2012

An Econometric Analysis of Patient Flows in the Cardiac Intensive Care Unit

Diwas Singh Kc
University of Pennsylvania

Christian Terwiesch
University of Pennsylvania

Recommended Citation

Kc, D., & Terwiesch, C. (2012). An Econometric Analysis of Patient Flows in the Cardiac Intensive Care Unit. *Manufacturing & Service Operations Management, 14* (1), 50-65. http://dx.doi.org/10.1287/msom.1110.0341

An Econometric Analysis of Patient Flows in the Cardiac Intensive Care Unit

Abstract

This paper explores the rationing of bed capacity in a cardiac intensive care unit (ICU). We find that the length of stay for patients admitted to the ICU is influenced by the occupancy level of the ICU. In particular, a patient is likely to be discharged early when the occupancy in the ICU is high. This in turn leads to an increased likelihood of the patient having to be readmitted to the ICU at a later time. Such "bounce-backs" have implications for the overall ICU effective capacity—an early discharge immediately frees up capacity, but at the risk of a (potentially much higher) capacity requirement when the patient needs to be readmitted. We analyze these capacity implications, shedding light on the question of whether an ICU should apply an aggressive discharge strategy or if it should follow the old quality slogan and "do it right the first time." By comparing the total capacity usage for patients who were discharged early versus those who were not, we show that an aggressive discharge policy applied to patients with lower clinical severity levels frees up capacity in the ICU. However, we find that an increased number of readmissions of patients with high clinical severity levels occur when the ICU is capacity constrained, thereby effectively reducing peak bed capacity.

Keywords

patient flow, health-care operations, capacity management, rework

Disciplines

Econometrics | Other Public Health

An Econometric Analysis of Patient Flows in the Cardiac ICU

Diwas Singh Kc
Goizueta Business School, Emory University
dkc@emory.edu

Christian Terwiesch
The Wharton School, University of Pennsylvania
terwiesch@wharton.upenn.edu

This paper explores the rationing of bed capacity in a cardiac intensive care unit (ICU). We find that the length of stay for patients admitted to the ICU is influenced by the occupancy level of the ICU. In particular, a patient is likely to be discharged early when the occupancy in the ICU is high. This in turn leads to an increased likelihood of the patient having to be readmitted to the ICU at a later time. Such "bounce-backs" have implications for the overall ICU effective capacity – an early discharge immediately frees up capacity, but at the risk of a (potentially much higher) capacity requirement when the patient needs to be readmitted. We analyze these capacity implications, shedding light on the question if an ICU should apply an aggressive discharge strategy or if it should follow the old quality slogan and "do it right the first time." By comparing the total capacity usage for patients who were discharged early versus those who were not, we show that an aggressive discharge policy applied to patients with lower clinical severity levels frees up capacity in the ICU. However, we find that an increased number of readmissions of patients with high clinical severity levels occur when the ICU is capacity constrained, thereby effectively reducing peak bed capacity.

1 Introduction

Numerous studies (Hall 2006, IOM 2007) have found that resource constraints often plague patient flows in hospitals in general and cardiac care in particular. A resource constraint at any one stage in the care process can lead to delays, congestion and overall reduction in patient throughput for the hospital. For example, in the cardiac care process, the surgical Intensive Care Unit (ICU) has often been identified as the process bottleneck. The ICU is an expensive resource with the cost of patient care being multiple times higher than in a regular ward (see e.g. Henning et al 1987). Consequently, many ICU's operate at high levels of occupancy, leading to increased waiting times upstream of the ICU and an overall reduction in patient throughput (see also McConnell et al 2005).

Given the scarce ICU capacity, hospitals are often forced to ration the available ICU beds. This means that when the ICU reaches its full occupancy, the healthiest (in relative terms) patient gets discharged, more or less independent of their absolute health condition. While such early discharges clearly increase patient throughput in the short-term, they have the potential to lead to medical complications and to increase the likelihood that a patient has to revisit the ICU in the future. Readmission of patients following reduced ICU length of stay has been a topic of interest in the health care literature, particularly following the rise of managed care over the last two decades. However, the subsequent impact of same-stay readmissions on the operational and financial performance of hospitals has not been examined.

It is this interplay between the medical variables (which determine the ICU length of stay of a patient) and the operational variables such as ICU occupancy and capacity rationing, that is at the heart of this paper. Based on medical records, billing records, and detailed operational flow data of 1365 cardiothoracic patients in a large US teaching hospital, we develop an econometric model of patient recovery, discharge from the ICU and potential readmission to the ICU. This allows us to make the following four contributions.

First, we estimate the impact of ICU occupancy on the ICU length of stay of a patient. This allows us to study the discharge pattern of the ICU. We show that a patient who is discharged from a busy ICU has an average length of stay that is 16% shorter compared to a patient (with similar medical conditions) that is discharged at a lower level of occupancy. Thus, the ICU rations its capacity when reaching full occupancy by discharging some patients early.

Second, we show that this capacity rationing behavior has serious medical implications. Specifically, we show that a patient who is discharged early has an increased likelihood of being readmitted to the ICU at some later time (creating a so called “bounce-back”) within the same hospital stay. Moreover, we find that patients have a dramatically longer length of stay in the ICU when they are admitted to the ICU for a second stay.

Third, we analyze the capacity implications of the hospital’s discharge pattern. An early discharge immediately frees up ICU capacity, but at the risk of a (much higher) capacity requirement upon readmission. By comparing the total capacity usage for patients who were discharged early versus those who were not, we show that an aggressive discharge policy frees up capacity in the ICU for lower severity patients. However, we find that an increased number of readmissions of high-severity patients occur when the ICU is capacity constrained, thereby effectively reducing peak bed capacity.

The remainder of this paper is organized as follows. We first discuss the recovery process of cardiac patients with a focus on the ICU followed by a review of the relevant literature. We then develop our models and provide a description of our data collection. In section 6, we present our estimation strategy and econometric specifications. Finally we present our results and conclude with a discussion of the implications of our findings.

2 Process Description

After a cardiac patient is admitted to the hospital, a number of pre-surgery diagnostic tests are conducted and the patient is prepared for surgery. The set of these activities is collectively referred to as the pre-operative stage. Immediately following surgery, the patient is taken to the intensive care unit (ICU). At this point, the patient is typically unconscious and is on breathing assistance via a ventilator. In the immediate post-operative stage, various medications are administered to sedate and stabilize the patient. The patient is under constant monitoring, often requiring a one-to-one patient-to-nurse ratio during the first twelve hours following surgery. A physician is also immediately available to attend to any complications that may arise.

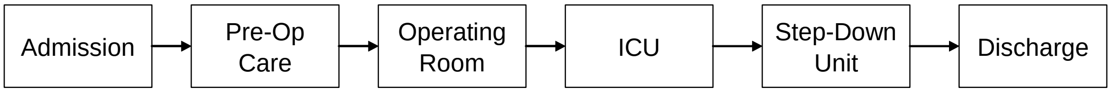

Figure 1: Patient Flow Process

Following discharge from the ICU, the patient is taken to a step-down unit, or “the floor”. The step-down unit has reduced intensity of treatment and monitoring. For example, the patient may no longer be on heavy medication and is typically no longer on ventilator support. Also, an attending physician may not be available immediately day and night.

The ICU is an expensive resource (Hall 2006), involving costly equipment and full-time dedicated staff. On the other hand, the step-down unit does not require as high a level of patient monitoring or equipment that is as costly. From a resource utilization point of view, it is thus less costly to have a patient in the step-down unit than in the ICU. As in many process flows, the most costly resource is usually the bottleneck. In this case, the ICU is capacity constrained, whereas the less costly step-down unit generally has excess capacity.

Therefore, in order to increase throughput from the system, the hospital needs to free up bed capacity in the ICU. If serious complications requiring increased level of care and monitoring arise while the patient is in the step-down unit, the patient is readmitted to the ICU. For the majority of patients, however, no significant complications arise, and after a period of stay in the step down unit, the patient is discharged from the hospital. A small minority of emergency patients may also be admitted directly into the ICU from the operating room without going through the pre-operative care process. The rest of the care process is similar to that for the elective care patients.

When the ICU is full, the decision maker is confronted with the following dilemma: whether to discharge an existing patient early, or to cancel the surgery for a scheduled patient (since this patient would immediately require an ICU bed). Both options are undesirable – discharging a patient early could lead to a bounce-back, whereas disrupting the surgery schedule is inconvenient and perhaps medically risky for the scheduled patient. It is thus theoretically possible to use procedure cancellations as a way to match supply with demand. However, from discussions with doctors from our research site, we learned that cancellations of surgical cases as a result of ICU occupancy are rare. Consequently, the only process flow control to deal with this variability in inflow and medically required ICU length of stay is the discharge decision.

This process of cardiac surgery and recovery is common across hospitals, including the hospital underlying this research study. The hospital is ranked as one of the top hospitals in cardiac surgery in the US and performs over 1200 cardiothoracic surgeries a year, including complex procedures such as heart transplants. Given this size, the hospital has an ICU dedicated to cardiothoracic care. The ICU has a total of 18 beds. Most cardiothoracic surgery patients spend between 1 and 5 days in the first visit to the ICU, and about 14% of the patients require readmission to the ICU within the same hospital stay.

3 Literature Review

Capacity planning in healthcare delivery has been an active and fruitful area for research in Management Science and Operations Research. Previous studies have looked at strategic decisions such as the sizing of capacity including beds, equipment and number of staff (e.g. Kwak and Lee 1997, Green and Meissner 2002, Huang 1995, Green et al 2006) as well as tactical decisions such as the scheduling of procedures (see e.g. Gerchak et al 1996). This stream of research is quite extensive and we refer the readers to Smith-Daniels et al (1988) and Green (2004) for more comprehensive overviews. Queueing theory is one of the most commonly used analytical methods to describe care processes because of the stochastic nature of demand and service in healthcare, as well as the ability to estimate performance measures such as waiting time, queue length, or turn-away probability.

A common assumption in the previous body of literature is that the service rate is drawn from a probability distribution (usually exponential) that is exogenous to and independent of the current state of the system, including the number of people waiting in line. However, there exist several papers, some analytical and some empirical, that challenge this independence assumption and postulate that resources should increase their service rate when the load on the system is high. This literature of dynamic queueing control started with a set of analytical models, including work by Bertsekas (2000), George and Harrison (2001), Stidham and Weber (1989) and Crabill (1972), that derive the optimal service rate that balances the costs of acceleration with the costs of waiting times. Collectively, this body of literature shows that it is optimal for resources to accelerate as the length of the queue increases.

From an empirical perspective, using data from two distinct healthcare services, patient transport and cardiothoracic surgery, Kc and Terwiesch (2009) validate that workers adapt to increasing levels of load in the system by increasing their service rate. The authors also show that such temporary service rate increases are not sustainable (i.e., workers become fatigued) and can have potentially serious quality implications. The results obtained by Kc and Terwiesch complement a set of prior lab experiments conducted by Schultz et al (1998, 1999) that establish that workers in an assembly line adjust their service rates in response to the amount of work in process inventory between the workers. Powell et al (2004) find that such worker-level state-dependent behavior has implications for the entire process flow.

Just like the literature reviewed above, the theory underlying our work also is in the tradition of optimal queueing control and we empirically investigate the relationship between system load and service rate. However, what sets the present paper apart from the prior literature is its focus on rework. We consider a service setting in which the server has the option to “rush” customers currently in service. While such rushing immediately increases service capacity, it comes at the risk that the customer has to be reworked at a later point in time. This, potentially much longer, rework has a negative impact on service capacity, leaving the server with a decision of “rushing now and reworking later” or “doing it right the first time.”

The quality management literature has taken a firm point on this decision. Rework is seen as one of the seven sources of waste initially observed by the Japanese production movement (see e.g. Ohno 1988). For example, in their work benchmarking automotive production plants across the world, Womack et al (1990) found that GM's Framingham plant spent over 40 hours on the average vehicle, including the rework of 1.3 defects per vehicle, while Toyota in its Takaoka plant only needed 18 hours, largely reflecting substantially less time wasted on rework. In his description of Toyota's production system, Liker (2004) emphasizes the importance of getting quality right immediately as opposed to relying on rework downstream in the assembly line. This quality paradigm has also been analyzed in healthcare operations. Tucker (2004) finds in her study of nursing work that nurses waste a large part of their time reworking what either they themselves or other members of the care process got wrong earlier on.

While our paper is written to contribute to the literature in Operations Management, we also draw on the medical literature to appropriately capture various patient level severity factors and their impact on clinical outcomes as well as ICU capacity consumption. In particular, we apply a widely used risk stratification method called EuroSCORE (Nashef et al 2002, Kurki et al 2002, Toumpoulis et al 2005, Horak et al 2009) to asses the impact of an early discharge on the likelihood of a bounce-back to the ICU. We also build on previous work in the medical literature to develop models of patient recovery (Peake et al 2006).

The impact on quality of care following reduced length of stay has a history in the health care literature. For example, Strauss et al (1986) find that patients tend to be discharged earlier when the ICU is more crowded. They find that the reduced length of stay due to bed availability has no effect on the rate of readmission to the ICU. Our study differs from Strauss et al in several ways. First, our outcome measure is in-hospital revisit rate from the ICU to the step-down unit and back. This is a very specific flow path, compared to Strauss et al, where the patient is discharged from the hospital. This distinction is important because the patient who is discharged from the hospital is generally in a more stable condition. Secondly, our length of stay is the time in the ICU. We are measuring the impact of the immediate post-surgery recovery time. Third, our focus is on the impact of occupancy-induced length of stay reduction on revisit rates. Occupancy allows us to perform a pseudo-randomization of patients to treatment and control groups, and as a result, handle unobserved heterogeneity. Bohmer et al (2002) examine the long term effects of ICU discharge policies and find that a decrease in ICU length of stay is not followed by a decrease in the long term quality of care, as measured by the post-discharge revisit rate to the hospital and the 30-day post-discharge mortality rate. Similarly, Obel et al (2007) examine the weekend effect on the quality of care for patients, and find that patients who are discharged close to a weekend had a greater likelihood of mortality.

Recently, the development of rapid response teams, which specialize in transferring patients with complications back into the ICU has been an area of interest (e.g. see Chan et al. 2008, Sherner 2009, Reynolds et al 2009) for hospitals. Our findings on the reduction in available capacity due to revisits could have implications for the implementation of rapid response teams. Finally, a related stream of literature examines the identification of low-severity patients for early discharge from the ICU (see Martin et al 2005 and Swenson 1992). Our analysis shows that low severity patients differ from high severity patients in their implications for the likelihood of bounce-back, and usage of capacity. These findings suggest the merit of further research into identifying patients for early discharge.

The prior literature, however, has not examined the subsequent impact of same-stay readmissions on the operational and financial performance of hospitals. What distinguishes our study from this previous work is the availability of micro-level operational data, which allows us to link clinical decisions more closely with the immediate outcomes. This allows us to focus on examining the effect of census-based occupancy measures on the risk adjusted early discharges from the ICU and subsequent revisit to the ICU from the step down unit during the same hospital stay. These same-stay revisits to the hospital are important because they have the potential to impact overall capacity utilization, throughput and thereby revenue.

4 Hypothesis Development

The time in the ICU following surgery is primarily one of stabilization and recovery; this recovery process involves major milestones such as the removal of ventilator assistance and the weaning off from heavy medication. Patients admitted to the ICU are heterogeneous in their medical conditions. In other words, the risk of complications and the case severity vary from patient to patient. A single bypass procedure performed on a 50 year old simply has a lower level of risk than a triple bypass surgery on an 80 year old. A higher case severity typically requires a longer time for recovery, i.e. a longer stay in the ICU. Thus, any analysis of patient length of stay requires us to account for the key indicators of severity, such as age and other patient risk factors.

However, we argue that medical factors alone are not the only determinants of patient length of stay. As the occupancy level in the ICU increases, fewer beds are available to accommodate the inflow of new patients from the operating room. If the ICU is full (i.e. all ICU beds are occupied), the hospital has to ration the ICU capacity. A patient admitted out of the operating room typically needs a bed immediately. Thus, a shortage of ICU beds leads to an existing patient having to be discharged from the ICU to accommodate the "fresh" new patient. The discharged patient is the healthiest from among the current ICU population. Note though, that in absence of the high ICU occupancy, this same patient would have spent a longer time in the ICU, making the discharge a result of operational variables as opposed to medical variables alone.

To formalize this logic, we hypothesize that a patient discharged from a busy ICU will have a shorter ICU length of stay (LOS_i) than a patient discharged from a less busy ICU. That is,

$$\frac{\partial LOS_i}{\partial OCCUPANCY_i} < 0 \qquad \text{(Hypothesis 1)}$$

where $OCCUPANCY_i$ is the occupancy level at the time of discharge of patient i.

From a medical perspective, a longer length of stay increases the likelihood of a more complete patient recovery in the ICU. A patient who is discharged early for operational reasons related to ICU occupancy, i.e. who would have spent a longer time in the ICU if it were for medical considerations alone, is at an increased risk of experiencing complications outside of the ICU. Holding medical risk factors constant, we therefore postulate that a patient who is discharged early has a higher likelihood of a revisit to the ICU. In other words, the greater the length of stay, the lower the likelihood of a patient bouncing back to the ICU:

$$\frac{\partial \Pr_i}{\partial LOS_i} < 0 \qquad \text{(Hypothesis 2)}$$

where $\Pr_i$ is the probability that patient i bounces back. The alternative hypothesis is that the early discharges have no effect on the likelihood of a bounce-back. In the medical literature, we find that the bounce-back rate is often taken to be a measure of the quality of care. If one takes this perspective on the quality of care, (Hypothesis 2) suggests that an early discharge has a negative impact on the quality of care.

Finally, consider the overall capacity implications of an early discharge. If a patient is more likely to bounce back to the ICU when discharged early, the discharge decision has implications for the total ICU capacity consumption of that patient. We define the total

ICU length of stay ($TOTAL\_LOS_i$) for patient i, including the initial length of stay as well as the future length of stay associated with a potential readmission, as:

$$TOTAL\_LOS_i = LOS_i + REVISIT_i$$

where $REVISIT_i$ is the revisit length of stay. $REVISIT_i$ takes a value of zero if a patient does not revisit.

Earlier, we postulated that a shorter initial length of stay (LOS_i) increases the likelihood of a bounce-back. If this is true, there exists an optimal LOS_i that minimizes expected $TOTAL\_LOS_i$. In other words, the ICU faces a trade-off between discharging a patient early ("rushing a patient"), in which case the initial length of stay is short, but the bounce-back probability is high, and following a more conservative discharge policy ("doing it right the first time"), in which case the initial length of stay is long, but the bounce-back probability is low. However, while operating under a capacity constraint, the ICU has to discharge patients early and is not able to achieve the optimal LOS for each patient that it admits. Thus we hypothesize that an early discharge from the ICU leads to an increase in $TOTAL\_LOS$.

$$\frac{\partial TOTAL\_LOS_i}{\partial LOS_i} < 0 \qquad \text{(Hypothesis 3)}$$

The total length of stay in the ICU, however, is not the only performance measure the hospital cares about. Under the diagnosis related group (DRG) payment system, a hospital is reimbursed a fixed payment amount depending on the diagnosis for the patient, irrespective of the actual cost incurred by the hospital. A hospital thus has little financial incentive to keep a patient longer in the ICU than medically necessary. The operational performance measure that maximizes hospital revenues is thus the overall patient throughput.

By definition, the early discharge of a patient as a result of capacity rationing happens at a time when the ICU is capacity constrained. Freeing up a bed in the ICU at that time and having the patient come back at some point in the future may or may not increase the patient throughput in the ICU. In particular, for a patient who is discharged early, if a future readmission occurs at a time when there exists excess ICU capacity, the early discharge helps increase the overall patient throughput. On the other hand, if the readmission occurs when the ICU is capacity constrained, the overall patient throughput could decrease.

We assess the throughput implications of the early discharge decision (long initial length of stay versus short initial length of stay) by estimating its impact on the peak ICU capacity.

Unlike our previous analysis of patient length of stay, this peak capacity calculation explicitly considers whether the ICU is capacity constrained at the time of the bounce-back or not. Let $BUSY(t)$ be an indicator function indicating whether the ICU is busy (and thus operating at peak capacity) at time t, and let $t_{i,initial}$ and $t_{i,revisit}$ be the starting times of initial visit and readmission of patient i to the ICU respectively. The peak capacity usage for a patient is thus given by:

$$TOTAL\_PEAK\_LOS_i = \int_{t_{i,initial}}^{t_{i,initial}+LOS_i} BUSY(t)dt + \int_{t_{i,revisit}}^{t_{i,revisit}+REVISIT_i} BUSY(t)dt$$

If an aggressive discharge decision decreases the total peak bed capacity consumption, it helps to increase the overall patient throughput in the ICU. In other words, if the ICU could improve its effective capacity, the hospital would be able to schedule more OR procedures. Put to the extreme, a patient staying in a half empty ICU for one week has a lower peak bed capacity consumption than a patient who just spends one day in an ICU that is full. The total peak capacity consumption thus depends on the incidence of a revisit, occupancy during the revisit, and the peak capacity saved from discharging the patient early. In practice, even though a patient bounces back, the revisit may occur when the ICU is not busy. Thus total peak capacity usage may decrease because the peak capacity saved from an early discharge more than compensate for future peak capacity usage. In other words "rushing and revisiting" can be used to smooth demand for peak bed capacity, under which lower-priority demand is satisfied later. Thus, we hypothesize that such early discharges of patients from the ICU reduce peak capacity consumption.

$$\frac{\partial TOTAL\_PEAK\_LOS_i}{\partial LOS_i} > 0 \qquad \text{(Hypothesis 4)}$$

Table 1 provides a brief description of the key variables that we analyse.

5 Data Collection

Our data was collected from the cardiothoracic intensive care unit at our research site. For each of the patients in our sample, we compiled data from three different sources - a medical database, a patient tracking system, and the patient billing records.

Our first source of data is the hospital's patient tracking system. This information system, NaviCare©, tracks patients and resources such as hospital beds and patient transporters in real time and supports the hospital in its patient flow management. Our research site was one of the first implementation sites of NaviCare in the country, providing us with access to patient flow data beyond what had previously been feasible. NaviCare generates timestamps for a set of events associated with the patient moving through the hospital, including the exact time the patient entered in and departed from the ICU. This information allows us to impute the length of stay for each individual patient. More importantly, since NaviCare allows us to track each individual patient's location at any given time in the hospital, we use this information to estimate an accurate, time-varying level of occupancy in the ICU. Prior to NaviCare, such micro-level data had not been available, and researchers typically had to rely on less accurate census data to estimate occupancy levels. In addition, the timestamp information allows us to test for potential seasonality associated with the time of admission.

The medical data was obtained from the cardiac surgery clinical database from the Society for Thoracic Surgeons (STS). This clinical database provides a comprehensive set of medical variables that enables us to capture the medical heterogeneity across patients. For example, the type of procedure, pre-existing conditions, age, gender and risk factors affect both the recovery time (and hence length of stay in the ICU) as well as the likelihood of developing complications that could lead to bounce-backs. These variables are used to adjust for patient severity using a widely used model called EuroSCORE, which includes a set of indicators for potential sources of complications such as previous cardiac surgery, an unstable angina, or a neurological dysfunction. We augment the EuroSCORE model to also include the New York Heart Association (NYHA) classification, which is a discrete measure of classifying the extent of heart failure. Finally, we used the patient billing records to determine the payer type, insurance status of the patient. In addition to medical variables, one might argue that hospitals discriminate the level of service they offer depending on the insurance status of the patient. Table 2 provides a comprehensive listing of the patient level clinical risk factors and non-clinical controls such as day of week and the type of payer (e.g. medicare, medicaid, private insurance, or self-pay).

We merge these three data sets based on a unique patient identifier to create a comprehensive and consolidated data set consisting of both, medical and operational variables. A total of 1365 patient admissions occurred from June 2006 to June 2007. This initial set of patients includes acutely severe (e.g. heart transplant) patients as well as patients who were admitted with primarily pulmonary conditions. In addition, some medical indicators were missing for several patients, rendering risk adjustment and the use of the EuroSCORE risk model inapplicable. Since the lengths of stay and likelihood of bounce-back for the acute case patients and those with missing observations could not be risk-adjusted, we do not include them in our statistical analysis. However, we do use these patients in estimating the ICU occupancy.

We observed that a few ($n = 42$) patients bounced back more than once. For these patients, we simply analyzed the effect of occupancy on the LOS of the patient during the first visit. Future visits obviously affect the future occupancy in the ICU, and so we used these revisits to estimate future occupancy. However, we did not examine the effect of future occupancy on the LOS reduction in future ICU visits. We did this for three reasons. First there is no clear theoretical or medical basis that we could draw on for the effect of the first bounce-back on the second. Secondly, and perhaps more importantly, the limited number of observations do not allow us to test any hypotheses that we may develop to examine the effect of the first bounce-back on the second. Finally, the nature of patient revisit is fundamentally different from the initial stay. The initial visit is characterized primarily by recovery from the "shock" of the surgery. The recovery is more likely characterized by further complications (e.g. infections). For these reasons we did not pool these stays in a single regression analysis. A total of 1036 patients had the complete set of clinical and operational (LOS, occupancy, bounce-back) data. Tables 3 and 4 provide summary statistics for these measures.

For our research design involving matching estimators, we employ a binary measure of $BUSY$. The ICU has a total of 18 beds. On any given shift, if the number of scheduled arrivals and the number of existing ICU patients exceed the total available bed capacity, patients have to be discharged early in order to accommodate the new arrivals. For each patient i in our sample, $BUSY_i$ is estimated at the time of discharge from the ICU; $BUSY$ is defined to be 1 if the sum of the number of newly arriving patients during the shift and the number of patients in the ICU at the time that patient i is discharged exceeds the total available capacity. Therefore from an operational perspective, a cutoff value of BUSY at 18 is an appropriate proxy for high occupancy.

6 Econometric Specifications

First and foremost, healing and recovery in the ICU require time. Consequently, everything else being equal, the longer a patient has spent in the ICU, the more likely she is to be ready for discharge. In other words, the hazard $(h_i(t))$ of patient i being discharged at any given time t increases with the time spent in the ICU. This assertion is in line with an extensive body of research in biostatistics, modeling the effect of time on patient recovery. Our interest is in examining the patient i's initial length of stay in the ICU (LOS_i) as a function of various medical and operational factors. We model the length of stay (LOS) of the patient in the ICU using the Weibull distribution, as the Weibull is commonly used in the biostatistics literature to model durations for patient recovery.

Assuming that the LOS has a Weibull distribution, we obtain the following econometric specification:

$$\log(LOS_i) = \mathbf{X}_i\boldsymbol{\beta} + \sigma\varepsilon_i \tag{1}$$

where the variables in $\mathbf{X}_i$ capture the various patient-level and system-level factors that affect the patient's length of stay. For example, patient-level variables are age or procedure type while system-level variables are the ICU occupancy, month of year or day of the week. $\boldsymbol{\beta}$ provides estimates for the effect of these covariates on the LOS. $\mathbf{X}_i$ also includes the dummy intercept term. σ is the scale parameter for the Weibull distribution, and ε_i denotes the error term.

6.1 Effect of Occupancy on Initial Length of Stay

In specification (1) above, the length of stay in the ICU in the absence of capacity constraints can be explained by the parameters in $\mathbf{X}_i$. To assess the effect of ICU occupancy on the length of stay, we append the binary variable $BUSY_i$ in the following expanded specification. We redefine $BUSY_i = 1$ to indicate that the ICU is at high occupancy at the time of *discharge* of patient i; 0 indicates otherwise:

$$\log(LOS_i) = \gamma BUSY_i + \mathbf{X}_i\boldsymbol{\beta} + \sigma\varepsilon_i \tag{2}$$

The coefficient γ provides us the estimate of the effect of high occupancy on the length of stay of the patient and tests (Hypothesis 1).

6.2 Effect of Early Discharge on Bounce-Back

To investigate whether an early discharge has an effect on the likelihood of a bounce-back, we start with the following model:

$$Y_i^* = \mathbf{X}_i \boldsymbol{\pi}^* + \eta^* LOS_i + u_i \tag{3}$$

$$BB_i = \mathbf{1}[Y_i^* > 0] \tag{4}$$

where Y_i^* is the unobserved state of health of patient i after being discharged from the ICU with length of stay LOS_i. Physicians may evaluate the wellness using various metrics for overall physiological function, response to medication, cognitive ability, etc. $\mathbf{X}_i$ includes patient level factors (such as age, gender, various measures of physiological functioning, emergency status, as well as day of admission) that have been identified in the medical literature to affect the rate of recovery and the likelihood of patient morbidity including revisits (see Table 2), and u_i captures unobserved patient heterogeneity. Although the actual state of health is a latent variable, we do observe the incidence of a bounce-back, captured by the binary variable BB_i. This model is thus estimated with the following probit specification:

$$\mathrm{Pr}_i = \Phi(\eta LOS_i + \mathbf{X}_i \boldsymbol{\pi}_{BB}) \tag{5}$$

where Pr_i denotes the probability that patient i has to revisit the ICU, and Φ is the cdf of the standard normal distribution. $\boldsymbol{\pi}_{BB}$ provides estimates for the various risk factors on the likelihood of a bounce-back. The above model is an extension of the EuroSCORE model, which we augment to include LOS_i, which is our key variable of interest. If a faster discharge (shorter length of stay) leads to an increased likelihood of a revisit, we expect η to be negative. This would provide support for (Hypothesis 2).

However, the presence of endogenous variables on the right hand side of equation (3) could bias our estimate of η. Unobserved patient level risk factors, for example would tend to increase LOS_i and also simultaneously increase the likelihood of an adverse outcome requiring a revisit. Such unobserved variables would have the effect of attenuating our estimate of η. In other words, η is an under-estimate of the effect of a faster discharge on the likelihood of a bounce-back. An appropriate instrumental variable strategy can be used to circumvent such endogeneity concerns, and to generate a consistent estimate of η. We include the following specification to describe our IV estimation strategy:

$$LOS_i = \mathbf{X}_i \boldsymbol{\pi}_{LOS} + \gamma BUSY_i + v_i \tag{6}$$

where $\boldsymbol{\pi}_{LOS}$ captures the effect of the patient level controls on the length of stay, and v_i accounts for unobserved heterogeneity that impact patient recovery. Since a busy ICU could lead to a shorter length of stay (LOS_i), while arguably having no impact on the unobserved factors underlying patient severity, occupancy in the ICU ($BUSY_i$) is a potential candidate as an instrumental variable for the effect of length of stay on the likelihood of a bounce-back. In other words, $BUSY_i$ provides exogenous variation in the length of stay that is unrelated to the patient's underlying conditions, and thus allows us to generate a consistent estimate for η.

$BUSY_i$ is an appropriate instrumental variable if *i*) it is correlated with the length of stay, and *ii*) it satisfies the exclusion restriction, i.e. it is independent of unobserved factors underlying the severity of the patient. We test for the first necessary condition for $BUSY_i$ to be an appropriate instrumental variable by establishing that the length of stay is correlated with the occupancy (to be discussed in the results section). To validate the second assumption of independence between case severity and occupancy, we computed the correlation between the level of pre-operative severity, as measured by the New York Heart Association Severity index, and the occupancy in the ICU. The resulting correlation coefficient is not statistically different from zero ($p = 0.32$) suggesting that the underlying patient severity is uncorrelated with the occupancy in the ICU. Moreover, we found in our discussion with the doctors working in cardiac care that the state of the ICU was not considered when scheduling new surgeries in pre-operative planning. This is because surgeries are scheduled days and weeks in advance. At that time predicting future ICU occupancy is simply not feasible. Note that emergency admissions, which account for a third of all admissions are, by definition, always random. Hence the severity of these patients is independent of occupancy in the ICU. In addition, we performed robustness checks to verify the lack of correlation between the arrival volume of patients and the occupancy in the ICU. Given this independence between case severity and occupancy, the assignment of patients to either a busy or a non-busy ICU is effectively a natural experiment. This observation allows us to generate an unbiased estimate η_{IV} of the effect of length of stay on the likelihood of a bounce-back. We estimate η_{IV} using the method outlined in Woolridge (2002 pp 472-477). In particular,

$BUSY_i$ will be used as an instrument for endogenous regressor LOS_i in (5), and η_{IV} will be estimated by instrumental variable probit maximum likelihood.

6.3 Capacity Implications of Discharge Decisions

Recall from our earlier discussion that an early discharge of a patient as a result of capacity rationing in a busy ICU has two effects. First, the early discharge reduces the initial length of stay of the patient. Second, it could increase the probability that the patient bounces back, thereby consuming ICU capacity at a later point. As a result of these two effects, the total capacity consumption ($TOTAL\_LOS_i$) might increase or decrease with an early discharge. Moreover, in addition to the total capacity consumption of a patient, the hospital is especially concerned about the total peak capacity consumption of a patient and how this changes with an early discharge ($TOTAL\_PEAK\_LOS_i$). In the extreme case, if all bounce backs occurred at times when the ICU is not busy, we could entirely ignore the extra days the patients spend in the ICU when they bounce back.

To study the capacity implications of an early discharge, we divide up the patient population into two groups, the group of patients discharged from a busy ICU, I_{BUSY}, and the group of patients discharged from a non-busy ICU, $I_{NONBUSY}$ where ($I_{BUSY} \cap I_{NONBUSY} = \emptyset$). As we argued above, the discharge of a patient from either a busy or a non-busy ICU is effectively a natural experiment and the severity of a patient is independent of the occupancy of the ICU.

For each patient, we define LOS_MED_i as the length of stay that the patient would have experienced if there were no capacity constraints in the ICU, i.e., this is the length of stay determined based purely on the medical risk factors of patient i. For the patients that were discharged from a non-busy ICU, this medically required length of stay corresponds to the actually realized length of stay ($LOS\_MED_i = LOS_i \; \forall i \in I_{NONBUSY}$). In contrast, for the patients that were discharged from a busy ICU, the medically required length of stay is not realized (and hence cannot be observed).

To estimate the medically required length of stay for patients discharged from a busy ICU, we match each patient in I_{BUSY} with one or several patients in $I_{NONBUSY}$ that have similar medical conditions. This is achieved by first computing the EuroSCORE model risk score, $p(\mathbf{X}_i)$, based on the set of medical variables, $\mathbf{X}_i$, discussed previously and then creating a set of matching patients, I_i, for each patient $i \in I_{BUSY}$ such that $p(\mathbf{X}_i) \approx p(\mathbf{X}_j)$ with $j \in I_i$. The detailed process of matching follows the method of matching estimators (see e.g. Rosenbaum and Rubin 1977 and Heckman et al 1999) and is described in the Appendix. The Appendix also describes the conditions necessary for $p(\mathbf{X}_i)$ to be a valid risk score. Once we have identified a set of matching patients, I_i, for a given patient i we can estimate the medically required length of stay, LOS_MED_i, as:

$$LOS\_MED_i = \frac{1}{n_{I_i}} \sum_{j \in I_i} LOS_j$$

where n_{I_i} is the number of patients in I_i.

Based on the difference in the length of stay LOS_i and the estimated medically required length of stay, LOS_MED_i, we can quantify the immediate capacity benefit from discharging patient i early from a busy ICU as:

$$\Delta LOS_i = LOS_i - LOS\_MED_i, \qquad \text{where } i \in I_{BUSY}$$

A negative value of ΔLOS_i implies that bed capacity was freed up by discharging the patient early.

An early discharge from a busy ICU does not only shorten the initial length of stay, but also increases the likelihood of a future bounce-back. We next examine the additional capacity consumption of such revisits to the ICU. Let $REVISIT\_MED_i$ be the additional time that a patient i spends in the ICU for a potential revisit (bounce-back). For each patient i discharged from a busy ICU ($i \in I_{BUSY}$), the expected revisit length of stay for a patient discharged from a non-busy ICU (and hence had experienced the medically required length of stay instead of being discharged early) can be computed by looking at patients I_i with similar medical conditions ($p(\mathbf{X}_i) \approx p(\mathbf{X}_j)\ \forall j \in I_i$) that were discharged from a non busy ICU:

$$REVISIT\_MED_i = \frac{1}{n_{I_i}} \sum_{j \in I_i} REVISIT_j$$

We expect the patients that were discharged early from a busy ICU ($i \in I_{BUSY}$) to have a larger ICU capacity consumption due to revisits. We can quantify this capacity loss caused by an increased amount of ICU capacity spent on revisits as:

$$\Delta REVISIT_i = REVISIT_i - REVISIT\_MED_i$$

The impact of early discharges on the total ICU capacity consumption of patient i ($TOTAL\_LOS_i$) is the net effect of the immediate capacity gains obtained from early discharges (ΔLOS_i) and the capacity losses resulting from more and / or longer revisits ($\Delta REVISIT_i$):

$$\Delta TOTAL\_LOS_i = \Delta LOS_i + \Delta REVISIT_i$$

We can also estimate the peak capacity consumption ($TOTAL\_PEAK\_LOS_i$) for each patient i by considering if the patient was discharged from a busy ICU and if the patient was re-admitted to a busy ICU. Recall that $BUSY_i$ is equal to 1 if the patient was discharged from a busy ICU and 0 otherwise. Similarly, we define the binary variable $REVISIT\_BUSY_i$ to equal to 1 if the ICU was busy at the time of readmission of patient i and 0 otherwise. Then, we can compute the impact of early discharges on the peak capacity consumption of patient i as:

$$\Delta TOTAL\_PEAK\_LOS_i = \Delta LOS_i * BUSY_i + \Delta REVISIT_i * REVISIT\_BUSY_i$$

Both $\Delta TOTAL\_LOS_i$ and $\Delta TOTAL\_PEAK\_LOS_i$ are the result of occupancy-induced change in the discharge decision. Since the values of ΔLOS_i and $\Delta REVISIT_i$ are significantly smaller than the initial stays and revisits, it is reasonable to assume that the occupancy is not likely to change drastically over the significantly smaller values of ΔLOS_i and $\Delta REVISIT_i$. Therefore, $BUSY_i$ and $REVISIT\_BUSY_i$ provide good approximations for the occupancy during the incremental lengths of stay (ΔLOS_i and $\Delta REVISIT_i$). Finally, we estimate the average capacity effect among patients who are similar in medical conditions. We do this by first dividing up the patient population discharged from a busy ICU into equally sized groups (G) based on their risk scores. We then construct equally weighted averages of $\Delta TOTAL\_LOS$ and $\Delta TOTAL\_PEAK\_LOS$ within each group. Our estimator $\Delta TOTAL\_LOS_G$ averages the overall capacity impact of early discharges on patients in group G and $\Delta TOTAL\_PEAK\_LOS_G$ estimates the peak capacity impact of early discharges on patients in G :

$$\Delta TOTAL\_LOS_G = \frac{1}{n_G} \sum_{i \in G} \Delta TOTAL\_LOS_i$$

$$\Delta TOTAL\_PEAK\_LOS_G = \frac{1}{n_G} \sum_{i \in G} \Delta TOTAL\_PEAK\_LOS_i$$

where n_G is the number of patients in G.

7 Results

We find that the occupancy level in the ICU has a significant impact on an admitted patient's length of stay. When we estimate equation (2) by the method of maximum likelihood, we find that the coefficient estimate (γ) for the explanatory variable indicating that the ICU is busy ($BUSY = 1$) is -0.169 (Table 5, Model 1). For a patient discharged from a busy ICU this corresponds to a length of stay that is 16% shorter than that for a comparable patient discharged from a low occupancy ICU. The effect of the occupancy on the length of stay is also evident from non-parametric Kaplan-Meier estimates of the aggregate survival functions generated for busy and non-busy estimates shown in Figure 2. This finding is in concord with the observations of physicians and nursing staff, who indicated to us that when the ICU gets busy, the least severe patients are discharged faster, as long as there is no significant risk to the patients. We also find that the insurance status of the patient (Model 1) or the effects of monthly and daily seasonality (Model 2) have no significant effect on the effect of occupancy on the length of stay (Hypothesis 1).

We next study the impact of the early discharges on the likelihood that a patient has to revisit the ICU. From our evaluation of the regression equation (5), we estimate the coefficient (η) for the explanatory variable measuring early discharge (LOS) to be -0.06 (Table 6). This provides support for (Hypothesis 2) that an early discharge is associated with an increased likelihood of a bounce-back. Our instrumental variable estimator η_{IV} is -0.76. For the average patient, the probability of a bounce-back is 14%. This corresponds to a normal distribution's z-statistic value of -1.08. The IV estimate of -0.70 suggests that for this average patient, an early discharge by day raises the z-statistic value to -0.32, which corresponds to a probability of bounce-back of 37.4%. An early discharge by one day is thus associated with an increase in the probability of a bounce-back by 23.4%. This larger estimate of η_{IV} is consistent with our speculation that unobserved patient-level risk factors would lead us to underestimate the impact of an early discharge on the likelihood of a bounce-back had we simply used the probit estimator in (5). The medical control variables that are statistically significant are all positive. Since the incidence of any of these medical controls is associated with higher severity levels, this finding is consistent with the established medical literature.

We next examine the impact of the increased rate of bounce-back on ICU capacity usage. Our summary statistics (Table 3) show that on average, revisits have a longer length of stay than first-time ICU visits - a median revisit lasts almost 3 days while a median first-time stay in the ICU lasts only 1.2 days. The second column of Table 7 displays the average of $\Delta TOTAL\_LOS_G$ for the patients each group G. Group 1 contains the least severe patients while group 4 contains the patients scoring in the highest range of the risk score. We should note that the severity level is based on the pre-surgery condition and diagnosis of the patient, not their medical condition at the time of discharge from the ICU. Given that the measures of severity (age, gender, and procedure type) are fixed during a hospital stay, and given that we do not observe time-varying health status of patients, our severity measure is fixed for a given patient's stay. For each of the groups, capacity is initially saved by discharging a patient early, as indicated by the negative values for ΔLOS_G. Similarly, $\Delta REVISIT_G$ is positive for all of the groups indicating that revisits take up valuable capacity. In general, $\Delta REVISIT_G$ is higher for patients in the higher risk categories. We also find that $\Delta TOTAL\_LOS_G$ has a statistically significant negative value for group 2, but a positive value for group 4. This means that the early discharge of group 2 patients freed up total bed days, despite the bounce-backs. On the other hand, for group 4 patients, the resulting bounce-backs are lengthy, resulting in a net increase in the total bed days used. One possible explanation for this effect is that complications of the low-severity patients can generally be handled in the step-down unit, obviating the need for a costly revisit to the ICU. However, any complications developed by higher-severity patients in the step-down unit call for an increased level of monitoring and a subsequent bounce-back to the ICU. Since the high severity patients are also associated with longer revisit stays, their net total length of stay ($\Delta TOTAL\_LOS$) increases.

In estimating $\Delta TOTAL\_PEAK\_LOS_G$ we do not find statistically significant results for groups 1 and 3. The peak capacity estimate $\Delta TOTAL\_PEAK\_LOS_G$ is positive for group 4, but not for group 2. In particular, peak capacity is reduced by 15.26 hours on average as a result of aggressive discharge in group 4. However, for group 2 patients, the early discharges increase peak capacity. Our results suggest that if an early discharge policy were to be adopted in an effort to increase throughput, they should be applied to group 2 patients. However, our analysis does not allow us to determine why this group of patients differs from the other groups, and we defer this examination to future research.

8 Model Validations and Robustness

In keeping with prior work in the medical literature, we used the Weibull distribution to describe the length of stay in model (2). One advantage of the Weibull model is that it provides the flexibility for the underlying hazard rates to be either increasing or decreasing. Nevertheless, the estimation of γ is sensitive to the distributional assumptions and the related underlying hazard rates. To provide a test of robustness for the validation of (Hypothesis 1), we use the Cox proportional hazard model. This semi-parametric approach frees us from having to make distributional assumptions and allows the hazard rate to vary with time. The instantaneous hazard rate $h(t)$ for a patient's discharge from the ICU can be expressed as:

$$h(t) = h_0(t) exp(\gamma_h BUSY + \mathbf{X}\boldsymbol{\beta}_h) \tag{7}$$

where $h_0(t)$ is the baseline hazard function that is allowed to be time-varying. γ_h provides an estimate for the effect of $BUSY$ on the hazard rate. We estimate (7) by the method of partial maximum likelihood (Cox 1972). We estimate γ_h to be 0.20 (Table 8). This corresponds to a hazard ratio of 1.226. In other words, regardless of the underlying evolution of the baseline hazard rate, we find that the instantaneous hazard rate of a patient's discharge from the ICU increases by 22.6% for a patient is in a busy ICU. Since the increase in hazard rate is equivalent to a reduction in the length of stay, this finding provides support for (Hypothesis 1).

A potential confounding effect in the estimation of γ in (2) arises if the hospital selectively operates on patients with lower anticipated ICU stay when the ICU is busy. This endogeneity could lead to a bias of our estimate for the coefficient of $BUSY$. We rule out the possibility of selection bias by estimating the correlation between occupancy in the ICU and the level of pre-operative severity, as measured by the New York Heart Association Severity index. The resulting correlation coefficient is not statistically different from zero. Thus there does not appear to be selective severity-based ICU admissions based on the occupancy level. One reason for this is that it is difficult (if not impossible) for the admitting personnel to predict the future ICU occupancy level at the time of scheduling elective procedures. We also examined the correlation between the number of admissions and the occupancy during a shift; we find that the correlation is statistically insignificant, a further indication of the lack of selective admissions based on ICU occupancy.

In our analysis, the explanatory variable $BUSY$ was estimated at the time of discharge. However, one could argue that it is not simply the occupancy at the time of discharge, but the also the occupancy in the ICU during the entire stay that determines the discharge decision. To further validate this definition of $BUSY$, we estimated equation (2) with $BUSY$ measured during times of admission. As further test for robustness, we also estimated BUSY at the start of the shift during which a patient was discharged. In estimating (2) and (5), we find that the coefficients for γ and η are negative, suggesting that our findings are robust to the times at which occupancy is estimated.

Finally, we find that controlling for day of week of admission or the month of admission does not have a significant impact on the discharge decision. In addition, the patient's insurance status (payer type) has a small impact on the length of stay.

When the ICU is busy, the alternative to discharging patients early is to cancel procedures. To examine whether procedure cancellations occurred, we investigated to what extent the patient arrivals were correlated with the occupancy in the ICU. We estimated the model: $AdmitVol_{Shift} = a + b * Occupancy_{Shift}$, where $AdmitVol_{Shift}$ is the admission volume for a given shift, and $Occupancy_{Shift}$ is the occupancy at the start of the shift. We found a lack of correlation (the intercept a and coefficient b is reported in Table A4 in the Apendix). This finding is plausible: patients are typically scheduled a few weeks in advance, and predicting future occupancy at the time when the surgery is scheduled is difficult.

When we discussed this issue with doctors from our research site, we obtained the following responses that confirmed our empirical examination: "There are times that we cancel elective surgical cases as a result of ICU occupancy- but this is very rare. I would say that it happens at most once a month." As far as early discharges are concerned, a doctor commented: "Of course, we do early discharges whenever we get full. We try to coordinate with the floors or look for other ICU beds, but when we are full and have a new patient arriving, what else do you expect us to do?" Similarly, his colleague from another teaching hospital in town observed: "It is quite rare to cancel operations. Getting in an extra patient when we are busy it usually works out somehow. There is always some slack in the system...either some patients can be pushed out or patients can board for a time in another ICU in the hospital before going to the floor. " And, another surgeon elaborated further: "Canceling a surgery is really rare. I remember one instance from two years ago. We were so full that we decided to cancel a scheduled OR procedure and it was a mess. The family of the patient, the doctor scheduled for the surgery, and the hospital administration, everybody was upset. We always find a way to fit a new patient in." Also, financial issues were pointed to as illustrated by yet another quote: "Elective procedures are typically associated with large revenues. You don't just go and cancel such procedures."

In this round of interviews, all interviewees confirmed that their discharge decision is strongly influenced by occupancy. Thus, although it is theoretically possible to use procedure cancellations as a way to match supply with demand, this does not happens frequently. Given the fixed inflow of patients and the capacity constraint, the only way to match supply with demand is the early discharge.

9 Conclusion and Future Research

In this paper we looked at the management of bed capacity in a cardiac intensive care unit. To determine the capacity needs of individual patients, we estimated a model of patient recovery that accounts for numerous patient-level risk factors. From this model, we found that the ICU rations its capacity during busy periods by discharging patients earlier.

However, we also found that an early discharge led to an increased likelihood of a patient revisit. That is, aggressively discharging patients to the step-down unit in order to free up capacity, led to an increase in likelihood of patients revisiting the ICU during the same hospital stay. In addition, we found that the revisits tended to incur long lengths of stay.

This observation raises the question of whether the ICU should keep patients longer the first time to reduce the probability of an incidence of revisit. Using the method of matching estimators, we estimated the additional length of stay needed for the initial patient visit, i.e. the "right first time" length of stay that would have been realized had the ICU not been busy. By comparing the total peak capacity usage for patients who were discharged early versus those who were not, we show that an aggressive discharge policy frees up peak capacity in the ICU only for lower severity patients. For the high-severity patients, however we find that an increased number of readmissions occur when the ICU is capacity constrained, thereby effectively reducing peak bed capacity. Thus, in our study of ICU capacity, the insights obtained from the quality management literature, favoring "to do it right the first time" dominate the benefits of capacity rationing for high severity patients. On the other hand, the hospital would be able to increase its patient throughput by selectively discharging the lower severity patients earlier.

We should note that this study is a first step towards quantifying the capacity trade-offs in discharge decisions. In practice, various unobserved factors determine the patient's recovery path and the exact circumstances governing the patient's discharge, although known to the medical experts at the time of discharge, are unknown to us as researchers. Many decisions are made not based on raw numbers and data, but on the basis of more subjective medical expertise developed over years of practice. Therefore, although we do not expect the results of our analysis to be the primary drivers of discharge decisions, we do hope that our findings can serve as additional information that the care providers can incorporate in their decision making.

Future research in Operations Management could look at ways to determine the optimal discharge policy with the objective of maximizing patient throughput. Future medical research is needed to build more sophisticated models of patient recovery that enable the hospital to customize the discharge decision to the medical profile of a patient while considering ICU occupancy (e.g. see Martin et al 2005 and Swenson 1992). In particular, various dimensions of patient recovery and quality of care both inside the ICU and after discharge need to be examined. Policy changes in the US, including the passage of the healthcare reform bill, could lead to increased demand for healthcare services. The ability to effectively manage the increase in volume of patients while operating under resource capacity constraints is likely to become even more important. In this paper, we address the short and medium term implications of discharge decisions from the ICU. However, it is important to also examine the effect on the long term wellbeing of a patient. For example, future studies could look at the effect on hospital revisits and rates of morbidity and mortality. One of the fascinating aspects of studying ICU operations is that both of these venues for future research essentially correspond to two sides of the same coin – only by building interdisciplinary models that combine medical variables with Operations Management decisions will we be able to improve the quality and productivity of our healthcare system.

References

Bertsekas, D. P. 2000. *Dynamic Programming and Optimal Control.* Athena Scientific.

Bohmer, R. M. J., Newell, J. and D. F. Torchiana. 2002. The Effect of Decreasing Length of Stay on Discharge Destination and Readmission after Coronary Bypass Operation. *Surgery* 132, no. 1: 10-16.

Centers for Disease Control and Prevention. 2005. Department of Health and Human Services. Justification of Estimates for Appropriation Committees.

Chan, P.S., Khalid, A, Longmore, L S., Berg, R. A., Kosiborod, M., J. A. Spertus. 2008. Hospital-wide Code Rates and Mortality Before and After Implementation of a Rapid Response Team *JAMA*. 300(21):2506-2513.

Cox, D. R. 1972. Regression Models and Life Tables. *Journal of the Royal Statistical Society Series B* **34** (2): 187–220.

Crabill, T. B. 1972. Optimal Control of a Service Facility With Variable Exponential Service Times and Constant Arrival Rate. *Management Science* **18** (9)

Department of Justice and Federal Trade Commission. 2004. Improving Health Care: A Dose of Competition. A Report by the Federal Trade Commission and the Department of Justice. July 2004.

EuroSCORE model. 2007. Retrieved October 25, 2007. http://www.euroscore.org

Gerchak, Y., D. Gupta, M. Henig. 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science,* **42**(3) 321–334

George, J. M., J. M. Harrison. 2004. Dynamic Control of a queue with adjustable Service Rate. *Operations Research* **49**(5) 720-731

Green,L.V. 2004.*Capacity planning in hospitals.*Handbook of Operations Research/Management Science Applications in Health Care, Kluwer Academic Publishers

Green, L.V., J. Meissner. 2002. Developing insights for nurse staffing. Columbia Business School, Working Paper

Green, L. V., S. Savin, B. Wang. 2006. Managing Patient Service in a Diagnostic Medical Facility. *Operations Research* **54** 11-25

Hall R. W (Ed). 2006. Patient Flow: Reducing Delays in Healthcare Delivery. International Series in Operations Research & Management Science. Springer

Heckman, J. J., Ichimura, H. and P. Todd. 1998. Matching as an Econometric Evaluation Estimator. *The Review of Economics Studies* **65**(2) 261-294

Henning, RJ, McClish D, Daly B, Nearman J, Franklin C, and Jackson D. 1987. Clinical characteristics and resource utilization of ICU patients: implications of organization of intensive care. *Crit Care Med*; 15: 264-269.

Horak, Jiri, KC, Diwas and Christian Terwiesch. 2009. Cardiothoracic Surgery Risk Stratification for Intra-hospital Decision Making and Inter-hospital Quality Comparisons. Wharton School Working Paper

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

Institute of Medicine (IOM). 2007. Hospital-Based Emergency Care: At the Breaking Point. Institute of Medicine of the National Academies

Kc, Diwas and Christian Terwiesch. 2009. Impact of Work Load on Service Time and Patient Safety: An Econometric Analysis of Hospital Operations. *Management Science* **55** (9):1486-1498

Kurki, T. S. 2002. Prediction of Outcome in Cardiac Surgery. *Mount Sainai Journal of Medicine* **69**(1, 2)

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

Liker, Jeffrey. 2004. The Toyota Way: 14 Management Principles from the World's Greatest Manufacturer. McGraw-Hill

Martin, C.M., Hill, A.D., Burns, K. and L.M. Chen. 2005. Characteristics and outcomes for critically ill patients with prolonged intensive care unit stays. *Crit Care Med*. 33(9):1922-7

McConnell, K. J., Christopher F. Richards, Mohamud Daya, Stephanie L. Bernell, Cody C. Weathers and Robert A. Lowe. 2005. Effect of Increased ICU Capacity on Emergency Department Length of Stay and Ambulance Diversion *Annals of Emergency Medicine*, Volume 45, Issue 5, Pages 471-478

Nashef, S. A., F. Roques, B. G. Hammill, E. D. Peterson, P. Michel, F. L. Grover, R. K. Wyse, T. B. Ferguson. 2002. Validation of European System for Cardiac Operative Risk Evaluation (EuroSCORE) in North American cardiac surgery *European Journal of Cardiothoracic Surgery* **22** 101- 105

Obel, N., Schierbeck, J., Pedersen, L, Storgaard, M., Pedersen, C., Sorensen, H. T. and B. Hansen. 2007. *Acta Anaesthesiol Scand* Vol. 51, 1225 - 1230

Ohno, T. 1988. The Toyota Production System: Beyond Large-Scale Production, Productivity Press, Portland.

Peake, S. L., J. L. Moran, D. R. Ghelani, A. J. Lloyd, M. J. Walker. 2006. The effect of obesity on 12-month survival following admission to intensive care: A prospective study. *Crit Care Med* 2006 Vol. 34, No. 12

Powell, S. G., K. L. Schultz. 2004. Throughput in Serial Lines with State-Dependent Behavior. *Management Science* **50**(8) 1095-1105

Reynolds, S. F., Bellomo, R. and K. Hillman. 2009. Rapid Response Team Implementation and Hospital Mortality Rates. *JAMA*. 301(16):1659.

Rosenbaum, P. R. and D. B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika.* **7**(1) pp 41-45

Rubin, D. B. 1977. Assignment to Treatment Group on the Basis of a Covariate. *J Ed. Statistics* **2** 1-26

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

Schultz, K. L., D. C. Juran, J. W. Boudreau. 1999. The Effects of Low Inventory on the Development of Productivity Norms. *Management Science* **45**(12) 1664-1678

Schultz, K. L., D. C. Juran, J. W. Boudreau, J. O. McClain, L. J. Thomas. 1998. Modeling and Worker Motivation in JIT Production Systems. *Management Science* **44**(12) 1595-1607

Sherner, J. H. 2009. Rapid Response Team Implementation and Hospital Mortality Rates. *JAMA*. 301(16):1658-1659.

Stidham, S., R. R. Weber. 1989. Monotonic and Insensitive Optimal Policies for Control of Queues with Undiscounted Costs *Operations Research* **87**(4)

Strauss, M. J., LoGerfo, J. P, Yeltatzie, J. A., Temkin, N. and L. D. Hudson. 1986. Rationing of Intensive Care Unit Services. *JAMA* 255(9)

Swenson, M. D. 1992. Scarcity in the intensive care unit: Principles of justice for rationing ICU beds. *The American Journal of Medicine* 92 (5):551-555

Toumpoulis, I. K., C. E. Anagnostopoulos, D. G. Swistel, J. J. DeRose, Jr. 2005. Does EuroSCORE predict length of stay and specific postoperative complications after cardiac surgery? *Eur Journal of Cardiothoracic Surgery* **27** 128-133

Tucker, A. 2004. The Impact of Operational Failures on Hospital Nurses and their Patients.

Journal of Operations Management **22**(2), 151-169.

Womack, J. P., D. T. Jones, and D. Roos. 1990. The Machine That Changed the World: The Story of Lean Production. New York: Rawson and Associates

Wooldridge, J. M. 2002. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.

Tables and Figures

Table 1: Operational Performance Variables

Measure	**Description and Coding**
i	Indicator variable for a patient
LOS	Length of stay of initial visit of patient
REVISIT	Revisit length of stay of patient
BB	Binary variable denoting the incidence of a revisit or a bounce-back of patient
OCCUPANCY	Occupancy in the ICU at the time of admission of patient
BUSY	Binary variable denoting whether occupancy will exceed the threshold during time of discharge of patient
TOTAL_LOS	Sum of initial and revisit lengths of stay
TOTAL_PEAK_LOS	Sum of peak capacity usage during initial and revisit stays

Table 2: Control Variables

Measure	Description and Coding
Age (AGE)	Patient ages less than 60 were coded 1, patients between 60 and 70 coded 2, patients between 70 and 80 coded 3, patients between 80 and 90 coded 4 and patients above 90 coded 5.
Gender (GENDER)	Females are coded 1 and males 0.
Chronic Pulmonary Disease (CH_PULM_DIS)	Indicates whether the patient is on medication for lung conditions or if the chronic lunge disease condition is moderate or severe.
Extracardiac Arteriopathy (EXT_ART)	Indicates the presence of vascular disease.
Neurological Dysfunction (NEUR_DYSF)	If a cerebrovascular disease exits, this explanatory variable is coded 1. The time of occurrence of the dysfunction, or the type of the cerebrovascular disease is ignored.
Previous Cardiac Surgery (PREV_CARD_SURG)	Indicator to denote if patient has had prior cardiac surgery. The type of cardiac surgery is not considered.
Serum Creatinine (SERUM_CREAT)	If the level is higher than 200mmol/l this risk factor is coded 1.
Active Endocarditis (ACT_ENDCRDT)	Indicator to denote that endocarditis is active.
Critical Preoperative State (CRIT_PRE_STATE)	This indicator variable denotes the pre-operative state of the patient (critical state or not). The factors that determine whether the patient is in critical state or not are the presence of arrhythmia (irregular heartbeats), cardiogenic shock, need for resuscitation, the need for an intra aortic balloon pump (IABP), or the use of nitrates administered through an I.V.
Unstable Angina (UNST_ANG)	Indicates syndrome that is intermediate between stable angina and a myocardial infarction.
Left Ventricular Dysfunction (LV_DYS)	Indicates whether ejection fraction is less than 30%
Recent Myocardial Infarction (RECENT_MYCR_INF)	Indicates whether myocardial infarction (heart attack) occurred in the last 90 days.
Pulmonary Hypertension (PULM_HYPER)	Indicates that the systolic pulmonary pressure exceed 60 mmHg.
Emergency (EMER)	Indicates status of patient at admission. Emergency is coded 1
Other than isolated CABG (OTHER_CAB)	Indicates whether in addition to a Coronary Artery Bypass Grafting, another type of heart procedure was performed.
Surgery on Thoracic Aorta (SURG_THOR)	Indicator for the presence of Aortic Aneurysm.
Post-infarction Septal Rupture (POSTINF_RUPT)	Indicates whether ventricular septum ruptured following a heart attack.
Day of Week (DAY)	Day of week of procedure
NYHA Classification (NYHA)	New York Heart Association Risk Classification (Ranging from 1 to 4)
Payer Type (PAYER)	Categorical Variable to denote Medicare, Medicaid, Insurance, Self-Pay or None

Table 3: Operational Variables Summary Statistics

Variable	Mean	Standard Deviation	Median
LOS (Days)	2.2	3.1	1.2
REVISIT on Bounce-back (Days)	4.2	4.6	2.8
OCCUPANCY (Beds)	15.4	2.6	16
BUSY	0.40	0.49	0.0
BB	0.14	-	-

N = 1365. Note: Summary Statistics Includes Pulmonary Patients

Table 4: Controls Summary Statistics

Variables	Mean
NYHA	2.2
AGE	62.3
GENDER	0.34
CH_PULM_DIS	0.14
ACT_ENDCRDT	0.15
NEUR_DYSF	0.17
RECENT_MYCR_INF	0.29
SERUM_CREAT	0.11
ACT_ENDCRDT	0.06
CRIT_PRE_STATE	0.33
UNST_ANG	0.17
LV_DYS	0.22
RECENT_MYCR_INF	0.1
PULM_HYPER	0.0037
EMER	0.3
OTHER_CAB	0.17
SURG_THOR	0.16
POSTINF_RUPT	0.005
DAY (Sun, Mon, Tu, Wed, Th, Fr, Sat)	(0.02, 0.19, 0.19, 0.18, 0.17, 0.2, 0.03)

N = 1036. Note: Pulmonary Patients Do Not Appear in Table 4

Table 5: Effect of Occupancy on Length of Stay

Coefficient	**Model 1**	**Model 2**	**Model 3**
Intercept	3.4399 (0.5105) ***	3.4 (0.4375)	3.3453 (0.4127) ***
BUSY	-0.1693 (0.0603) ***	-0.1715 (0.0601) ***	-0.1827 (0.0603) ***
NYHA 1	-0.2512 (0.0953) ***	-0.2445 (0.0946) ***	-0.2559 (0.0947) ***
NYHA 2	-0.0952 (0.0941)	-0.0967 (0.0936)	-0.1121 (0.0935)
NYHA 3	0.0714 (0.0899)	0.0674 (0.0897)	0.0629 (0.0901)
AGE 1	-0.1308 (0.3936)	-0.123 (0.3918)	-0.2107 (0.3929)
AGE 2	0.0424 (0.3942)	0.0575 (0.3934)	-0.026 (0.3946)
AGE 3	0.0716 (0.3924)	0.0818 (0.3925)	-0.0037 (0.3938)
AGE 4	0.114 (0.3962)	0.1248 (0.3966)	0.0304 (0.398)
GENDER	0.1592 (0.0567) ***	0.1589 (0.0564) ***	0.1698 (0.0564) ***
CH_PULM_DIS	0.0966 (0.0718)	0.0931 (0.0717)	0.1001 (0.0715)
EXTRA_CARD_ART	0.0706 (0.0745)	0.069 (0.0746)	0.0821 (0.0742)
NEUR_DYSF	0.0026 (0.067)	0.0012 (0.0669)	0.0284 (0.0665)
RECENT_MYCR_INF	0.1112 (0.1315)	0.1146 (0.1314)	0.1447 (0.1314)
PREV_CARD_SURG	0.0344 (0.0589)	0.0427 (0.0577)	0.0477 (0.0576)
SERUM_CREAT	0.214 (0.1081) **	0.2086 (0.1076) **	0.1921 (0.1077) **
ACT_ENDCRDT	0.0417 (0.1046)	0.0624 (0.1025)	0.0598 (0.1025)
CRIT_PRE-STATE	0.1623 (0.065) **	0.1617 (0.0647) **	0.1575 (0.0651) **
UNST_ANG	-0.1171 (0.1245)	-0.1218 (0.1243)	-0.1564 (0.1236)
LV_DYS	0.0503 (0.0734)	0.0515 (0.0729)	0.0336 (0.073)
PULM_HYPER	0.4889 (0.4095)	0.493 (0.4091)	0.4804 (0.4088)
EMER	0.2313 (0.0779) ***	0.2403 (0.0769) ***	0.2494 (0.0748) ***
OTHER_CAB	0.267 (0.115) **	0.2726 (0.1149) **	0.2713 (0.1148) **
SURG_THOR	0.2094 (0.0697) ***	0.217 (0.0687) ***	0.2257 (0.0686) ***
POSTINF_RUPT	0.3805 (0.33)	0.3498 (0.3298)	0.3562 (0.3302)
CAB	0.0165 (0.1054)	0.0114 (0.1051)	0.0052 (0.1055)
AORTIC_VALVE	-0.1724 (0.0633) ***	-0.1737 (0.0632) ***	-0.1778 (0.0625) ***
MITRAL_VALVE	0.0059 (0.0722)	-0.0007 (0.0722)	0.0025 (0.0723)
PULM_VALVE	0.1247 (0.3324)	0.1148 (0.3326)	0.118 (0.3327)
INSURANCE	-0.0175 (0.2678)		
MEDICAID	0.2919 (0.4482)		
MEDICARE	-0.0013 (0.2737)		
DAY	Included	Included	Not Included
MONTH	Included	Included	Included
LogLikelihood (Pr > Chi-Sq)	< 0.001	< 0.001	< 0.001

*Standard errors are shown in parentheses. ***, **, and * denote statistical significance at the 1%, 5% and 10% confidence levels respectively*

Table 6: Effect of Early Discharge on Likelihood of Bounceback

Coefficient	Probit Estimate (se)	Probit IV Estimate (se)
intercept	-1.48 (0.37) ***	0.52 (0.71)
LOS (η)	- 0.061 (0.039) *	- 0.762 (0.075) ***
GENDER	0.11 (0.11)	0.22 (0.072) ***
CH_PULM_DIS	0.035 (0.14)	0.16 (0.093) *
ACT_ENDCRDT	0.054 (0.21)	-0.037 (0.14)
EXTRA_CARD_ART	0.30 (0.14) **	0.25 (0.14) *
NEUR_DYSF	-0.012 (0.31)	0.033 (0.088)
RECENT_MYCR_INF	0.058 (0.244)	0.11 (0.16)
SERUM_CREAT	0.24 (0.18)	0.035 (0.18)
CRIT_PRE_STATE	0.26 (0.12) **	0.293 (0.14) ***
UNST_ANG	0.19 (0.21)	- 0.023 (0.18)
LV_DYS	0.13 (0.14)	0.16 (0.09) *
EMER	0.23 (0.14) *	0.29 (0.14) **
OTHER_CAB	0.14 (0.15)	0.22 (0.10) **
SURG_THOR	0.40 (0.13) ***	0.19 (0.19)
POSTINF_RUPT	0.43 (0.64)	0.62 (0.49)
LogLikelihood (Pr > Chi-Sq)	< 0.001	< 0.001

*Standard errors are shown in parentheses. ***, **, and * denote statistical significance at the 1%, 5% and 10% confidence levels respectively*

Table 7: Effect of Busy Admission on Length of Stay (Hours)

Group (G)	ΔLOS_G	$\Delta\text{REVISIT}_G$	$\Delta(\text{REVISIT* REVISIT\_BUSY})_G$	$\Delta\text{TOTAL\_LOS}_G$	$\Delta\text{TOTAL\_PEAK\_LOS}_G$
1	-0.69	3.10	2.16	2.41	1.48
	(2.48)	(2.78)	(1.90)	(3.98)	(3.16)
2	-11.4 ***	3.28	1.39	-8.10 *	-9.98 ***
	(2.33)	(3.35)	(1.27)	(4.24)	(2.72)
3	-3.15	6.26 *	5.93 ***	3.11	2.78
	(3.19)	(3.69)	(1.89)	(4.99)	(3.49)
4	-4.44 ***	33.83 ***	19.71 ***	29.39 ***	15.26 **
	(3.06)	(9.63)	(6.43)	(10.65)	(6.86)

*Standard errors are shown in parentheses. ***, **, and * denote statistical significance at the 1%, 5% and 10% confidence levels respectively. The 1036 risk-adjusted patients are split uniformly across groups 1-4. 541 of the patients fall into the BUSY designation.*

Table 8: Effect of Occupancy on Hazard Rate of Discharge

Coefficient	Model 1	Model 2	Model 3
BUSY (γ_h)	0.204 ***	0.206 ***	0.215 ***
	(0.079)	(0.079)	(0.078)
Payor Type	Included	Not Included	Not Included
Monthly and Daily Seasonality	Included	Included	Not Included
LogLikelihood (Pr > Chi-Sq)	< 0.001	< 0.001	< 0.001

*Standard errors are shown in parentheses. ***, **, and * denote statistical significance at the 1%, 5% and 10% confidence levels respectively. Control variables (provided in the online appendix) are not displayed.*

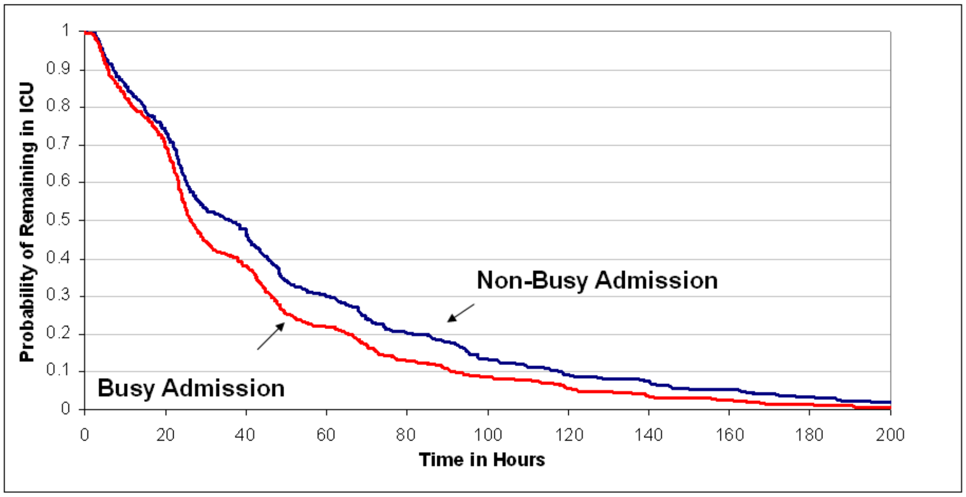

Figure 2: Effect of Occupancy on Length of Stay

Appendix: Risk Score Matching

We use Y_i to denote the outcome variable (e.g. the length of stay). Let Y_{i1} represent the value of the outcome variable when patient i is discharged from a busy ICU ($BUSY_i = 1$), and let Y_{i0} represent the outcome value when patient i is discharged from a non-busy ICU ($BUSY_i = 0$). Then, the observed outcome for patient i is:

$$Y_i = BUSY_i * Y_{i1} + (1 - BUSY_i) * Y_{i0}$$

The effect of ICU occupancy on the outcome is

$$\tau_i = Y_{i1} - Y_{i0}$$

and the effect we would like to identify is the effect of a busy admission (or the "treatment" effect) on the outcomes for patients discharged from a busy ICU:

$$[\tau | BUSY = 1] = E[Y_{i1} | BUSY_i = 1] - E[Y_{i0} | BUSY_i = 1]$$

However, Y_{i0}, which is the outcome that would have been realized *had* patient i been discharged from a non-busy ICU, is not directly observed for the patient i who is actually discharged from a busy ICU. We estimate this counterfactual Y_{i0} by assuming that after conditioning on observable covariates $\mathbf{X}_i$, there is no selection bias in the assignment of a patient to either a busy or a non-busy ICU. In other words, assuming selection on observables

$\mathbf{X}_i$, the assignment of patients to a busy or a non-busy ICU is effectively random. Since $BUSY$ is an instrumental variable (as discussed previously), the assignment of patients to a busy or a non-busy ICU is effectively a natural experiment, and in the words of Rosenbaum and Rubin (1983), the ignorability condition is satisfied. That is,

$$E[Y_{ij}|\mathbf{X}_i, BUSY_i = 1] = E[Y_{ij}|\mathbf{X}_i, BUSY_i = 0], \quad for\ j = 0, 1$$

However, estimation of $E[Y_i|\mathbf{X}_i, BUSY_i]$ is difficult if the dimension of $\mathbf{X}_i$ is large. We thus use the method of Propensity Score Matching (Proposition 2, Rosenbaum and Rubin 1983) to show that conditional on the propensity score (or a balancing score that balances the observed covariates amongst the busy and non-busy admission groups), each individual has the same probability of assignment to a busy ICU, as in a randomized experiment. In the notation of Rosenbaum and Rubin (1983), if $p(\mathbf{X}_i)$ is the propensity score, then $X_i \perp BUSY_i \mid p(\mathbf{X}_i)$.

We use the cumulative probability distribution of the initial length of stay explained using the full set of variables used in the $\mathbf{X}_i$ in (2) and (5) as the propensity score. In order for $p(\mathbf{X}_i)$ to be a valid propensity score, two conditions need to be met. First, the ignorability condition needs to be satisfied. As indicated above, the assignment of patients to a busy or non-busy ICU is effectively a natural experiment and the ignorability condition is trivially satisfied. Secondly, there needs to be common support in the propensity scores across the two groups of patients. From Figure 3 we find that there is good overlap in the propensity scores across the two busy and non-busy admissions, and this condition is satisfied.

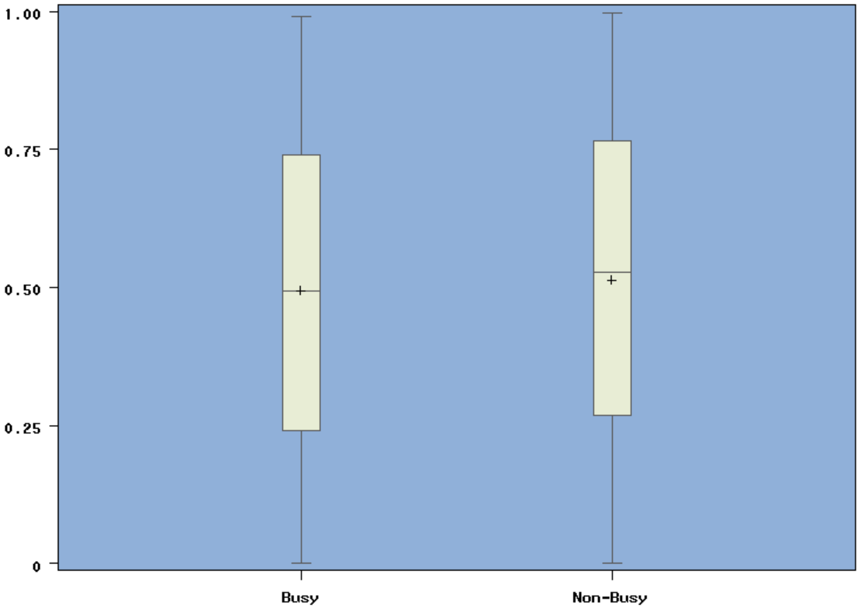

Figure 3: Boxplot of Propensity Scores of Busy versus Non-Busy Admissions

Next, given the estimated propensity score (which we now label p_i), we estimate the non-parametric regression $E(Y_i \mid BUSY_i = j,\ p(\mathbf{X}_i)) \quad for\ j = 0, 1$ using the method of matching estimators (Heckman et al 1997). Our matching estimator takes the form

$$\hat{\alpha} = \frac{1}{n_1} \sum_{i \in I_1} [Y_{i1} - \hat{E}(Y_{i0} | BUSY_i = 1, p_i)] \tag{8}$$

where

$$\hat{E}(Y_{i0} | BUSY_i = 1, p_i) = \frac{1}{n_{I_0}} \sum_{j \in I_0} Y_{0j}$$

I_1 denotes the set of patients discharged from a busy ICU, and I_0 is the set of patients discharged from a non-busy ICU. n_1 is the number of patients in I_1 and the match for each patient $i \in I_1$ is an equally weighted average over the outcomes of their counterpart patients in the non-busy admission group, which consists of n_{I_0} members that fall within a 5 percentile range of p_i.