p)h(Q-1, Z_C, Z_{SC})] \\ &+ Z_{SC}\mu_{SC}h(Q, Z_C, Z_{SC} - 1) + (v - \theta - \lambda - Z_C\mu_C - Z_{SC}\mu_{SC})h(Q, Z_C, Z_{SC})]\end{aligned}$$

—If $Z_C < B_I$ (note that, necessarily, $Z_C + Z_{SC} = B_I + B_S$)

$$\gamma + h(Q, Z_C, Z_{SC}) = \min\{w_{SC} + h(Q-1, Z_C+1, Z_{SC}-1),\ \frac{1}{v}[\theta Q[w_C + h(Q-1, Z_C, Z_{SC})] + \lambda h(Q_{\max} \wedge Q + 1, Z_C, Z_{SC}) + Z_C\mu_C[ph(Q, Z_C-1, Z_{SC}+1) + (1-p)h(Q-1, Z_C, Z_{SC})] + Z_{SC}\mu_{SC}h(Q, Z_C, Z_{SC}-1) + (v - \theta - \lambda - Z_C\mu - Z_{SC}\mu_{SC})h(Q, Z_C, Z_{SC})]\}$$

We start by exploring a number of properties of the optimal differential cost function, $h(Q, Z_C, Z_{SC})$.

Lemma 1*[Monotonicity] All else being equal, the differential cost function, $h(Q, Z_C, Z_{SC})$, is non-decreasing in the number of Critical patients in service, Z_C. That is,*

$$h(Q, Z_C, Z_{SC}) \leq h(Q, Z_C + 1, Z_{SC})$$

for any Z_C and Z_{SC} such that $Z_C + 1 \leq B_I$ and $Z_C + Z_{SC} + 1 \leq B_I + B_S$.

PROOF: The proof is via a coupling approach. Consider two systems which start in states $S = (Q, Z_C + 1, Z_{SC})$ and $\bar{S} = (Q, Z_C, Z_{SC})$. Consider a coupling of the systems starting at state S and $\bar{S}$ wherein both systems witness identical sample paths for patient arrivals, service times, and abandonment for the patients they have in common. For instance, if there is an arrival in the S system, there is also one in the $\bar{S}$ system. Now assume that the system starting at S uses an optimal policy whereas the system starting at state $\bar{S}$ 'mimics' the actions of the S system (call this policy $\bar{\pi}$), so that if the S system chooses to BUMP a Semi-critical patient from the ICU, the $\bar{S}$ system will also BUMP a Semi-critical patient should such a patient be available; this will occur irrespective of whether or not this BUMP is called for (i.e. whether or not a new patient has arrived and there are no available beds). In the event that the $\bar{S}$ system needs to BUMP a Semi-critical patient and the S system does not have any Semi-critical patients to be bumped, the $\bar{\pi}$ policy will just WAIT. It is easy to see that $\bar{\pi}$ is an admissible randomized non-anticipatory policy: starting at state $\bar{S}$ one adds a 'virtual' Critical patient so that the total number of Critical and Semi-critical patients (real and virtual) in the $\bar{S}$ system are identical to the number in the S system. One then employs an optimal policy, and simulates service completion for the virtual patient (it may eventually become a Semi-critical patient, at which point its service completion is simulated as such). It is easy to see that under our coupling, $Q(S_t) \geq Q(\bar{S}_t)$, $Z_C(S_t) \geq Z_C(\bar{S}_t)$ and $Z_{SC}(S_t) \geq Z_{SC}(\bar{S}_t)$ for all t. Moreover, the cost incurred under the $\bar{\pi}$ policy is identical to that of the π^* policy, expect for any cost incurred by the virtual Critical patient (if it was bumped after it transitioned to a Semi-critical patient). Thus, we have the desired result. □

Corollary 2*[Monotonicity] All else being equal, the differential cost function, $h(Q, Z_C, Z_{SC})$, is non-decreasing in the number of Semi-critical patients in service, Z_{SC}. That is,*

$$h(Q, Z_C, Z_{SC}) \leq h(Q, Z_C, Z_{SC} + 1)$$

for any Z_C and Z_{SC} such that $Z_C + Z_{SC} + 1 \leq B_I + B_S$.

Lemma 2*[Cost of Semi-critical patient] The cost of having an additional Semi-critical patient in the system is no more than the cost of bumping that patient,* w_{SC}*. That is,*

$$h(Q, Z_C, Z_{SC} + 1) \leq w_{SC} + h(Q, Z_C, Z_{SC})$$

for any Z_C *and* Z_{SC} *such that* $Z_C \leq B_I$ *and* $Z_C + Z_{SC} + 1 \leq B_I + B_S$*.*

PROOF: This follows from properties of the optimal differential function, h. In state $(Q, Z_C, Z_{SC} + 1)$, it is an admissible policy to BUMP a Semi-critical patient, without necessarily admitting a new Critical patient. Such an action would incur cost w_{SC} and lead to the instantaneous state transition: $(Q, Z_C, Z_{SC} + 1) \rightarrow (Q, Z_C, Z_{SC})$. Let h^π denote the differential cost of this policy which is equal to $w_{SC} + h(Q, Z_C, Z_{SC})$. The result follows by the optimality of the differential cost function h, which must be minimal:

$$h(Q, Z_C, Z_{SC} + 1) \leq h^\pi(Q, Z_C, Z_{SC} + 1) = w_{SC} + h(Q, Z_C, Z_{SC})$$

□

Lemma 3*[Cost of Critical patient] The cost of having an additional Critical patient in service is no more the the cost of abandonment,* w_C*. That is,*

$$h(Q, Z_C + 1, Z_{SC}) \leq w_C + h(Q, Z_C, Z_{SC})$$

for any Z_C *and* Z_{SC} *such that* $Z_C + 1 \leq B_I$ *and* $Z_C + Z_{SC} + 1 \leq B_I + B_S$*.*

PROOF: The proof is via a coupling approach identical to that used in the proof of Lemma 1. Again, we denote the states by $S = (Q, Z_C + 1, Z_{SC})$ and $\bar{S} = (Q, Z_C, Z_{SC})$. We define a policy π' as one that 'mimics' the actions of the $\bar{S}$ system, so that if the $\bar{S}$ system chooses to BUMP a Semi-critical patient from the ICU, the S system will also BUMP a Semi-critical patient should such a patient be available. We modify this policy depending on which one of two events occurs first:

1. A Critical patient completes ICU service in the S system, but not in the $\bar{S}$ system. It is easy to see that if the patient leaves the system, then $S_t = \bar{S}_t$ for all remaining time t. Thus, the π' policy is precisely equal to the optimal policy for the $\bar{S}$ system. If the patient does not leave the system and transitions to the Semi-critical state, the π' policy will BUMP this new Semi-critical patient, incurring cost $w_{SC} \leq w_C$.

2. A Critical patient is admitted to the ICU in the $\bar{S}$ system. The π' policy will *not* admit this Critical patient so that $Z_C(S_t) = Z_C(\bar{S}_t)$ and $Z_{SC}(S_t) = Z_{SC}(\bar{S}_t)$, but $Q(S_t) = Q(\bar{S}_t) + 1$. The π' policy continues to mimic the optimal policy for the $\bar{S}$ system and ignores the extra Critical patient waiting in the queue. It will eventually abandon, resulting in cost of w_C.

Thus, under the π' policy, the differential cost incurred in the S system is at most w_C (depending on which event occurs first) plus the differential cost incurred by the $\bar{S}$ system which uses the optimal policy. The result then follows by the optimality of the differential cost function h, which must be minimal:

$$h(Q, Z_C + 1, Z_{SC}) \leq h^{\pi'}(Q, Z_C + 1, Z_{SC}) \leq w_C + h(Q, Z_C, Z_{SC})$$

□

We are now prepared to show the desired result. Our goal is to understand what is the implication on the relationship between w_C and w_{SC} of the optimality of the policy that chooses the action BUMP whenever $Z_C < B_I$ and $Q > 0$. To do so, we utilize the policy iteration algorithm (Bertsekas 2005). We consider the case where there are Semi-critical patients in the full ICU and a new Critical patient has arrived: $S = (1, Z_C, Z_{SC}), Z_C + Z_{SC} = B_I + B_S$. By assumption, it is always optimal to BUMP. We explore the implication of this optimality on w_C/w_{SC}.

We must consider the differential costs of bumping, COST_BUMP, versus waiting, COST_WAIT in a number of cases. By assumption COST_BUMP $\leq$ COST_WAIT:

1. $Z_{SC} \geq B_S + 2$: In this case, at least 2 Semi-critical patients can be bumped.

$$\begin{aligned} w_{SC} \; + \; & \frac{1}{v}[\lambda h(1, Z_C + 1, Z_{SC} - 1) + (Z_C + 1)\mu_C[ph(0, Z_C, Z_{SC}) + (1-p)h(0, Z_C, Z_{SC} - 1)] \\ & + \; (Z_{SC} - 1)\mu_{SC} h(0, Z_C + 1, Z_{SC} - 2) + (v - \lambda - (Z_C + 1)\mu_C - (Z_{SC} - 1)\mu_{SC})h(0, Z_C + 1, Z_{SC} - 1)] \\ & \leq \frac{1}{v}[\theta[w_C + h(0, Z_C, Z_{SC})] + \lambda h(2, Z_C, Z_{SC}) + Z_C\mu_C[ph(1, Z_C - 1, Z_{SC} + 1) + (1-p)h(0, Z_C, Z_{SC})] \\ & \qquad + Z_{SC}\mu_{SC} h(0, Z_C + 1, Z_{SC} - 1) + (v - \theta - \lambda - Z_C\mu_C - Z_{SC}\mu_{SC})h(1, Z_C, Z_{SC})] \end{aligned}$$

Because we know that bumping is optimal, we have that $h(1, Z_C, Z_{SC}) = w_{SC} + h(0, Z_C + 1, Z_{SC} - 1)$; $h(2, Z_C, Z_{SC}) = 2w_{SC} + h(0, Z_C + 2, Z_{SC} - 2)$ and $h(1, Z_C - 1, Z_{SC} + 1) = w_{SC} + h(0, Z_C, Z_{SC})$. This gives us:

$$\begin{aligned} w_{SC} \; + \; & \frac{1}{v}[\lambda[w_{SC} + h(0, Z_C + 2, Z_{SC} - 2)] + (Z_C + 1)\mu_C[ph(0, Z_C, Z_{SC}) + (1-p)h(0, Z_C, Z_{SC} - 1)] \\ & + \; (Z_{SC} - 1)\mu_{SC} h(0, Z_C + 1, Z_{SC} - 2) + (v - \lambda - (Z_C + 1)\mu_C - (Z_{SC} - 1)\mu_{SC})h(0, Z_C + 1, Z_{SC} - 1)] \\ & \leq \frac{1}{v}[\theta[w_C + h(0, Z_C, Z_{SC})] + \lambda[2w_{SC} + h(0, Z_C + 2, Z_{SC} - 2)] + Z_C\mu_C[p[w_{SC} + h(0, Z_C, Z_{SC})] + (1-p)h(0, Z_C, Z_{SC})] \\ & \qquad + Z_{SC}\mu_{SC} h(0, Z_C + 1, Z_{SC} - 1) + (v - \theta - \lambda - Z_C\mu_C - Z_{SC}\mu_{SC})[w_{SC} + h(0, Z_C + 1, Z_{SC} - 1)]] \end{aligned}$$

This gives us:

$$\begin{aligned} (\theta + Z_C\mu_C(1-p) + Z_{SC}\mu_{SC})w_{SC} + (Z_C + 1)\mu_C[ph(0, Z_C, Z_{SC}) \; + \; & (1-p)h(0, Z_C, Z_{SC} - 1)] \\ + (Z_{SC} - 1)\mu_{SC} h(0, Z_C + 1, Z_{SC} - 2) \leq \; & \theta w_C + (\theta + Z_C\mu_C)h(0, Z_C, Z_{SC}) \\ & + (\mu_C - \theta + (Z_{SC} - 1)\mu_{SC})h(0, Z_C + 1, Z_{SC} - 1) \end{aligned}$$

Using Lemma 2, we have

$$\begin{aligned} (\theta + Z_C\mu_C(1-p) + Z_{SC}\mu_{SC})w_{SC} + (Z_C + 1)\mu_C[ph(0, Z_C, Z_{SC}) \; + \; & (1-p)[-w_{SC} + h(0, Z_C, Z_{SC})]] \\ + (Z_{SC} - 1)\mu_{SC}[-w_{SC} + h(0, Z_C + 1, Z_{SC} - 1)] \leq \; & \theta w_C + (\theta + Z_C\mu_C)h(0, Z_C, Z_{SC}) \\ & + (\mu_C - \theta + (Z_{SC} - 1)\mu_{SC})h(0, Z_C + 1, Z_{SC} - 1) \end{aligned}$$

which gives us

$$(\theta - \mu_C(1-p) + \mu_{SC})w_{SC} \; \leq \; \theta w_C + (\theta - \mu_C)h(0, Z_C, Z_{SC}) + (\mu_C - \theta)h(0, Z_C + 1, Z_{SC} - 1)$$

Using Lemma 3 and 1, we have

$$\begin{aligned} & \mu_C \geq \theta \; : \; (\theta - \mu_C(1-p) + \mu_{SC})w_{SC} \leq \theta w_C + (\theta - \mu_C)h(0, Z_C, Z_{SC}) + (\mu_C - \theta)[w_C + h(0, Z_C, Z_{SC})] \\ & \mu_C < \theta \; : \; (\theta - \mu_C(1-p) + \mu_{SC})w_{SC} \leq \theta w_C + (\theta - \mu_C)[w_C + h(0, Z_C + 1, Z_{SC} - 1)] + (\mu_C - \theta)h(0, Z_C + 1, Z_{SC} - 1) \end{aligned}$$

So that the optimality of bumping implies that:

$$\frac{w_C}{w_{SC}} \geq \begin{cases} \frac{\theta - \mu_C(1-p) + \mu_{SC}}{\mu_C}, \; \mu_C \geq \theta; \\ \frac{\theta - \mu_C(1-p) + \mu_{SC}}{2\theta - \mu_C}, \; \mu_C < \theta. \end{cases} \implies \frac{w_C}{w_{SC}} \geq \frac{\theta - \mu_C(1-p) + \mu_{SC}}{\max\{\mu_C, 2\theta - \mu_C\}}$$

2. $Z_{SC} = B_S + 1$: In this case, at most 1 Semi-critical patient can be bumped.

$$\begin{aligned} w_{SC} \; + \; & \frac{1}{v}[\lambda h(1, Z_C+1, B_S) + (Z_C+1)\mu_C[ph(0, Z_C, B_I+1) + (1-p)h(0, Z_C, B_S)] \\ & + \; (v - \lambda - (Z_C+1)\mu_C)h(0, Z_C+1, B_S)] \leq \frac{1}{v}[\theta[w_C + h(0, Z_C, B_S+1)] + \lambda h(2, Z_C, B_S+1) \\ & \qquad + Z_C\mu_C[ph(1, Z_C-1, B_S+2) + (1-p)h(0, Z_C, B_S+1)] + \mu_{SC}h(0, Z_C+1, B_S) \\ & \qquad + (v - \theta - \lambda - Z_C\mu_C - \mu_{SC})h(1, Z_C, B_S+1)] \end{aligned}$$

Because we know that bumping is optimal, we have that $h(1, Z_C, B_S+1) = w_{SC} + h(0, Z_C+1, B_S)$; $h(2, Z_C, B_S+1) = w_{SC} + h(1, Z_C+1, B_S)$ and $h(1, Z_C-1, B_S+2) = w_{SC} + h(0, Z_C, B_S+1)$. This gives us:

$$\begin{aligned} w_{SC} \; + \; & \frac{1}{v}[\lambda h(1, Z_C+1, B_S) + (Z_C+1)\mu_C[ph(0, Z_C, B_S+1) + (1-p)h(0, Z_C, B_S)] + (v - \lambda - (Z_C+1)\mu_C)h(0, Z_C+1, B_S)] \\ & \leq \frac{1}{v}[\theta[w_C + h(0, Z_C, B_S+1)] + \lambda[w_{SC} + h(1, Z_C+1, B_S)] + Z_C\mu_C[p[w_{SC} + h(0, Z_C, B_S+1)] + (1-p)h(0, Z_C, B_S+1)] \\ & \quad + \mu_{SC}h(0, Z_C+1, B_S) + (v - \theta - \lambda - Z_C\mu_C - \mu_{SC})[w_{SC} + h(0, Z_C+1, B_S)] \end{aligned}$$

This gives us:

$$\begin{aligned} w_{SC} \; + \; & \frac{1}{v}[\lambda h(1, Z_C+1, B_S) + (Z_C+1)\mu_C[ph(0, Z_C, B_S+1) + (1-p)h(0, Z_C, B_S)] + (v - \lambda - (Z_C+1)\mu_C)h(0, Z_C+1, B_S)] \\ & \leq \frac{1}{v}[\theta w_C + (v - \theta - Z_C\mu_C(1-p) - \mu_{SC})w_{SC} + \theta h(0, Z_C, B_S+1) + \lambda h(1, Z_C+1, B_S) + Z_C\mu_C h(0, Z_C, B_S+1) \\ & \quad + (v - \theta - \lambda - Z_C\mu_C)h(0, Z_C+1, B_S)] \end{aligned}$$

This gives us:

$$\begin{aligned} (\theta + Z_C\mu_C(1-p) + \mu_{SC})w_{SC} \; + \; & (Z_C+1)\mu_C[ph(0, Z_C, B_S+1) + (1-p)h(0, Z_C, B_S)] \\ & \leq \theta w_C + \theta h(0, Z_C, B_S+1) + Z_C\mu_C h(0, Z_C, B_S+1) + (\mu_C - \theta)h(0, Z_C+1, B_S) \end{aligned}$$

Using Lemma 2, we have

$$\begin{aligned} (\theta + Z_C\mu_C(1-p) + \mu_{SC})w_{SC} \; + \; & (Z_C+1)\mu_C h(0, Z_C, B_S+1) - (Z_C+1)\mu_C(1-p)w_{SC} \\ & \leq \theta w_C + \theta h(0, Z_C, B_S+1) + Z_C\mu_C h(0, Z_C, B_S+1) + (\mu_C - \theta)h(0, Z_C+1, B_S) \end{aligned}$$

which gives us

$$(\theta - \mu_C(1-p) + \mu_{SC})w_{SC} \; + \; (\mu_C - \theta)h(0, Z_C, B_S+1) \leq \theta w_C + (\mu_C - \theta)h(0, Z_C+1, B_S)$$

Using Lemmas 1-3 and Corollary 2, we have

$$\begin{aligned} & \mu_C \geq \theta : (\theta - \mu_C(1-p) + \mu_{SC})w_{SC} + (\mu_C - \theta)h(0, Z_C, B_S+1) \leq \theta w_C + (\mu_C - \theta)[w_C + h(0, Z_C, B_S+1)] \\ & \mu_C < \theta : (\theta - \mu_C(1-p) + \mu_{SC})w_{SC} + (\mu_C - \theta)[w_C + h(0, Z_C+1, B_S)] \leq \theta w_C + (\mu_C - \theta)h(0, Z_I+1, B_S) \end{aligned}$$

So that the optimality of bumping implies that:

$$\frac{w_C}{w_{SC}} \geq \begin{cases} \frac{\theta - \mu_C(1-p) + \mu_{SC}}{\mu_C}, \; \mu_C \geq \theta; \\ \frac{\theta - \mu_C(1-p) + \mu_{SC}}{2\theta - \mu_C}, \; \mu_C < \theta. \end{cases} \implies \frac{w_C}{w_{SC}} \geq \frac{\theta - \mu_C(1-p) + \mu_{SC}}{\max\{\mu_C, 2\theta - \mu_C\}}$$

□

PROOF OF PROPOSITION 2:

1. Suppose that $\limsup_{N\to\infty} \frac{\lambda(r_I\mu_C p + r_S\mu_{SC})}{N r_I r_S \mu_C \mu_{SC}} \leq 1$. By the equivalence between (4) and (5), there exists a sequence of bed allocations $(B_I, B_S) := (B_I^N, B_S^N)$ such that $\limsup_{N\to\infty} [\max\{\rho_C(B_I, B_S), \rho_T(B_I, B_S)\}] \leq 1$. Since the number of Critical patients in the ICU behaves like an $M/M/B_I + M$ queue, with traffic intensity $\rho_C \leq 1$, we have that, by (Garnett et al. 2002, Theorem 4) with $\beta > -\infty$, the rate of abandonment is equal to $[\lambda - B_I\mu_C]^+ + o(N) = o(N)$.

As for the Semi-Critical patients, the arrival rate into this state is equal to $p\mu_C EZ_C$, where EZ_C stands for the expected steady-state number of ICU beds that are occupied by critical patients. The service rate is equal to $(B_S + B_I - EZ_C)\mu_{SC}$. By Little's law, $EZ_C = (\lambda - o(N))/\mu_C$, where the $o(N)$ term is contributed by the Critical patient abandonment rate. The bumping rate is hence equal to

$$[p\mu_C EZ_C - (B_S + B_I - EZ_C)\mu_{SC}]^+ = \mu_{SC}\left[\mu_T(\lambda + o(N)) - (B_S + B_I)\right]^+ = o(N).$$

2. Suppose now that $\liminf_{N\to\infty} \frac{\lambda(r_I\mu_C p + r_S\mu_{SC})}{N r_I r_S \mu_C \mu_{SC}} > 1$. In this case, we have that for any sequence of bed allocation (B_I, B_S), either $\liminf_{N\to\infty} \rho_C > 1$, or $\liminf_{N\to\infty} \rho_T > 1$, or both. If $\limsup_{N\to\infty} \rho_C > 1$, then by (Garnett et al. 2002, Theorem 4) with $\beta = -\infty$, we have that the rate of abandonment is $O(N)$. Else, if $\limsup_{N\to\infty} \rho_C \leq 1$, then by 1. the abandonment is $o(N)$. Therefore, the bumping rate is again equal to

$$[p\mu_C EZ_C - (B_S + B_I - EZ_C)\mu_{SC}]^+ = \mu_{SC}\left[\mu_T(\lambda + o(N)) - (B_S + B_I)\right]^+ = O(N).$$

If neither of these cases applies, the argument works analogously when considering converging subsequences such that either $\lim_{N\to\infty} \rho_C > 1$ or $\lim_{N\to\infty} \rho_C \leq 1$.

□

PROOF OF THEOREM 1: Suppose that (5) holds in the limit. That is, assume that

$$\liminf_{N\to\infty} \frac{\lambda(r_I\mu_C p + r_S\mu_{SC})}{N r_I r_S \mu_C \mu_{SC}} > 1. \tag{18}$$

Additionally, assume that the system operates in the ID regime and that (9) and (10) hold. Let $\hat{U}^N := \hat{Z}_C^N + \hat{Z}_{SC}^N$. And suppose that $\hat{U}^N(0) = 0$. It is our goal to show that for any $\epsilon > 0$,

$$P\left\{\inf_{0\leq t\leq 1} \hat{U}^N(t) < -\epsilon\right\} \to 0, \text{ as } N \to \infty.$$

The proof follows along the lines of Reiman (1984). Fix $\epsilon > 0$ and let

$$\tau_N = \inf\{t \geq 0;\ \hat{U}^N(t) < -\epsilon\} \text{ and } \tau'_N = \sup\{t \leq \tau_N;\ \hat{U}^N(t) \geq -\epsilon/2\}.$$

During $[\tau'_N, \tau_N]$ there are empty beds in either the ICU or SDU (or both), so no bumping will occur. In particular, during this interval

$$Z_C^N(t) + Z_{SC}^N(t) = Z_C^N(\tau'_N) + Z_{SC}^N(\tau'_N) + A^N(\tau'_N, t) + \Phi^N(\tau'_N, t) - D_C^N(\tau'_N, t) - D_{SC}^N(\tau'_N, t),$$

where, for $s < t$, $A^N(s,t)$ is the number of critical patients that arrived directly into the ICU (and did not wait in queue) during $(s,t]$, $\Phi^N(s,t)$ is the number of critical patients arrivals into the ICU from the queue in $(s,t]$. Also,

$D_C^N(s,t]$ is the number of critical patients who have completed their stay in the ICU and did not switch to a semi-critical state during $(s,t]$. Finally, $D_{SC}^N(s,t)$ is the number of service completions of semi-critical patients in $(s,t]$. More specifically, let S_i, $i=1,2,3$ be independent unit Poisson processes, then

$$\begin{aligned} A^N(s,t)+\Phi^N(s,t) &= S_1\left(\int_s^t \lambda 1_{\{Z_C^N(r)<B_I\}}+\mu_C Z_C^N(r)1_{\{Z_C^N(r)=B_I,\ Q>0\}}\cdot dr\right)=(t-s)\cdot(\lambda+o(\lambda)),\\ D_C^N(s,t) &= S_2\left((1-p)\mu_C\int_s^t Z_C^N(r)\cdot dr\right)=(t-s)\cdot((1-p)\lambda+o(\lambda)),\\ D_{SC}^N &= S_3\left(\mu_{SC}\int_s^t Z_{SC}^N(r)\cdot dr\right)\le S_3\left(\mu_{SC}\int_s^t\left(B_S^N+B_I^N-Z_C^N(r)\right)\cdot dr\right)\\ &= (t-s)\cdot\left(\frac{\mu_{SC}r_S}{r_I}\left(Nr_I-\frac{\lambda}{\mu_C}\right)+o(\lambda)\right). \end{aligned} \tag{19}$$

Recall that the ICU is operating in the QED regime with respect to Critical patients; therefore, $\mu_C Z_C^N=\lambda+o(\lambda)$ and $B_I-Z_C^N=o(\lambda)$. Finally, we have:

$$\begin{aligned} P\left\{\inf_{0\le t\le 1}\hat{U}^N(t)<-\epsilon\right\} &\le P\left\{\inf_{0\le s\le t\le 1}\tfrac{A^N(s,t)+\Phi^N(s,t)-D_C^N(s,t)-D_{SC}^N(s,t)}{\sqrt{\lambda}}<-\epsilon/2\right\}\\ &= P\left\{\inf_{0\le s\le t\le 1}\tfrac{\frac{t-s}{\mu_C r_I}\cdot(\lambda\cdot(pr_I\mu_C+\mu_{SC}r_S)-\mu_{SC}\mu_C r_S r_I N)+o(\sqrt{\lambda})}{\sqrt{\lambda}}<-\epsilon/2\right\}\\ &\to 0,\ \text{ by (18).} \end{aligned}$$

□

B. A System with Readmissions

We now consider a stylized model which explicitly accounts for patient readmissions. We consider the following setup:

1. N nurses are reserved to treat first-time arrivals. These nurses can be allocated amongst B_I ICU and B_S SDU beds as desired. Any reference to system state will be understood to correspond to the number of *first-time* Critical and Semi-critical patients.

2. A first-time Critical patient who abandons from the ICU queue returns for ICU treatment with probability p_C, and has 'readmission' ICU LOS which is exponentially distributed with mean L_C^R. We let $w_C=p_C L_C^R$.

3. A first-time Semi-critical patient who is bumped from the ICU returns for ICU treatment with probability p_{SC}, and has readmission ICU LOS which is exponentially distributed with mean L_{SC}^R. We let $w_{SC}=p_{SC}L_{SC}^R$.

4. The readmission queue is served First-Come-First-Serve by C beds. Readmitted patients are treated in the ICU until they are stable enough to be transferred to the Ward, i.e. they do not go through the SDU. Readmitted patients will not abandon from the readmission queue, nor can they be bumped from the ICU.

In practice, readmitted patients tend to be much sicker, with higher mortality rates and longer LOS (Snow et al. 1985, Durbin and Kopel 1993, Chen et al. 1998). Thus, it is desirable to provide high quality care for these readmitted patients, which we capture by requiring they are treated in the ICU and cannot abandon or be bumped.

The total number of ICU beds in this setting is $C+B_I$. Given the N nurses to treat first-time arrivals, our goal is to determine the allocation of nurses to the ICU and SDU (B_I and B_S) such that we minimize the number of nurses, C/r_I, required to staff the readmission queue so that the queue remains stable. That is, if we let $\{W_n\}$ denote the waiting time the n^{th} readmitted patient experiences, we require that for any subsequence of $\{W_n\}$ there exists a sub-subsequence which converges to a random variable which is finite almost surely.

We start by examining the stability condition of the readmission queue. Let $\{\sigma_n, T_n\}$ denote the service requirement and interarrival time for readmitted patient n under some allocation of nurses between the ICU and SDU. Then the stability condition stems from a classical result of Loynes (1963), which requires that $E[\sigma_0]/E[T_0] < C$ for the readmission queue to be stable. We let π denote the steady-state distribution of the first-time patients, where the state $S = (Q, Z_C, Z_{SC})$. For notational compactness, we suppress the dependence of this distribution on the nurse allocation. Relating the stability condition to our original problem setting of Section 2 we have:

Lemma 4*The readmission queue is stable if and only if:*

$$\limsup_{T\to\infty} \frac{1}{T}\left[\sum_{n=0}^{\infty} [w_C a_n + w_{SC} b_n]\, 1_{\{t_n \le T\}}\right] < C$$

PROOF: As in the proof of Proposition 1, and focusing first on first-time patients, we let $v = \lambda + Q_{\max}\theta + B_I \max\{\mu_C, \mu_{SC}\} + B_S\mu_{SC}$ be the maximum possible transition rate. We can determine the probability that the next event is a Critical patient abandonment or a Semi-critical patient bumping:

$$P(\text{Abandonment or Bumping}) = \sum_S \pi_S \frac{[\theta Q + \lambda 1_{\{Z_C = B_I + B_S - Z_{SC}, Z_{SC} > B_S\}}]}{v} \tag{20}$$

If an abandonment (bumping) occurs, the patient's readmission ICU LOS is L_C^R (L_{SC}^R) with probability p_C (p_{SC}) and 0 otherwise; that is, we formally assume that *all* the abandoning and bumped patients are readmitted, but some of them have an ICU LOS of 0. The interarrival time of events is exponentially distributed with rate v. Additionally, the number of events until an abandonment or bumping is Geometrically distributed with mean $1/P(\text{Abandonment or Bumping})$. Thus, the interarrival time of readmitted patients is:

$$E[T_0] = \frac{v}{\sum_S \pi_S \left[\theta Q + \lambda 1_{\{Z_C = B_I + B_S - Z_{SC}, Z_{SC} > B_S\}}\right]} \tag{21}$$

Finally, the expected service requirement of readmitted patients is:

$$E[\sigma_0] = \frac{\sum_S \pi_S \left[\frac{\theta Q}{v} w_C + \frac{\lambda 1_{\{Z_C = B_I + B_S - Z_{SC}, Z_{SC} > B_S\}}}{v} w_{SC}\right]}{\sum_S \pi_S \left[\theta Q + \lambda 1_{\{Z_C = B_I + B_S - Z_{SC}, Z_{SC} > B_S\}}\right]} \tag{22}$$

Combining equations (21) and (22) gives the desired stability condition.

$$\begin{aligned} \frac{E[\sigma_0]}{E[T_0]} &= \sum_S \pi_S \left[\theta Q w_C + \lambda 1_{\{Z_C = B_I + B_S - Z_{SC}, Z_{SC} > B_S\}} w_{SC}\right] \\ &= \limsup_{T\to\infty} \frac{1}{T}\left[\sum_{n=0}^{\infty} [w_C a_n + w_{SC} b_n]\, 1_{\{t_n \le T\}}\right] < C \end{aligned} \tag{23}$$

□

We can see that given an allocation of nurses, the readmission queue is stable when there are enough beds C to serve the readmission load. By specifying the costs of abandonment and bumping to be the readmission load associated with these events, the stability condition is to have enough beds C such that it is greater than the optimal average abandonment and bumping costs. Thus, to minimize the number of beds (and, subsequently, nurses) necessary to stabilize the readmission queue, the N nurses dedicated to first-time patients should be allocated such that the average abandonment plus bumping cost is minimized, as captured in equation (2).