ICU Admission Control: An Empirical Study of Capacity Allocation and its Implication on Patient Outcomes

Song-Hee Kim, Carri W. Chan, Marcelo Olivares, and Gabriel Escobar

September 7, 2013

Abstract

This work examines admission control of the Intensive Care Unit (ICU) which provides care for a hospital’s most critically ill patients. We focus on how congestion can impact ICU admission decisions and ultimately patient outcomes. We build an econometric framework to explain how ICU admission decisions are made in practice, which captures key trade-offs in allocating beds to patients with heterogeneous medical needs and stochastic arrival patterns. In addition to describing the actual admission policy used by hospitals, this econometric model provides instrumental variables that can be used to identify the effect of endogenous ICU admission decisions on patient outcomes. We estimate these models using patient data from an integrated healthcare delivery system with nearly 200,000 hospitalizations. We show that busy ICUs –defined as exceeding the 95th percentile value of the observed bed occupancy distribution– are associated with lower chance of admission (a 53% decrease on average) which can further lead to significant health implications. In turn, providing ICU care can improve patient outcomes substantially: hospital length of stay decreases by 1.2 days and the likelihood of readmission drops by 3.4%. Using our empirical results, we study the performance of an admission policy that minimizes adverse outcomes based solely on objective metrics available to all patients admitted from the ED and compare it with the current policy used by hospitals in our study. Because the current policy also uses the discretion of physicians (which is not captured in the available objective metrics), we see that the policy based on objective criteria can sometimes underperform, but not always. Additionally, we find that slight adjustments to the current hospital’s policy, which account for the dynamic nature of the admission control problem while still exploiting physician discretion in the admission decisions, can improve outcomes without increasing costs.

Keywords: healthcare delivery, empirical operations management, dynamic programming, capacity allocation, admission control, congestion, quality of service
1 Introduction

Intensive Care Units (ICUs) are specialized inpatient units that provide care for the most critically ill patients. They are extremely expensive to operate, consuming 15-40% of hospital costs (Brilli et al. 2001, Halpern et al. 2007, Reis Miranda and Jegers 2012) despite comprising less than 10% of the inpatient beds in the U.S. (Joint Position Statement 1994, Halpern et al. 1994). Most hospital ICUs operate near full capacity (Green 2003, Pronovost et al. 2004), making ICU beds a limited resource which must be rationed effectively. In this work, we evaluate how hospital managers are making ICU admission decisions currently, and examine what could be changed to improve the process, how to generate the necessary information to help make these decisions, and how these decisions should vary under different scenarios.

The obvious criteria for ICU admission is that very sick and unstable patients should be treated in the ICU, while stable patients do not require ICU care. However, determining the most unstable patients is a complex task that is subject to high variability depending on the training and experience of the particular physician on staff (Boumendil et al. 2012, Chen et al. 2012, Mullan 2004). A critical care task force established ICU admission, discharge, and triage standards that are highly subjective in nature; the task force even admits that “[t]he criteria listed, while arrived at by consensus, are by necessity arbitrary” (Task Force of the American College of Critical Care Medicine, Society of Critical Care Medicine 1999). Indeed, the medical community has started to point to a need to develop systematic criteria for ICU care; see Kaplan and Porter (2011) and Chen et al. (2013).

Critical to this need, our work is the first to estimate the benefit of ICU care for all medical patients admitted to the hospital through the Emergency Department (ED). We focus on patients admitted through the ED, who typically exhibit high uncertainty in the volume and severity of incoming patients, and whose care is the most likely to be affected by not only each patient’s medical severity but also hospital operational factors. For ethical reasons, it is not possible to run a field experiment to randomize ICU treatment to patients to estimate this benefit. Prior research has used observational data to measure the impact of ICU treatment on patient outcomes (e.g., Sprung et al. (1999), Shmueli et al. (2004), Simchen et al. (2004), Simpson et al. (2005), Iapichino et al. (2010), Kc and Terwiesch (2012), Louriz et al. (2012) ). We also utilize data from 15 hospitals covering over 190,000 hospitalizations (of which we consider the admission decisions of over 70,000 patients).

Working with observational data to answer our questions brings an important econometric challenge: the decision to admit a patient to the ICU is endogenous and this can generate biases in estimating the benefit of ICU admission. Specifically, there are discretionary patient health severity factors which are accounted for by the deciding physicians but unobserved in the data; this unobservable information that goes into the admission decision will be positively correlated with ICU admission and adverse patient outcomes, generating a positive bias in the estimate of the causal effect of ICU care on patient outcomes. Kc and Terwiesch (2012) and Shmueli et al. (2004) propose using the congestion level of the ICU (which can affect patients’ access to ICU care) as an instrumental variable (IV) to address this endogeneity problem. To be a valid IV, ICU congestion should affect patient outcomes only through its effect on the access to ICU care. But since hospital resources are shared among patients, a congested inpatient unit could directly impact the patient’s recovery during his stay in the unit, invalidating the required exogeneity assumption of the IV. Unlike
these prior studies, our data has detailed information on every unit each patient visits, which allows us to separate the
effect of ICU congestion on the admission decision from its effect during the patient’s hospitalization period, thereby
validating the IV identification strategy. Based on these detailed data, we also construct and test additional IVs based
on physician’s behavioral aspects that influence the admission decision. Many U.S. hospitals have started to collect
data similar to the one used in this work, and so the proposed methodology is applicable in other hospital settings.
Moreover, the fact that our study covers 15 hospitals of different sizes, specialties, and locations helps to validate the
robustness and generalizability of our results.

Hospitals face the tradeoff of admitting a new incoming patient to the ICU versus reserving a bed for a more
critical patient that could arrive in the near future. Optimizing this ICU admission criteria requires estimating the
cost of denying ICU care to every incoming patient; as discussed in Kaplan and Porter (2011), objective criteria to
characterize these re-routing costs is generally lacking. Using the estimates of the econometric analysis described
above, we characterize the admission control policy which optimally trades-off these competing objectives using an
analytic model that is based on objective and observable metrics available for every patient admitted via the ED. In
this context, an important contribution to the prior work by Shmueli et al. (2003) is that we formalize a number of
structural properties of the optimal ICU admission control policy. In particular, we demonstrate the optimality of
congestion-dependent admission control; such results have not been derived in the prior literature. This brings an
important theoretical basis for the ICU admission policy presented in this work.

Most importantly, we compare the performance of the derived optimal policies to the current policy used at the
hospitals in our study. While the admission criteria we develop are optimal given the objective patient risk metrics
available in our econometric study, current admission criteria used by hospitals also leverage doctors’ discretion to
make admission decisions. This discretion has potential to be highly informative in assessing the costs of denying ICU
admission but may be hard to record into objective patient metrics. Hence, it is possible for the optimal policies we
develop, which are based on objective metrics alone, to underperform relative to the currently used admission policies
which also exploit unobservable patient characteristics taken into account by the doctors. We use our estimated model
of the hospital’s current admission policy to simulate the current system and compare its relative performance vis-à-
vis a system which uses our derived optimal policy. We find that the proposed optimal admission policies that use
objective patient severity metrics can outperform the current policy on certain measures of patient outcomes, but not
all of them. For this reason, we also propose some alternative policies which adjust the current hospital admission
policy by accounting for the dynamics involved in the ICU admission decision while still taking advantage of doctors’
discretionary assessment of patient risk; these modifications can help to improve system performance on all of the
patient outcomes studied. An interesting managerial insight from this analysis is to determine under which settings it
is useful to make centralized admission decisions based on objective criteria alone versus allowing for local decision
making which incorporate discretionary criteria. While this question has received interest in other areas of Operations
Management (e.g., see Anand and Mendelson (1997) and Phillips et al. (2013)), to the best of our knowledge it has
not been studied in the healthcare operations literature.

Closest to our work is Shmueli et al. (2003) that examine the impact of denied ICU admission on mortality. They
consider patients who have already been referred for ICU admission and use an IV approach to measure how ICU admission decreases mortality for patients of different severity levels. Focusing on a sub-sample of patients pre-selected for ICU care has several drawbacks which we can address in our research design. First, Shmueli et al. (2003) use a severity measure (APACHE II) to measure the impact of ICU admission. This metric is generally assigned based on data available within the first 24 hours of ICU stay (Strand and Flaatten 2008), and so is not possible to use when considering which (of all) ED patients should be referred to the ICU. We instead develop admission criteria using metrics available to all patients in the ED. Second, their ICU admission criteria cannot be generalized to the (much larger) cohort of patients admitted from the ED. (In their study, 84% of patients are admitted to the ICU whereas in our sample, only 9.9% are admitted.) In particular, the benefit of ICU care may be exaggerated in Shmueli et al. (2003) because they only consider patients whose physicians have already determined that they require ICU care, whereas we are able to identify patients who will and will not benefit greatly from ICU care. Third, there is likely substantial variation in which patients will be recommended for ICU admission across hospitals and physicians due to heterogeneity in physicians’ backgrounds, training, and opinions as documented in Mullan (2004), Weinstein et al. (2004), Fisher et al. (2004), O’Connor et al. (2004). In a sequel study to Shmueli et al. (2003), Shmueli and Sprung (2005) explicitly discuss that the admission policy in the ICU they are studying does not maximize the benefits of the ICU, and that “the discrepancies actually originate from [an] inappropriate referral policy.” Our study provides criteria to use before any subjectivity in the pre-selection process can play a role. Fourth, we make important contributions by studying a number of different patient outcomes beyond mortality. This becomes important when the impact on mortality is similar across many patients, but highly variable in other outcomes such as length-of-stay (LOS) and readmission. Accurately quantifying these effects is necessary when determining the optimal ICU admission decision.

Our work contributes to the growing literature in healthcare operations management studying mechanisms to manage ICU capacity. Allon et al. (2013) study how congestion in inpatient units can result in increased ambulance diversion. Thus, they examine provision of care via preventing patients arrivals to the hospital, rather than examining the best inpatient unit for a patient following hospital admission, as we do. Kc and Terwiesch (2012) consider discharging current patients when the cardiac surgical ICU is busy and measure the implication of these speed-ups on patient readmission to the ICU and total length of stay. We argue that admission and discharge decisions are fundamentally very different and that they utilize different information and criteria. Hence, the detailed understanding of the discharge decision established in Kc and Terwiesch (2012) cannot provide insight into the admission decision we study here.

Our work is also related to previous empirical and analytical work in healthcare operations management that studies the effect of workload and congestion on healthcare productivity, albeit in different settings than the ICU admission decision. On the empirical side, Kc and Terwiesch (2009) show that hospital congestion can accelerate patient transportation time within the hospital; Green et al. (2013) find that nurse absenteeism rates in an ED are correlated with anticipated future nurse workload levels; Kc and Staats (2012) show that surgeon experience leads to better outcomes; Jaeker and Tucker (2013) report that the length of inpatient stays depends on current workload as well as the predictability and the pressure level of the incoming workload; and Batt and Terwiesch (2012) find
workload-dependent service times in the ED.

In summary, we make the following key contributions:

- **Management of ICU admissions:** We build an important foundation for a systematic decision support for ICU admission decisions. Using a large patient-level dataset of over 190,000 hospitalizations across 15 hospitals, we quantify the cost of denied ICU admission and use this to provide insight into which patients to recommend for ICU admission under various conditions. This work is the first to develop purely objective ICU admission criteria (i.e. we do not pre-select ICU-eligible patients). We compare the derived optimal admission policies with the current hospital admission policies; we discuss under what circumstances it is useful to base admission decisions on objective metrics of patient risk alone versus allowing for discretionary criteria in the admission decision.

- **Patient Outcomes:** In order to derive the optimal admission policy, we quantify the impact of ICU admission on a number of patient outcomes including hospital LOS, hospital readmission, and patient transfers to higher levels of care. We demonstrate that the impact of ICU admission is highly variable for different patients and different outcomes. Thus, it is important to have an understanding of all of these when making admission decisions. We also make methodological contributions in this context, improving upon previously developed instrumental variable approaches to address endogeneity biases that arise in this estimation problem.

The rest of the paper is organized as follows. Section 2 describes the context of the problem and the data used in this empirical study. Section 3 develops the econometric model of ICU admission decisions and provides its estimation results. Section 4 studies the effect of admission decisions on various patient outcomes. Section 5 analyzes the robustness of our results under the presence of alternative mechanisms to handle ICU congestion. Section 6 uses the empirical results to develop a simulation study to compare the performance under the current ICU admission policy used by hospitals with alternative approaches. Section 7 summarizes our main contributions and provides guidelines for future research.

## 2 Setting and Data

We employ a large patient dataset collected from 15 hospitals, comprising of nearly 200,000 hospitalizations over the course of one and a half years. The hospitals are within an integrated healthcare delivery system, where insurers and providers fall under the same umbrella organization. The majority of patients treated within the system’s hospitals are insured via this same organization. This allows us to ignore the potential impact insurance status may have on the care pathway of individual patients. However, we expect that our results can be extended to other hospitals that treat patients with heterogeneous insurance coverage.

In these 15 hospitals, inpatient units are broadly divided according to varying levels of nurse-to-patient ratios, treatment, and monitoring. The ICUs have a nurse-to-patient ratio of 1:1 to 1:2. There are two other kinds of inpatient units: general wards with ratios 1:3.5 to 1:4 and intermediate care units with ratios 1:2.5 to 1:3 (not all hospitals have
intermediate care units). While there is some differentiation within each level of care, the units are relatively fungible, so that if the medical ICU is very full, a patient may be admitted to the surgical ICU instead.

Patient-level information in our dataset includes patient age, gender, admitting diagnosis, hospital, two severity of illness scores (one based on lab results and comorbidities and the other a predictor for in-hospital death). In addition, we collect operational data that includes every unit each patient visits along with unit admission and discharge dates and times. Since we have an inpatient dataset, we do not have information on patients who are discharged directly from the ED.

In the rest of this section, we first describe different mechanisms that can be used to manage ICU capacity as well as related work in this subject. We then describe the sample selection procedure for the data used in this study.

2.1 Managing ICU capacity

Within the Operations Management (OM) and medical literature, several empirical studies have examined how hospitals utilize adaptive mechanisms to navigate periods of high ICU congestion. When a hospital does not have sufficient downstream bed capacity, surgical cases may be either delayed or canceled (Cady et al. 1995). When a new patient requires ICU care, but there is no available bed, he may be delayed and board in another unit, such as the ED or the post-anesthesia care unit (Ziser et al. 2002, Chalfin et al. 2007). An econometric study by Louriz et al. (2012) shows that a full ICU is the main factor associated with late ICU admission. Furthermore, Allon et al. (2013) shows that ED boarding caused by a congested ICU is an important factor driving ambulance diversion.

A mechanism that has received considerable attention from the OM and medical communities is to speed up the treatment of current ICU patients to accommodate new, potentially more critically ill patients. Anderson et al. (2011) investigate daily discharge rates from a surgical ICU at a large medical center, and find higher discharge rates on days with high utilization and more scheduled surgeries. Kc and Terwiesch (2012) study the effect of ICU occupancy level on discharge practices in a cardiac surgical ICU. They find that congested ICUs tend to speed-up the treatment of their patients and that these affected patients tend to be readmitted to the ICU more frequently.

Yet another alternative to manage ICU capacity is to control the admission of patients. During periods of high congestion, some patients who may benefit from ICU care might be denied access because the ICU is full or all available beds are being reserved for more severe incoming patients. ICU congestion is an important factor affecting ICU admission decisions (Singer et al. 1983, Strauss et al. 1986, Vanhecke et al. 2008, Robert et al. 2012). Other studies have obtained similar results in international hospitals: Escher et al. (2004) in Switzerland, Azoulay et al. (2001) in France, Shmueli et al. (2004), Shmueli and Sprung (2005) and Simchen et al. (2004) in Israel, and Iapichino et al. (2010) in seven countries, including Italy, Canada, and UK.

The above discussion suggests that all of these mechanisms – delayed ICU admission, speed-ups, and ICU admission control – are used to manage ICU capacity in various settings. However, it is hard to find standards for when and

---

1I.e. chronic diseases, such as diabetes, that may complicate patient care and recovery.

2These multiple severity of illness scores reflect the complexity in defining objective severity of illness measures. Table 1 explains patient characteristics in detail.
how these mechanisms should be used; often there is substantial subjectivity in defining best practices. In a recent exploratory study, Chen et al. (2013) discuss the lack of standards in the field and point to a need to utilize Electronic Health Records to gain a better understanding of who benefits from ICU care in order to facilitate improved ICU triage decision making.

Indeed, our study utilizes data from a comprehensive Electronic Medical Records system. We focus on the ICU admission decision for patients that were admitted to the hospital through the ED to a medical service; in our data, about 55% (52%) of patients admitted to the hospital (ICU) are admitted via the ED to a medical service. The admission process works as follows. If an ED physician believes a patient is eligible for ICU admission, an intensivist will be called to the ED for consultation. While the intensivist has the ultimate decision about whether to admit the patient from the ED, the decision is typically a negotiation between the two physicians as to what the individual patient’s needs are and what resources (e.g. ICU versus non-ICU beds) are available. The medical necessity of a patient plays a key role in the ICU admission decision, but the assessment of this necessity likely differs across physicians depending on his/her background and training (Mullan 2004, Weinstein et al. 2004, Fisher et al. 2004, O’Connor et al. 2004). Moreover, most hospitals lack a universal metric that characterizes the severity of patients admitted via the ED, making it challenging to determine which patients should be considered for ICU care.

Most of the aforementioned studies on ICU admission control use patient severity measures which are based on scoring systems available only after patients are admitted to an ICU (Strand and Flaatten 2008). Examples include the Acute Physiology and Chronic Health Evaluation II (APACHE II) scores (Shmueli et al. 2004, Shmueli and Sprung 2005), Simplified Acute Physiology Score (SAPS II) (Iapichino et al. 2010, Simchen et al. 2004), Simplified Therapeutic Intervention Scoring System (TISS) (Simchen et al. 2004) and Mortality Prediction Model (MPM) (Louriz et al. 2012). These measures of patient severity are not available for a typical ED patient and hence, as argued by Franklin et al. (1990), they cannot be used to decide which patients should be routed to the ICU. In contrast, the hospitals we analyze use a uniform metric of patient severity available for all admitted patients: the Laboratory Acute Physiology Score (LAPS) (see Escobar et al. (2008) for details and validation of this metric). Previous work by van Walraven et al. (2010) show that LAPS is a reasonable predictor of patient length of stay and mortality. Utilizing this measure, we can analyze ICU admission decisions for all ED patients, and not just the patients who have been pre-screened for admission under subjective criteria, as done in prior work.

Our work takes an important step towards quantifying the costs/benefits of ICU admission. Currently, most hospitals lack such measures, making it practically impossible to develop rigorous, evidence-based ICU care standards (Kaplan and Porter 2011). Although our focus is admission control, we conduct additional empirical analysis that accounts for other mechanisms mentioned above (see Section 5).

2.2 Data Selection

Figure 1 illustrates our data selection process. Hospitals in our dataset come from an integrated healthcare delivery system and had heterogeneous sizes of inpatient units. Because defining congestion in a small ICU is challenging and different mechanisms might be used to allocate beds in small ICUs, we consider only the patients who are treated in
hospitals with ICUs of ten or more beds. There were 15 such hospitals; among them the average ICU size was 21 beds (the largest having 37 beds), and the average percentage of ICU beds among inpatient beds was 12.9% with minimum of 9.3% and maximum of 21.5%.

We utilize patient flow data from all of 192,409 patient visits in the selected 15 hospitals (indicated by one star in Figure 1) to derive the capacity and instantaneous occupancy level of each inpatient unit. Because our dataset consists of patients admitted and discharged within the 1.5 year time period, we restrict our study to the 12 months in the center of the period to avoid censored estimation of capacity and occupancy. We exclude patients who experienced inter-hospital transport as it is difficult to determine whether it was due to medical or personal needs. Because of the reasons explained in Section 2.1, we focus on the patients who are admitted via the ED to a medical service. The sizes of the inpatient units were quite stable over our study period. However, four hospitals had a small change in the capacity of the intermediate care unit and we exclude patients who are hospitalized during these rare occurrences of intermediate care unit reorganization (such as reducing the number of beds). Our final dataset consists of 70,133 hospitalizations, as indicated by two stars in Figure 1.

3 Impact of Congestion on ICU Admission

This section develops an empirical model to study the ICU admission decision process for ED patients. To develop this empirical model, we first present a stylized model of ICU admission control, which is similar to the model developed by Shmueli et al. (2003). Section 3.1 describes this model and characterizes the structure of the optimal solution. In particular, the model captures the effect of ICU congestion on the decision to admit new patients to the ICU. This structure is then used in Section 3.2 to develop an empirical model of ICU admission control that can be estimated using our data; this model’s main objective is to measure the effect of congestion and other operational/behavioral factors on ICU admission decisions. Section 3.3 discusses the estimation results of this model.3

3.1 The ICU Admission Control Problem

We model the ICU admission control problem through a discrete-time/finite-horizon version of the Erlang Loss Model used in Shmueli et al. (2003). There is a finite-horizon, $T$, and time periods are discretized and indexed by $t$. In each time period, a single patient arrives for potential ICU admission with probability $\lambda$. If a new patient arrives, his risk, $p \in (0, \bar{p}]$, is randomly distributed according to a cumulative distribution function, $F(\cdot)$. We use $p = 0$ to denote the absence of a new patient. All patients who arrive are candidates for ICU admission, and the decision is whether to route the patient to the ICU or to a non-ICU (the model excludes patients discharged directly from the ED). In each period, patients being treated in the ICU are discharged with probability $\mu$, irrespective of their risk level and length

3 Although the empirical model is based on an analytical model of admission control, we do not conduct a structural estimation. The analytical model is used to gain insights on the relevant factors that affect the admission decision which are useful to specify the empirical model. We do not estimate primitives of the analytical model or assume that the hospital admissions are made optimally, as is typically done with a structural estimation approach.
of stay in the ICU (i.e. the service time in the ICU is memoryless). This probability, $\mu$, can be viewed as an average probability over all patient risk levels. Moreover, patient discharge is exogenous, i.e. there is no speed-up in the ICU.\footnote{As discussed in Section 2.1, other mechanisms may be used. However, in order to focus on the tradeoff between admitting now versus saving space for a potentially more severe patient, we only examine admission control. Via numerical analysis, we found that the qualitative results extend when speed-ups are incorporated.} There is a limited capacity of ICU beds, given by $B$. Because we are concerned with the allocation of ICU beds, we assume there is ample space in the other inpatient units to care for all patients.

Let $x \in [0, B]$ denote the number of patients currently being treated in the ICU. When a new patient arrives, the doctor must decide whether to admit or reroute the arriving patient to the ICU, given the current available capacity $B - x$. If there are no available ICU beds ($x = B$), the new patient must be denied ICU admission. If the patient is denied admission, a cost, $\Phi(p) \geq 0$ is incurred (when no patients arrive this cost is zero, $\Phi(\emptyset) = 0$). If the patient is admitted to the ICU, no cost is incurred. For the purpose of our discussion, this cost captures the clinical cost for admitting a patient to a non-ICU (e.g., this could be the increase in readmission risk due to denied admission). Section 6 shows how to estimate this clinical cost from our empirical models. We assume that being denied admission to the ICU is more detrimental to more severely ill patients:

**Assumption 1** The cost for denying ICU admission, $\Phi(p) \geq 0$, is non-decreasing in $p$.

A policy is defined as a decision rule that chooses whether to admit (A) or reroute (R) an incoming patient, for each possible state characterized by the severity of the incoming patient $p$, the number of occupied ICU beds $x$, and the current period $t$. The online appendix derives structural properties of the optimal policy that minimizes the expected total costs during the time horizon $T$. The key results can be summarized in the following theorem.

**Theorem 1** The optimal admission control policy satisfies the following properties:

1. It is a threshold policy: there exists a threshold $\kappa(x, t)$ – which depends on the current bed occupancy and time period – such that a new patient is admitted if and only if $p \geq \kappa(x, t)$.

2. In any time period $t$, the threshold $\kappa(x, t)$ is non-decreasing in the number of patients in the ICU, $x$.

3. For rerouting costs of the form $\Phi(p) + C$, the optimal threshold is non-decreasing in $C$.

The next section develops an empirical model based on the insights provided by this analytical model of admission control.

### 3.2 Econometric Model for ICU Admission

Following the notation in Section 3.1, let $p_i$ denote the designated risk level of patient $i$ upon arrival to the ED. Although we do not observe $p_i$ directly in the data, there are several observable metrics that are presumably related to a patient’s designated risk. We let $X_i$ be a row-vector of covariates containing these metrics (in our application these are age, gender, two kinds of severity scores, and admitting diagnosis) as well as seasonality controls (month, day, and
time of admission) and an intercept; details can be found in Table 1. In addition, there are some additional risk-related metrics which are observed by the hospital but are not contained in the data; we denote these by the unobservable term $u_i$. Patient $i$’s designated risk level is modeled as:

$$ p_i^* = X_i \theta + u_i, \quad (1) $$

where $\theta$ is a column-vector of parameters to be estimated. Since $p_i^*$ is not observed in the data, it will be treated as a latent factor affecting the admission decision (the star in the notation emphasizes the latent nature of the risk level measure). Based on the first result from the admission control problem in the previous section, we assume the admission decision follows a threshold policy. Let $\kappa_i$ denote the admission threshold, which could vary across patients and is modeled as:

$$ \kappa_i = Z_i \alpha + e_i, $$

where $Z_i$ is a set of observable covariates that affect the admission threshold (to be specified later in the section); the term $e_i$ denotes other unobservable factors determining the threshold. Admission to the ICU occurs when $p_i^* \geq \kappa_i$.

Defining the error term $\xi_i = u_i - e_i$, we can model the admission decision $r_i$ through the latent model:

$$ r_i = \begin{cases} 
\text{admit to ICU} & \text{if } X_i \theta - Z_i \alpha + \xi_i \geq 0, \\
\text{re-route to Ward} & \text{otherwise}. 
\end{cases} \quad (2) $$

Assuming the error term $\xi_i$ follows a Standard Normal distribution, the model becomes a Probit regression and the vector parameters $(\theta, \alpha)$ can be estimated via Maximum Likelihood Estimation (Wooldridge 2010).

Next, we define the covariates in $Z$. The first covariate is motivated by result 2 of Theorem 1. We include a variable which captures the observed ICU occupancy level around the time of the admission decision of patient $i$. Since the actual time of the admission decision is unknown, we use the ICU occupancy one hour before the time patient $i$ is discharged from the ED, as illustrated in Figure 3. We consider the ICU occupancy level at this specific time for two main reasons: 1) the decision of where to admit a patient can change during the entire ED boarding time – defined as the elapsed time between the time the hospital admission decision is made and the time the patient leaves ED – so we want to capture the occupancy closest to when the final admission decision is made and; 2) admissions are not instantaneous, so we use the occupancy an hour before the actual physical arrival of the patient to the inpatient unit. We also tried alternative measures of occupancy (e.g., 2 hours before first inpatient unit admission) that yield similar, but slightly weaker results.

It is also important to account for the non-linear effect of ICU occupancy on the admission threshold. This can be done by including the ICU occupancy in levels (using indicator variables), where the levels should be defined so that they capture the relevant changes to the admission threshold. Therefore, we used the data to identify at which levels of occupancy the ICU admission rates get adjusted. Figure 2 illustrates this for when the occupancy is at the 95th percentile observed in the data and higher. The x-axis indicates 20 different patient severity classes defined by their LAPS score, a reasonable proxy of patient severity (see Table 1 for details). For each severity group, the graph shows the actual percentage of patients who are admitted to the ICU among ED patients who saw, during the
last hour in the ED, low occupancy in the ICU (below the 95th percentile) and among the patients who saw high ICU occupancy (95th percentile and above); note that all 40 points in this graph have enough observations, with the smallest sample size being 144 patients. This figure shows that ICU admission decisions for patients at all severity levels are affected by ICU occupancy; among patients in the same severity group, a lower percentage of patients who saw high ICU occupancy was sent to the ICU compared to the patients who saw low ICU occupancy level. We repeated the exercise for other cutoffs of ICU congestion: at the 90th, 85th and 75th percentiles. The change in admission rate was much smaller and non-existent for some groups of patients. While we considered multiple different measures of ICU occupancy, based on this analysis, we defined \( ICU_{Busy_i} \) as the dummy variable indicating an ICU bed utilization greater than or equal to the 95th percentile one hour prior to patient \( i \)'s discharge from the ED.

In addition to \( ICU_{Busy_i} \), another set of covariates was included in \( Z \) to capture behavioral factors affecting the ICU admission decisions. The first behavioral variable, \( RecentDischarge_i \), accounts for recent discharges from the ICU and is motivated by the anecdotal evidence we gained from interviews with doctors. ICU discharges typically release the nurse who has been monitoring the discharged patient. The intensivist in charge may have an incentive to “preserve the nurse hours” by demonstrating a continuous demand for those nurses even after patients are discharged.\(^5\) This would lead to higher ICU admission rates right after one or more ICU discharges. Note that this behavior is different from the speed-up effect reported in Kc and Terwiesch (2009) because it can also be manifested when discharges are not “forced” to occur faster. It is also different from the ICU occupancy effect because it can operate when the ICU has low utilization. To measure \( RecentDischarge_i \), we count the number of all ICU discharges in the 3-hr window before patient \( i \)'s admission to the first inpatient unit. In the sample, 56% of the patients see no recent ICU discharges, 27% see one discharge, and 11% see two discharges. Because bigger ICUs would naturally have more recent discharges, we divide the number of recent ICU discharges by the ICU capacity of each hospital to use it as \( RecentDischarge_i \).

The second behavioral variable, \( RecentAdmission_i \), accounts for the number of recent admissions of ED patients to the ICU. Since ICU beds are shared between ED and elective patients, a high number of recently admitted ED patients may reduce the bargaining power of the ED physician in his negotiation with the intensivist. To measure \( RecentAdmission_i \), we consider ICU admissions in the 2-hr window before patient \( i \)'s admission to the first inpatient unit, but count as a recent admission only if the patient is admitted via the ED to a medical service (excluding those that go to surgery, as in that case the negotiation may involve the surgeon). Because of shift changes, we do not expect the impact of expending negotiation power to propagate for extended periods of time. In our data, 84% of the patients see no recent admission and 14% see one recent admission. Similar to \( RecentDischarge_i \), we divide the number of recent admissions by the ICU capacity of each hospital to define \( RecentAdmission_i \).

The third behavioral variable, \( LastAdmitSeverity_i \), measures the severity of the last patient admitted to the ICU from the ED. The motivation for including this variable is that the most recent admit serves as a reference point in the negotiation process: if he was a very severe patient, this may cause the ED physician to require a new patient to be

\(^5\)This behavior is related to supply-sensitive demand that has been shown in the medical literature. For instance, see Wennberg et al. (2002) and Baker et al. (2008).
very sick to recommend ICU admission. We defined \textit{LastAdmitSeverity}, as a dummy variable indicating whether the last patient admitted to the ICU had a LAPS score greater than or equal to the 66\textsuperscript{th} percentile value of the observed LAPS distribution. Summary statistics of the covariates for all the patients in our sample are described in Table 2, also grouped by whether they were admitted or not to the ICU.

Finally, \(\eta_{h(i)}\) is a hospital indicator variable which controls for any differences across hospitals that can affect the thresholds for ICU admission. In particular, result 3 from Theorem 1 suggests that the thresholds depend on the structure of the rerouting cost (\(\Phi(p)\) in the model). For example, some hospitals in our sample have intermediate units of care with nurse-to-patient ratios that are inferior to the ICU but are above the general ward. The cost of denying ICU admission to a patient can be reduced by admitting this patient to an intermediate care unit.

As we later discuss in Section 4, the fact that the ICU admission, \(r_i\), is affected by unobservable patient severity factors generates some challenges in estimating the causal effect of ICU admission on patient outcomes. But because the covariates \(Z_i\) include factors unrelated to a patient’s severity condition that affect patient outcomes, these provide potential instrumental variables to identify the causal effect that we seek to estimate. We provide further details in Section 4.

### 3.3 Estimation Results of the ICU Admission Model

Table 3 summarizes the estimation results for the Probit model using the selected sample of patients. The upper part of Table 3 reports the coefficients for the \(Z_i\) covariates. Due to space limitations, the bottom panel of the table displays the estimated coefficients for a selected group of coefficients for patient severity factors (\(X_i\)); the complete set of estimates is provided in the online appendix.

The coefficient for \(ICUBusy_i\) is negative and highly statistically significant, providing strong evidence that higher ICU occupancy leads to lower probability of being admitted to the ICU. The estimated value translates to the probability of ICU admission decreasing from 0.104 to 0.049 on average—a 53\% decrease—when the ICU occupancy increases above the 95\textsuperscript{th} percentile.

The coefficient for \(RecentDischarge_i\) is positive and statistically significant. This implies that, for an average patient, the admission probability increases from 0.092 to 0.103—a 12\% increase—when the ICU has recently discharged one ICU patient. An additional recent discharge increases the admission probability to 0.115. This is consistent with the mechanism suggesting that intensivists have incentives to admit after a recent discharge to maintain demand for the nurses in the ICU. Recall that discharges can happen when there is ample space in the ICU, so a recent discharge does not necessarily correspond to a demand-driven discharge to make room for the incoming patient.

The coefficient for \(RecentAdmission_i\) shows a statistically significant negative effect; this is consistent with a reduction in the bargaining power of the ED physicians in negotiating the ICU admission of a patient when other patients have been recently admitted. For an average patient, the admission probability decreases from 0.100 to 0.093—a 7\% decrease—when the ED has recently sent one patient to the ICU.

\(LastAdmitSeverity_i\) has a statistically significant negative coefficient. This implies that the severity of the last admitted patient affects the reference point for the severity threshold at which patients should be admitted. For an
average patient, the admission probability decreases from 0.101 to 0.095—a 6% decrease—when the most recent admit is a very severe patient (which we define to be having a LAPS score greater than or equal to the 66th percentile value of the observed LAPS distribution).

The coefficients for hospital indicator variables also showed statistically significant results. An F-test of joint significance of the hospital indicator variables rejects the null hypothesis that all hospital indicators are zero, with p-value less than $10^{-4}$.

ICUBusy$_i$ has the largest impact on the ICU admission rate among all the operational and behavioral factors studied. We also estimated specifications that exclude RecentAdmissions$_i$, RecentDischarges$_i$, and LastAdmitSeverity$_i$, and the coefficient of ICUBusy$_i$ was similar in magnitude and significance.

In summary, the empirical results show that, although medical necessity plays a key role in ICU admissions, operational factors such as the ICU occupancy also determines which patients receive the ICU care. At high levels of congestion, patients that would otherwise receive ICU care are not admitted, and this effect persists even when the ICU is not completely full. A related important question is to quantify the effect of this admission policy on patient outcomes. The next section develops an econometric model to analyze this empirical question.

4 Impact of ICU Admission on Patient Outcomes

In this section, we study how access to ICU care affects several patient outcomes. To do so, we begin by defining several measures of patient outcomes of interest in Section 4.1. In addition to the traditional measures, such as mortality and hospital length of stay, we were able to construct other useful measures which exploit the rich information provided in our data covering the complete path a patient follows within and after the hospital stay. Next, Section 4.2 develops an econometric model to measure the impact of ICU care on these outcomes. The main challenge in this estimation is to account for the endogeneity in ICU admission decisions, for which we use the Instrumental Variables (IVs) estimation as an identification strategy. Section 4.3 reports the main results of this estimation and 4.4 shows additional analysis showing the robustness of these results to alternative specifications.

4.1 Measuring Patient Outcomes

To quantify the benefit of ICU care, we focus on four types of patient outcomes: (1) in-hospital death ($Mortality$), (2) hospital readmission ($Readmit$), (3) hospital length of stay (LOS) ($HospLOS$), and (4) transfer-up to a higher level of care ($TransferUp$). $Mortality$, $Readmit$, and $HospLOS$ are fairly standard patient outcomes used in the medical and OM communities (e.g. Iezzoni et al. (2003) and Kc and Terwiesch (2009)). We consider one additional measure of patient outcome, $TransferUp$, for the following reason. Typically, a patient will be transferred to an inpatient unit with lower level of care or be discharged from the hospital as his health state improves. Being transferred up to the ICU can be a sign of physiologic deterioration and such patients typically exhibit worse medical conditions (Luyt et al. 2007, Escobar et al. 2011). Accordingly, a $TransferUp$ event is defined as a patient’s transfer to the ICU from
an inpatient unit with lower level of care.\footnote{ICU readmission, which qualifies as a TransferUp event, has also been shown to lead to higher mortality and length of stay (Durbin Jr and Kopel 1993).} Note that patients who were admitted to and directly discharged from the ICU can never show this event, and so we study TransferUp over the subset of patients who visited the general ward at least once during their hospital stay.

Defining readmission requires specifying a maximum elapsed time between consecutive hospital discharges and admissions. As this elapsed time increases, it becomes less likely that the complications were related to the care received during the initial hospitalization. Hence, after discussions with doctors, we defined a relatively short time window for hospital readmission – within the first two weeks following hospital discharge. When analyzing Readmit, we did not include patients with in-hospital death as they cannot be readmitted.

We let HospLOS measure the time from admission to the first inpatient unit until hospital discharge time, excluding the ED boarding time. A complication in analyzing HospLOS is that its histogram reveals “spikes” every 24 hours. This is because of a narrow time-window for hospital discharge: more than 60% of the patients are discharged between 10am and 3pm, whereas admission times are less concentrated and demonstrate a markedly different distribution (a similar issue was reported in Armony et al. (2011) and Shi et al. (2012) on data from other hospitals). To avoid this source of measurement error, we measure HospLOS as the number of nights the patient stayed in the hospital. In studying HospLOS, we include patients who died during their hospital stay. The results are similar if we exclude patients with in-hospital death.

Table 4 provides summary statistics of our patient outcome variables.

### 4.2 Econometric Model for Patient Outcomes

An ideal thought experiment to examine the implications of ICU admission on patient outcomes would be randomizing treatments to patients by allocating patients to the ICU and non-ICU units regardless of their severity condition. Of course, such an experiment would be impossible in practice due to ethical concerns. This limits us to work with observational data, which brings important challenges to the estimation, as we now describe.

Let $y_i$ denote a measure capturing a patient outcome of interest (e.g., HospLOS). There is extensive work in the medical literature that provides several patient severity measures that are useful in predicting patient outcomes. For example, Escobar et al. (2008) and Liu et al. (2010) illustrate how severity measures based on automated laboratory and comorbidity measures can be used to successfully predict in-hospital mortality and hospital length of stay, respectively. As before, let $X_i$ denote those patient severity factors as well as seasonality controls that are observed in the data. We also control for hospitals, and let $\omega_{h(i)}$ denote the coefficients for a set of hospital indicator variables where $h(i)$ is patient $i$’s hospital. We model patient outcome $y_i$ as a random variable with distribution $f(y_i|\beta_1, \beta_2, r_i, X_i, \omega_{h(i)})$, where the parameters $(\beta_1, \beta_2)$ capture the effect of the admission decision $r_i$ (defined in equation (2)) and $X_i$ on the patient outcome, respectively. For example, this distribution could be given by a model of the form:

$$
\log(y_i) = \beta_1 r_i + X_i \beta_2 + \omega_{h(i)} + \varepsilon_i, 
$$

(3)
with the error term $\varepsilon_i$ following a normal distribution so that $y_i$ is log-normally distributed. In this example, we have a linear regression with Gaussian errors, but our framework allows for more general specifications (e.g., binary patient outcomes).

The linear regression example (3) is useful to illustrate the main estimation challenge. A naive approach to estimate the effect of ICU admission on $y_i$ is to include the actual admission of the patient ($r_i$) as a covariate in the regression (3) and interpret the Ordinary Least Square (OLS) estimate of $\beta_1$ as the causal effect of ICU admission on the outcome. This approach ignores that the admission decisions are endogenous; patient severity conditions that are unobservable in the data, such as the cognitive state of the patient, are likely to affect admission decisions. Figure 4 illustrates this endogeneity issue in further detail. Note that both admission decisions and patient outcomes are affected by $X_i$ and $\xi_i$. Because $\xi_i$ is unobserved, it will be absorbed as part of the error term $\varepsilon_i$ of model (3). Hence, the covariate $r_i$ in the outcome model will be positively correlated with the error term $\varepsilon_i$, violating the strict exogeneity assumption required for consistent estimation through OLS. This endogeneity could introduce a positive bias in the estimate of the effect of ICU admission on patient outcomes, underestimating the value of ICU care (because we expect $\beta_1$ to be negative).

An alternative is to use the Instrumental Variables (IVs) estimation to obtain consistent estimates of this linear regression model. A valid instrument should be correlated with the admission decision $r_i$ but unrelated to the unobserved patient severity factors $\varepsilon_i$ determining the outcome $y_i$. The set of covariates, $Z_i$, accounting for operational factors and behavioral aspects of the admission model (2) are potential instruments because: (1) they should be unrelated to the patient-specific risk factors of a new incoming patient; and (2) they do affect the admission decision, as validated with the empirical results of Section 3.3.

Validating the Instrumental Variables

The results in Table (3) show that $ICUBusy_i$ is a statistically powerful instrument, in the sense that it explains significant variation in the admission decision $r_i$. However, for $ICUBusy_i$ to be a valid instrument it also has to be uncorrelated with the unobservable factors $\varepsilon_i$ that affect patient outcomes. Kc and Terwiesch (2012) describe a potential mechanism that could lead to a violation of this assumption. They show that readmission rates tend to be higher for patients who experienced high ICU occupancy level during their ICU stay. Moreover, the same effect could apply to other inpatient units visited by the patient.

To overcome this issue, we used the detailed information in our data about the complete care path of patients to control for the congestion levels that a patient experienced in each of the visited inpatient units during his hospital stay. Specifically, let $D_i$ be the set of days patient $i$ stayed in the hospital (after leaving the ED) and $Occ_{i,d}$ the occupancy of the inpatient unit where patient $i$ stayed in day $d$. The average occupancy of the inpatient units visited by the patient during his hospital stay is defined as $AvgOccVisited_i = \frac{1}{|D_i|} \sum_{d \in D_i} Occ_{i,d}$ (see Figure 3 for details on the time-line where this measure is calculated from).\footnote{We define capacity of an inpatient unit as the 95th percentile of the bed occupancy distribution of that unit to compute $Occ_{i,d}$, because in many occasions, the maximal capacity is rarely observed as hospitals may temporary expand their standard capacity by a few beds in extreme circumstances (this was also pointed out in Armony et al. (2011) and Jaeker and Tucker (2013)). Given this definition, it is possible to have $Occ_{i,d}$ above 100%. The average $AvgOccVisited_i$ was 0.84 with median of 0.86 in our dataset.} We include $AvgOccVisited_i$ as an additional control variable in the outcome
model (in addition to the patient severity factors $X_i$). $AvgOccVisited_i$ is not perfectly correlated with $ICUBusy_i$ because the latter is measured before the patient is physically moved to the inpatient unit and the occupancy level typically varies during a patient’s hospitalization period; in our sample, the correlation between the two measures is 0.24. Separating the effect of occupancy on the admission decision from its effect during the inpatient hospital stay is essential to have a proper IV identification strategy. Note that previous works using ICU congestion as an instrument (e.g., Kc and Terwiesch (2012), Shmueli et al. (2004)) were not able to account for the congestion during the patient’s hospital stay.

Another mechanism that could invalidate the use of $ICUBusy_i$ as an IV is when periods of high congestion coincide with the arrival of very severe patients; this is what happens, for example, during an epidemic. We tested this potential mechanism by analyzing the relationship between hospital occupancy and the LAPS score, a validated measure of patient severity, and found no correlation between the two. Although this does not prove that the instrument $ICUBusy_i$ is uncorrelated with the unobservable factors affecting outcomes, there is no reason to believe that they would be related to occupancy given that reasonable observable proxies of severity are not (this approach was also used by Kc and Terwiesch (2012) to validate a similar instrument).

Overall, our analysis provides substantial support validating the use of $ICUBusy_i$ as an IV. With this IV approach, the identification is driven by comparing differences in outcomes among patients who have similar observable characteristics captured by $X_i$ but received different treatments because of the different levels of ICU occupancy at the time of their admission to an inpatient unit. Although this is not a perfectly randomized experiment, this identification strategy provides a valid approach to estimate the effect of ICU admission on patient outcomes.

The behavioral factors included in the $Z$ covariates of the admission model in Equation (3)–$RecentDischarge_i$, $RecentAdmission_i$, and $LastAdmitSeverity_i$– are also valid IVs because: (1) they should be unrelated to the new incoming patient’s specific risk factors; and (2) they indeed affect the admission decisions, as reported in Table 3. To validate the first point, we looked at the correlation between these three factors and the LAPS score of the incoming patient and found no correlation. This is expected given the randomness in the arrival process of new incoming ED patients. Nevertheless, because results in Table 3 suggest that these IVs are less powerful than $ICUBusy_i$ (in the sense that they explain less variation in the ICU admission decision), we also considered specifications that had $ICUBusy_i$ alone as an IV and the results were similar.

Estimation methods

When the patient outcome is modeled via a linear regression as in (3), we can use a standard two stage least squares (2SLS) approach to implement the IV estimation. But because admission decisions and some of our patient outcomes are discrete, a more efficient estimation approach is to use nonlinear parametric models via the Full Maximum Likelihood Estimation (FMLE) (Wooldridge 2010), described in detail next.

We provide two estimation models depending on whether the patient outcome is measured as a binary or a counting variable. We first consider the three binary patient outcomes $Mortality$, $TransferUp$ and $Readmit$. To model each
of these outcomes, we use a Probit model defined by a latent variable:

\[ y_i^* = \beta_1 r_i + X_i \beta_2 + \omega_{h(i)} + \beta_3 \text{AvgOccVisited}_i + \epsilon_i \]  

where \( y_i^* \) is the latent variable. The additional control \( \text{AvgOccVisited}_i \) captures the effect of the congestion during the hospital stay of the patient, as previously discussed. To account for the endogeneity in ICU admission decisions \( r_i \), we allow for the error term \( \epsilon_i \) to be correlated with the unobservable factors affecting admission (\( \xi_i \) in equation (1)) by assuming that the random vector \( (\xi_i, \epsilon_i) \) follows a Standard Bivariate Normal distribution with correlation coefficient \( \rho \) (to be estimated along with the other parameters of the model). Note that this requires a joint estimation of the ICU admission model (2) and the outcome model (4). The model becomes a Bivariate Probit which can be estimated via the Full Maximum Likelihood Estimation (FMLE) (Cameron and Trivedi 1998). The endogeneity of the admission decision \( r_i \) can be tested through a likelihood ratio test of the correlation coefficient \( \rho \) being different from zero.

The patient outcome defined by \( \text{HospLOS}_i \) is a count variable of the number of nights a patient stays in the hospital. A Poisson model could be used to model this count variable, but preliminary analysis of \( \text{HospLOS}_i \) reveals over-dispersion (Table 4 shows the mean of \( \text{HospLOS}_i \) is 3.9 while the variance is 24.0). Hence, we use the Negative Binomial regression, which can model over-dispersion using the parametrization developed in Cameron and Trivedi (1986). We use the extension developed by Deb and Trivedi (2006) to include a binary endogenous variable – the ICU admission decision \( r_i \) – into the negative binomial regression, which is estimated jointly with the Probit model (2). The negative binomial regression includes the same covariates as in (4). The next section describes the estimation results of all the outcome models.

### 4.3 Patient Outcome Model Estimation Results

In this section, we discuss the results of the patient outcome models, which are summarized in Table 5 and Table 6. As discussed in Section 4.2, we estimate the admission decision and patient outcome model jointly to account for the endogeneity of the admission decisions. In all cases the estimates of the admission decision model are similar to those reported in Table 3, so they are omitted for brevity. For space limitations, Table 5 and Table 6 show only the coefficient and the marginal effects of \( r_i \) (i.e., whether the patient was admitted to the ICU or not), which is the main focus of this analysis. Each row corresponds to a different outcome (the dependent variable).

In Table 5, the coefficients of \( r_i \) are negative and significant in all models except Mortality, suggesting that admitting a patient to the ICU reduces the chance of having an adverse outcome. (Later we discuss possible explanations for the lack of significance in the Mortality outcome model). The table also displays the average marginal effect (AME), which is the average expected absolute change in the outcome (among all patients) when a patient is admitted to the ICU instead of the Ward. The average relative change (ARC) is also reported, which is AME divided by the average outcome when a patient is not admitted to the ICU. The magnitude of the effect is substantial. For instance, admitting a patient to the ICU reduces the likelihood of hospital readmission by 32% on average.

The column “Test \( \rho = 0 \)” show the p-values of the test with the null hypothesis of exogeneity of the ICU admission
decision, which is equivalent to a likelihood ratio test against the model where the correlation coefficient $\rho$ between the admission and outcome models’ errors is restricted to be zero. The estimates of $\rho$ are reported in the column “$\rho$ (SE).” In all models, the null hypothesis of exogeneity of the ICU admission decision is strongly rejected. Hence, the results suggest that accounting for the endogeneity of the ICU admission decision is important to obtain consistent estimates of the effect of ICU care on patient outcomes.

We now assess the magnitude of the bias induced by neglecting the endogeneity of the admission decision in the estimation. The right panel of Table 5 (“Without IV”) shows the estimates ignoring the endogeneity of the admission decision, which are significantly different from those estimated with IVs (left panel). All cases exhibit positive biases on the coefficients when ignoring the admission decision endogeneity. This is consistent with the endogeneity problem discussed in Figure 4. ICU patients tend to be more severe, and because part of the patient severity is unobserved and therefore cannot be controlled for, the naive estimates (without IVs) tend to underestimate the benefit of ICU admission. In some cases the bias is so severe that it leads to a positive correlation between being admitted to an ICU and experiencing adverse outcomes.

While Table 5 reports the average marginal effect on outcomes from admitting a patient to the ICU, it is also useful to study how this effect varies across different patient severity classes. The Laboratory-based Acute Physiology Score (LAPS, described in Table 1) is used as a proxy for patient severity; its 30th and 70th percentile values are used to divide patients into three severity of illness groups: low, medium, and high. Table 6 shows the average benefit from ICU admission – the reduction in adverse outcomes when admitting the patient to the ICU relative to the general ward – by patient class. For instance, admitting a patient from the low-severity group to the ICU decreases his/her predicted probability of readmission from 0.077 to 0.050 on average (a reduction of 0.026). The table shows that, in general, the benefits are larger for higher patient severity classes; in the example, the average readmission probability decreases from 0.143 to 0.099 (a reduction of 0.043) for patients in the high-severity group.

In all of our estimates, we could not find a significant effect of ICU admission on mortality rates, which was at first surprising given the magnitude of the effect for other outcomes. A possible explanation of this relates to the IV estimation approach when the effects on the outcome are heterogeneous across patients. The estimation with valid IVs provides an unbiased effect of the average effect of ICU admission on patient outcomes over the subset of patients that are affected by the instrument. In our context, this includes patients whose ICU admission decision was affected by the ICU congestion one hour prior to their ED discharge. Figure 2 shows that this set includes patients from a broad class of severity – the ICU admission rate drops significantly when the ICU is congested and this is observed for patients with widely different severity classes. However, there is a good reason to believe that for patients with risk of dying, their ICU admission decisions are not influenced by the ICU congestion. In essence, there are many factors which come into play once a patient is assessed to have high risk of in-hospital death. First, if a patient is at high risk of death, and ICU care and monitoring could substantially reduce these risks, ICU congestion is unlikely to have much effect on the patient being admitted into the ICU. This would suggest that patients with very high severity always get into the ICU; however, we know this is not true. As mentioned in Reignier et al. (2008), it is often the case that patients with Do-Not-Resuscitate (DNR) orders are likely deemed “too sick for ICU treatment” and not admitted to the ICU.
Obviously, patients with DNR are also very likely to die, which is why the DNR was executed in the first place. Thus, of the patients who are very likely to die, other factors, such as DNRs, are more important in determining ICU care than the busyness of the ICU. For these patients, there is little compliance with the instrument. For the remaining patients, if they are likely to die if denied ICU care, then they will be sent to the ICU regardless of the state of the ICU. Because of how mortality risks and DNRs play into the ICU admission decision, it is not possible for us to estimate the effect of denied ICU admission on mortality.

4.4 Robustness Analysis and Alternative Model Specifications

Some of the controls of patient severity – LAPS and \( \hat{P}(\text{Mortality}) \) – are included with piece-wise linear functions to account for their possible non-linear effects on admission decisions and patient outcomes. We tried different specifications of these functions and the results were similar.

In the ICU admission model, we tested alternative measures to capture the level of occupancy in the ICU. As discussed in Section 3.2, our data analysis suggests that most of the adjustment to the ICU admission rate occurs when ICU occupancy goes above the 95\(^{th}\) percentile; hence, \( ICU_{\text{Busy}} \) was defined as a binary variable indicating occupancy levels above this threshold. This measure accounts for the differences in ICU sizes across the hospitals in the sample. In addition, we tested other specifications in which we interact several hospital characteristics with \( ICU_{\text{Busy}} \) to account for potential heterogeneous effects: these included measures of hospital size (dividing hospitals into groups by size), the presence of an intermediate care unit at the hospital, as well as with different shifts (7am-3pm, 3pm-11pm, and 11pm-7am). In all cases the estimated average effect of ICU occupancy on ICU admissions was similar to what was obtained in the main results.

In our model, we control for month of admission to capture potential seasonal effects and also hospital fixed effects to account for variations across hospitals. It is possible that there are time-varying hospital characteristics, which would not be controlled for with our month and hospital fixed effects. Thus, we also included hospital-month fixed effects and found that while these effects do seem to be statistically significant, accounting for them does not change our main results.

In defining \( Recent_{\text{Discharge}} \) and \( Recent_{\text{Admission}} \) in the ICU admission model, we use the 3-hr and 2-hr time windows, respectively. We experimented with shorter and longer time windows. For \( Recent_{\text{Discharge}} \), we observed that the effect persisted even when we consider a 8-hr time window (which we consider as the maximum duration since shifts change every eight hours). For \( Recent_{\text{Admission}} \), increasing the time window gave us weaker results, and the effect of this variable disappeared when we considered time windows longer than three hours. The estimates of the other model coefficients were robust to these alternative specifications.

We also examined other factors which may affect the admission decision, such as the severity of the patients currently in the ICU. Because our measures of severity are taken at the time of hospital admission (not ICU admission), this measure may not be very accurate, especially as we cannot account for how patient severity improves or deteriorates during their ICU stay. Nonetheless, when we control for the average severity of patients in the ICU, we find that 1) a patient is less likely to be admitted to the ICU when there are many severe patients and 2) the main results (e.g.
impact of a busy ICU on admission and the effect of admission on outcomes) of our estimations are robust to these alternate specifications.

We use the Full Maximum Likelihood Estimation (FMLE) to estimate our patient outcome models. While being more efficient, the FMLE imposes strong parametric assumptions on the distribution of outcomes. We did some validation of these assumptions for the count variable HospLOS. We observe over-dispersion – i.e., the unconditional variance is 24.0 while the mean value is 3.9 – and no evidence of zero-inflation (only 5.9% had hospital LOS equal to 0). Hence, the negative binomial model seems an appropriate model for this outcome.

For Readmit, recall that we have set the time window of two weeks after discussions with doctors. We have tested shorter and longer time windows and the results for two week time window were the strongest.

Recall that all the outcome models include the covariate AvgOccVisited, to control for the average occupancy level through the path of a patient. We considered other alternatives to measure the effect of this factor: (i) the daily average occupancy of all the inpatient units in the hospital during the patient’s hospital stay; (ii) the maximum occupancy level experienced by the patient in an inpatient unit during his hospital stay; (iii) the average number of inpatients in the hospital during the patient’s hospital stay over the maximum possible number of inpatients (without differentiating amongst different inpatient units); and (iv) the average occupancy level of inpatient units at the time the patient was discharged from the first inpatient unit he visited. All of these alternative definitions gave results that were consistent with those reported in our main specification.

When analyzing TransferUp, we included all patients in the estimation model as long as the patient had been to a non-ICU at least once. But patients who had in-hospital death may have a lower probability of a transfer-up event. Hence, we excluded patients with in-hospital death in TransferUp model and found that the results were similar. For the HospLOS model, recall that we measured it by the number of nights a patient stayed in the hospital after being discharged from the ED. We also tried defining HospLOS as LOS rounded to the nearest day, and the results were similar. We also estimated the outcome models excluding patients with in-hospital death (this was done with all outcome models except Mortality), and the results were again similar.

5 Accounting for Alternative Mechanisms that Control ICU Congestion

Although the results seem to be robust to alternative specifications, it is possible that the effect we attribute to ICU admission may be in part capturing the effect of other mechanisms used by the hospitals to manage ICU capacity. In this section, we consider two such alternative mechanisms. The first mechanism, which has been studied by Anderson et al. (2011) and Kc and Terwiesch (2012), is to shorten or “speed-up” the time a patient stays in the ICU to make room for new severe patients. Furthermore, Kc and Terwiesch (2012) show that this speed-up increases the probability of readmission of those patients, which is one of the patient outcomes we analyze in this study. Because this mechanism is more likely to be used when the ICU is busy, it is correlated to our main IV and can therefore confound our estimation of the effect of ICU admission on patient outcomes. The second mechanism is ED boarding: a congested ICU can extend the time a patient spends in the ED waiting to be transported to an inpatient unit. Typically EDs have limited
resources in terms of providing adequate patient care relative to an ICU or general ward\(^8\), and hence increasing ED boarding time for patients who have already been recommended for admission may negatively affect their quality of care. Because ED boarding is indeed correlated with ICU congestion, it may be confounding our estimation. This section analyzes these two alternative mechanisms – ICU speed-up and ED boarding – with the objective of identifying their effect separately from the causal effect of ICU admission on outcomes that we seek to estimate.

### 5.1 Speed up in the ICU

Kc and Terwiesch (2012) document the presence of an ICU speed-effect in a sample of patients undergoing cardiac surgery (which are admitted to a cardiac-specific ICU). Since our patient cohort is different from theirs (ours has patients admitted via the ED and did not go through surgery), it is important to validate this effect by replicating that methodology in our sample. The methodology is described briefly here but see Kc and Terwiesch (2012) for further details.

Define \(\text{firstICU }\text{LOS}_i\) as the ICU length of stay during patient \(i\)’s first ICU visit and \(\text{BUSY}_i\) as the bed utilization of the ICU at the time patient \(i\) was discharged from this ICU visit. Because our dataset does not have information on the number of scheduled arrivals, our definition of \(\text{BUSY}_i\) is not the same as in Kc and Terwiesch (2012). Instead, we let \(\text{BUSY}_i\) be 1 if the number of existing ICU patients at the time patient \(i\) is discharged from the ICU exceeds the 95\(^{th}\) percentile of occupancy.\(^9\) We estimate the association between ICU occupancy and length of stay through the following regression:

\[
\log(\text{firstICU }\text{LOS}_i) = \gamma \text{BUSY}_i + \beta X_i + u_i, \quad (5)
\]

where \(X_i\) is a vector of observable patient characteristics that describe the patient’s severity of illness. A negative \(\gamma\) suggests that high ICU congestion leads to a shorter ICU LOS – a speed-up effect. However, using the estimation results of this regression, we cannot reject that the null hypothesis that \(\gamma = 0\), with \(p\)-value of 0.47 (the coefficient value was -0.02 with standard error of 0.03).

To further validate the methodology used, we replicated the analysis over a different cohort of patients which is comparable to the one studied in Kc and Terwiesch (2012). We collected additional data from surgical patients that went to the ICU and ran regression (5) using this sample. Here, we do find a significant negative \(\gamma\) coefficient, -0.13 with standard error of 0.04. In other words, a congested ICU reduces the ICU length-of-stay by 5.5 hours, a 12\% reduction for the average patient. Our estimation effectively replicates the results of Kc and Terwiesch (2012) but at the same time we show that the speed-up effect is not present in the patient cohort (ED patients admitted to a medical service) in our study. Hence, ICU speed-up is unlikely to be confounding our main results on the effect of ICU admission on patient outcomes.

---

\(^8\)California requires 1:3 nurse-to-patient ratio for EDs, which is lower than that of ICUs but higher than that of general wards. Moreover, the primary purpose of an ED is to stabilize patients, rather than to provide supportive care as given in inpatient units.

\(^9\)We have tried various specifications for defining \(\text{BUSY}\), such as using different cutoff points for occupancy level and including future arrivals in a certain time window, and the results were consistent. In addition, we have tried hazard rate models–Weibull and Cox proportional hazard models–with \(\text{BUSY}\) measure included as both time-invariant and time-varying, and the results were consistent.
It is also interesting to see how the mechanisms to manage ICU capacity may vary across patient types. This was also reported in the work by Chen et al. (2013), showing that in contrast to non-cardiac patients, severity scores have little impact on the admission decision for cardiac patients.

5.2 ED Boarding Time

ED boarding – patients waiting in the ED to be admitted to an inpatient unit – tends to increase when the inpatient unit where the patient was admitted to is more congested. Hence, patients who are admitted to the ICU during high periods of ICU congestion may have waited a longer time in the ED. Since the ED has less adequate resources to take care of the patient, this additional waiting time in the ED may have direct implications on the patient outcome. This suggests that ICU congestion may influence patient outcomes through two different mechanisms: (i) the ICU admission decision, which is captured through model (2) and; (ii) the ED boarding time. Consequently, for ICU congestion to be a valid instrumental variable in isolating the effect of ICU admission on patient outcomes, we need to control for the effect of ED boarding time in the outcome model.

ED boarding time is defined as the time between the decision to admit the patient until the patient is discharged from the ED and physically moved to the inpatient unit. The decision on where to admit the patient (ICU versus general ward) is made somewhere in between, and might change during the ED boarding time. Figure 3 illustrates this time-line. We have data on the actual boarding time, but not the time at which the final decision of where to admit the patient is made. Also, we do not have any information on how much of the boarding time is due to waiting for a bed in the corresponding inpatient unit. If a patient’s ICU admission has been delayed (shown by long ED boarding time), the patient’s outcomes might be adversely affected by not receiving timely care. This effect can be captured by including total ED boarding time as an additional covariate in the outcome model.

The main challenge in this approach is that ED boarding time is also affected by unobservable patient characteristics and therefore is endogenous in the outcome model. Very severe patients will have shorter ED boarding time as they require urgent care in a specialized unit. Hence, in order to properly account for ED boarding, we need an exogenous instrumental variable that affects ED boarding time but is unrelated to the severity of the patient. The instrument we use is the “average level of bed occupancy of the inpatient unit the patient goes to after the ED”, labeled $\text{FirstInpatientOcc}$, where the average is taken during the time the patient is boarding in the ED. The logic is similar to our $\text{ICUBusy}$ instrument: if the patient was routed to an inpatient unit but this unit was busy when the patient was in the ED, then the patient probably had to stay a longer time in the ED waiting for a bed. A regression of the logarithm of ED boarding time on $\text{FirstInpatientOcc}$ shows a positive and highly significant effect. A 10% increase in the inpatient occupancy increases ED boarding time by 18%. Details of the results and controls included in the regression are shown in the online appendix.

Next, we estimate each of the outcome models developed in section 4.2 including $\log(\text{ED Boarding Time})$ as an additional covariate. This model measures the marginal effect of ED Boarding and ICU admissions, partialling out the effect of each variable separately; that is, it measures the effect of ICU admission above and beyond any effect caused by ED boarding. Since both covariates are endogenous, we use $\text{ICUBusy}$ and $\text{FirstInpatientOcc}$ as IVs
(note that the two instruments are not perfectly correlated because \textit{FirstInpatientOcc} measures the occupancy of the ward for the patients that are routed there). Since the outcome models are not linear, we use a control function approach to implement this IV estimation. The estimation is carried out in two steps: (1) we first estimate a linear regression with \textit{log(ED Boarding Time)} as the dependent variable and the IVs and controls as covariates; and (ii) we calculate the residuals of this regression and include the residuals and \textit{log(ED Boarding Time)} as additional covariates in the outcome model. See Wooldridge (2010) for more details on the control function approach.

Table 7 reports the estimated coefficients for ICU admission and \textit{log(ED Boarding Time)} for the different outcome models. The results for the effect of ICU admission are similar to those obtained in Table 5. The effect of ED boarding is not significant in most patient outcomes, except for \textit{TransferUp} where the coefficient is positive and significant. The main conclusion from this analysis is that the effect of ICU admission on patient outcomes prevails after we partial-out the effect of ED boarding time.

6 Managerial Insights and Performance Evaluation

One of the primary motivations for our study was to provide insight into how ICU admission decisions should be made. To this end, we developed a dynamic model and derived the optimal admission policy in Section 3.1. Such a policy requires measures of heterogeneous benefits of ICU admission for patients with varying severity, and so we estimated these primitives in Section 4. We now leverage these results to examine the performance of various admission policies. In particular, we are interested in studying whether admission criteria that are based on objective metrics of patient risk can outperform the current hospital admission policies.

Current admission policies use additional discretionary information from doctors in the admission decision (which cannot be accounted for in policies that use objective metrics alone), and it is not clear whether this practice is helpful. To address this issue, we (i) investigate the difference in patient outcomes when using the current policy (employed at the hospitals in our study) versus the optimal policies based on objective metrics of patient severity, (ii) study the robustness of admission policies across alternative performance criteria, such as readmission, transfer-up events and hospital length-of-stay, and (iii) analyze alternative admission criteria which adjusts the current admission decision to better account for the dynamics involved in the ICU admission decision while still exploiting the valuable information provided by doctors’ discretion. A key managerial insight from our study is to understand the circumstances under which allowing for discretionary criteria in the admission is useful versus making admission decisions on objective criteria alone.

6.1 Simulation Model

We start by describing the simulation model we use to evaluate the performance of the various admission policies. We consider a $B = 21$ bed ICU,\footnote{The average ICU capacity –defined as the 95\textsuperscript{th} percentile of the occupancy distribution– was 20 beds. There were three (out of 15) hospitals that had 20 beds as the capacity, and their average maximal capacity (i.e., the 100\textsuperscript{th} percentile) was about 21 beds.} so that if there are 21 patients in the ICU upon arrival, the new patient \textit{must} be rerouted.
The mean service time of each patient in the ICU is 60 hours, which is close to the empirical average for the patients in our sample. The arrival rate $\lambda$ is 3 patients per hour, which we have delicately chosen so that the simulated setting is consistent with the regime of the hospitals in our study that admit approximately 10\% of the inpatients to the ICU under the current policy. We simulate a full year, with one month of warmup, over 1,000 iterations.

When a patient arrives, we select $X_i$—the patient and seasonal characteristics from our estimation models in Sections 3.2 and 4.2—uniformly at random across all 70,133 patients in our sample. To calculate the benefit of ICU admission, we must specify an objective metric that characterizes a patient’s overall health outcome. In our context, there are three possible outcome measures as defined in Section 4—$HospLOS$, $TransferUp$, and $Readmit$ (we do not study mortality since the estimates for that outcome were imprecise and not statistically significant). Given $X_i$, we show in detail how to estimate the patient outcome for one of these measures, $Readmit$; the calculation for the other metrics is similar.

We define the outcome for patient $i$ according to the outcome model specified by Equation (4):

$$Readmit_i\{X_i, \epsilon_i, r_i\} = \begin{cases} 
1\{y_i^* > 0|X_i, \epsilon_i, r_i = 1\}, & \text{if admitted to the ICU;} \\
1\{y_i^* > 0|X_i, \epsilon_i, r_i = 0\}, & \text{if not admitted to the ICU.}
\end{cases}$$

Note that the benefit for patient $i$ is random and depends on the error term $\epsilon_i$ in Equation (4), which is assumed to be a standard normal random variable. Each patient’s outcome is determined by generating the error term $\epsilon_i$, and then comparing whether the aggregate risk of the patient (inclusive of the admission decision) is greater than 0. That is, assuming patient $i$ is not admitted to the ICU, if $X_i \beta_2 + \epsilon_i > 0$, then he will require a readmission. Otherwise, he will not be readmitted. Similarly, assuming the patient is admitted to the ICU, if $\beta_1 + X_i \beta_2 + \epsilon_i > 0$, then he will require a readmission and will not be readmitted otherwise. Because $\beta_1 < 0$, readmissions occur less frequently when admitted to the ICU. Note that in simulating the patient outcome, we use the hospital fixed effect for Hospital A (whose characteristics most closely correspond to our simulation setup), and the mean value of $AvgOccVisited$ seen in the data.

### 6.2 Admission Control Policies

We now describe each of the admission control policies we examine:

**Current Policy:** From our analysis in Section 3.2, we have empirically estimated the ICU admission model used in practice across the hospitals in question. As seen in Equation (2), the ICU admission decision depends on patient and seasonal characteristics, unobservable factors, and operational factors such as ICU occupancy, captured by $X_i$, $\xi_i$ and $Z_i$, respectively. While $\xi_i$ is not observed in our data, it is observable to the physicians who make the ICU admission decision. Thus, we assume $\xi_i$ is available to the current policy in our simulation. Additionally, note that $\xi_i$ is correlated with $\epsilon_i$, the unobservable factors in the outcome model, which is also accounted for in the simulation.

The current policy is simulated as follows for each patient arrival: we generate the unobservables $\epsilon_i$ and $\xi_i$ from a bivariate normal distribution with correlation values estimated from the data (reported in Table 5 for the various outcome models). If there is no available bed in the ICU, the patient is rerouted and the resulting patient outcome is
simulated as described in Equation (6). If there is space in the ICU, we determine whether the patient is admitted to the ICU according to the threshold policy in Equation (2) which depends on observed risk factors and $\xi_i$. Because the threshold for each hospital may vary due to the hospital fixed-effects, we specifically consider the admission policy used in Hospital A. Mimicking the current policy in Hospital A, we increase the admission threshold $\kappa_i$ by the $\alpha$-coefficient associated with $ICUBusy$ (estimated in Table 3 and described in Section 3.3) when the occupancy is equal to 20 beds, which corresponds to the 95th percentile occupancy at this hospital. For the other operational factors ($RecentDicharge$, $RecentAdmission$, and $LastAdmitSeverity$), we use the average values seen in the data.

**Optimal Policy:** We now consider the optimal policy derived in Section 3.1. Our analytic model requires patient risk, $p$, and cost function, $\Phi(p)$, as inputs. Again, we describe the method to derive model primitives for $Readmit$, noting the other outcome models will follow similarly. For each patient $i$, we can use the outcome model specified by Equation (4) to assess the risk of readmission when admitted versus rerouted and use this to derive $\Phi(p_i)$. Our goal in this work is to develop objective measures for ICU admission control, and thus, we consider the risk which can be measured via observable characteristics (captured by $X_i$) only and ignore the unobservable component $\xi_i$ in the admission policy.

For each patient $i$ with observable risk factors $X_i$, we calculate the cost of denied ICU admission as follows:

$$\Delta Readmit_i = E(Readmit_i|X_i, r_i = 0) - E(Readmit_i|X_i, r_i = 1)$$

(7)

where the expectation is calculated using Equation (4) and accounting for the fact that $\epsilon_i$ has a Standard Normal distribution. Note that we again use the hospital fixed effect for Hospital A and the mean value of $AvgOccVisited$ seen in the data.

In order to solve for the optimal policy, we must have a finite number of $p$ values. To do so, we divide the costs into ten equally sized groups where the cutoff points are the decile values and label them $p = 1, 2, \ldots, 10$. To estimate $\Phi(p)$ for each group, we assign the group’s cost as the average of $\Delta Readmit_i$ in each group.11 With $\Phi(p)$, we can use dynamic programming to calculate the optimal occupancy-dependent thresholds. When a new patient arrives and there is an available bed in the ICU, we examine the expected cost of denied ICU admission, as given by Equation (7), and admit the patient if his risk score is above the optimal occupancy-dependent threshold, and reroute otherwise. Again, the resulting patient outcome is simulated as described in Equation (6).

**Current Policy with adjusted admission threshold:** The current policy used in the hospital is based on the observable measures of risk $X_i$, as well as the unobservable measures $\xi_i$ and is characterized by the estimated threshold policy $X_i \theta + \xi_i \geq \kappa_i$. Because the term $\xi_i$ is correlated with the patient outcome, policies that ignore this discretionary assessment of risk – such as the optimal policy described above – could underperform relative to the current policy. An alternative to consider is to use the same risk assessment as in the current policy (accounting for $\xi_i$), but adjusting the admission threshold $\kappa_i$. The objective of this adjustment is to improve the admission decisions by making the policy better account for the trade-off of admitting patients now versus reserving a bed for a more severe incoming patient. This adjusted policy cannot do worse than the current admission policy, but could underperform relative to the

---

11That is, $\Phi(p) = \sum_{\xi\in\text{risk group}} \Delta Readmits_i / (#\text{patients in the risk group})$ for $p = 1, 2, \ldots, 10$. 

24
optimal admission policy described above. We use exhaustive search to find the optimal values of the two thresholds that correspond to occupancies below the 95th percentile value (ICU is not busy) and greater than or equal to the 95th percentile value (ICU is busy).

### 6.3 Results and Discussion

Table 8 summarizes simulated patient outcomes under the alternative policies we consider. The first result column labeled “BASE - 21 Beds” under “Current Policy” illustrates the performance of the current policy in a 21-bed ICU. Under the current policy (estimated from our econometric analysis), there were on average 2,641 hospital readmissions, 745 transfer-ups to the ICU from the general wards, and patients spent 87,545 hospital days over the course of a year. We note that our simulation results were well aligned with what we observe in the data (reported in Tables 2 and 4): in our simulations, approximately 10% of the patients were admitted to the ICU, 11% of the patients experienced readmissions, and 3% experienced transfer-up events. In the second result column labeled “22 Beds”, we also report the change in performance of the current policy when we increase the ICU capacity by one bed. Increasing the ICU bed capacity by one bed could be quite expensive; we roughly estimate this cost to be $0.8 million per year based on a $3,164 expense per ICU day (Aloe et al. 2009). In examining alternative admission policies, we will examine if some of the improvements in patient outcomes can be achieved without this high investment cost of increasing capacity.

The third result column labeled “Optimal Polices Minimizing Each Outcome” provides the performance of the optimal policy based on objective measures alone. If we use the optimal occupancy-dependent threshold derived from the cost function $\Phi(p)$ for \textit{Readmit}, we will observe 24.3 fewer readmissions on average compared to the current policy. This saving is more than five times the gain achieved by increasing the ICU capacity by one bed. On the other hand, if we use the optimal policy derived from the cost function $\Phi(p)$ for \textit{Transferup} or \textit{HospLOS}, we will see more transfer-ups (14 more transfer-ups on average) and hospital days (1,515.6 more days) compared to the current policy. That is, policies that ignore the discretionary assessment of the doctors perform worse than the current policy in terms of the number of transfer-ups or hospital LOS. Indeed, Table 5 shows high correlation, $\rho$, between the admission decision and the two patient outcomes (0.32 for \textit{Transferup} and 0.56 for \textit{HospLOS}), suggesting that doctors’ discretionary assessment in making the ICU admission decision captures a significant portion of unobserved factors that affect patient outcomes.

Having observed that doctors’ discretionary assessment could be useful, we now consider alternative policies that adjust the current hospital admission policy simply by changing the thresholds. Note that under the current policy, the admission threshold $\kappa_i$ increases by 0.48 when $\text{ICUBusy} = 1$ (see Table 3). We want to see whether moving $\kappa_i$ up or down, even when $\text{ICUBusy} = 0$, would give us better performance. That is, under these policies, we are still taking advantage of doctors’ discretionary assessment of patient risk, but are trying to optimize the admission thresholds given the same risk measure. Under the fourth (fifth) result column labeled “Current Policy with Adjusted Threshold”, we show the performance of such policies that minimizes the number of transfer-ups (hospital LOS). We see that the performance does improve by a non-negligible amount: we save 1.1 transfer-ups and 58.6 hospitals days on average when each policy is used. These values correspond to 17% (25%) of the benefit in terms of transfer-up
we get by adding one more bed. That is, slight adjustments to the current policy could capture much of the benefit of adding an ICU bed, without incurring the high costs of such a resource investment. We note that the value of local decision making that incorporates discretionary criteria has started receiving interest in other areas of Operations Management. For instance, Phillips et al. (2013) examine the use of pricing discretion of dealers in the auto lending industry, and have shown that dealers’ discretion help adjust prices in a way that improves profitability. To the best of our knowledge, our study is the first to address this issue in the healthcare operations literature.

Lastly, we discuss the robustness of admission policies across different patient outcome measures. When we use the optimal policy to minimize the number of readmissions, we decrease the number of readmissions at the expense of 61.8 more transfer-ups and 2,265.2 more hospital days. This suggests that the observed patient characteristics that determine readmission versus transfer-up and hospital LOS are very different, and highlights the inherently multi-objective nature of the ICU admission problem. One needs to be careful in setting the objective, as the cost of optimizing for the wrong patient outcome can be substantial. What is noticeable about the current policies with adjusted thresholds is that the policies minimizing transfer-up and hospital LOS (unlike that for readmissions) actually improve all patient outcomes. We note that this could further support the use of the current policy with adjusted threshold in certain settings.

7 Conclusion

We have examined the impact of ICU congestion on a patient’s care pathway and the subsequent effect on patient outcomes. We focused on medical patients who are admitted via the emergency department: a large patient cohort that comprises more than half of the patients admitted to the hospital. This is the first study to provide objective metrics that can be used by ED doctors and intensivists to decide which patients to admit to the ICU from the ED. We empirically found that the ICU congestion can have a significant impact on ICU admission decisions and patient outcomes and provided systematic and quantitative measures of the benefit of ICU care on various patient outcomes. Furthermore, we provide a detailed characterization of the optimal ICU admission policy based on objective measures of patient severity and show how to compute these policies for different patient outcome measures using empirical data, dynamic programming and simulation methods. Via simulation experiments, we were able to compare the performance of admission policies based purely on objective criteria (calculated from our empirical estimation) vis-à-vis the performance of the current admission policies used by the hospital network in our study. We showed that for certain outcome measures, using optimal policies based on objective metrics alone can outperform current hospital policies. For other outcome measures, we found that the discretionary criteria used by doctors is useful and helps to improve system performance relative to the decision based solely on objective criteria. We believe this is the first work to study the impact of doctors’ discretionary criteria on system performance in a healthcare setting.

From an estimation perspective, our instrumental variable approach can be extended to estimate the effect of other operational decisions. It is often the case that the effect of operational decisions on service outcomes is hard to estimate because of endogeneity bias. Our identification strategy of using operational and behavioral factors as
instrumental variables and carefully controlling for factors that would invalidate the instrument can be further utilized in related questions. We believe the present work can be easily applied to study capacity allocation and the impact of the occupancy level of available resources in many other healthcare settings. For instance, the differentiated levels of care can be among different ICU units. Rather than having only one type of ICU, many hospitals have specialized ICUs such as cardiac, surgical and medical ICUs, and the level of nurse-to-patient ratios and level of treatment might differ. However, they are sometimes shared when the occupancy levels are high in some of these units. Our model can be applied to estimate how the admission control to these different types of ICUs are done and whether it has an impact on patient outcomes.

We acknowledge that our study has several limitations, which in turn suggests future research directions. First, our dataset is limited in that all hospitals belong to one healthcare organization and that the majority of the patients are insured via this same organization. It would be interesting to look at other types of hospitals, which would enable us to explore features such as the difference between paying and non-paying patients. Second, in Section 3.1, we introduce a stylized model of ICU admission with constant arrival rate of inpatients and constant departure rate of ICU patients. We believe it serves its role of giving us insights on the impact of operational and medical factors on ICU admission control. However, when this model is again used in Section 6 to compare the performance of alternative admission policies, we might be able to get more accurate estimates if the model incorporated time-varying arrival rates, departure rates that depends on patient severity, and readmissions to the ICU and to the hospital. We note that incorporating these features adds new analytic challenges and that it is an active area of ongoing research (e.g., see Feldman et al. (2008) and Yom-Tov and Mandelbaum (2012)). Lastly, we hope to tease out and quantify the impact of the different adaptive mechanisms discussed in Sections 2.1 and 5–delays and boarding, speed-up, admission control, surgery cancellation and blocking via ambulance diversion– in terms of patient outcomes and hospital costs, depending on patient admission types and diagnosis. Building an analytic model that includes the complex interplay between different adaptive mechanisms on patient outcomes might prove useful in developing decision support tools for ICU admission, discharge, and capacity planning.

References


Boumendil, Ariane, Derek C Angus, Anne-Laure Guitonneau, Anne-Marie Menn, Christine Ginsburg, Khalil Takun, Alain Davido, Rafik Masmoudi, Benoit Doumenc, Dominique Pateron, et al. 2012. Variability of intensive care admission decisions for the very elderly. PloS ONE 7(4) e34387. 1


Cameron, A.C., P.K. Trivedi. 1986. Econometric models based on count data. comparisons and applications of some estimators and tests. Journal of Applied Econometrics 1(1) 29–53. 4.2

Cameron, A.C., P.K. Trivedi. 1998. Regression analysis of count data. Cambridge University Press. 4.2


Deb, P., P.K. Trivedi. 2006. Maximum simulated likelihood estimation of a negative binomial regression model with multinomial endogenous treatment. Stata Journal 6(2) 246–255. 4.2


Fisher, E. S., D. E. Wennberg, T. A. Stukel, D. J. Gottlieb. 2004. Variations In The Longitudinal Efficiency Of Academic Medical Centers. *Health Affairs* 1, 2.1

Franklin, C., E.C. Rackow, B. Mamdani, G. Burke, M.H. Weil. 1990. Triage considerations in medical intensive care. *Archives of internal medicine* 150(7) 1455. 2.1


Liu, V., P. Kipnis, M.K. Gould, G.J. Escobar. 2010. Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. *Medical care* 48(8) 739. 4.2


van Walraven, C., G.J. Escobar, J.D. Greene, A.J. Forster. 2010. The kaiser permanente inpatient risk adjustment methodology was valid in an external patient population. *Journal of Clinical Epidemiology* 63(7) 798–803. 2.1


8 Appendix: Figures and Tables

Figure 1: Data Selection

* indicates patient cohort used to determine the capacity and occupancy level of inpatient units. ** indicates the patient cohort used to estimate our econometric models.

Figure 2: Observed percentage of ICU admission for patients with different severity levels characterized by LAPS

Triangles show admission rates during periods of high ICU occupancy (95th percentile and above) and circles for other periods (below 95th percentile).
Figure 3: Time-Line for ED Patient Flow Process

Figure 4: Relationship between ICU admission decision and patient outcome - Endogeneity of admission decision
Table 1: Patient Characteristics and Seasonality Control Variables ($X_i$)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description and Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Patient ages less than 39 were coded 1, 40-64 coded 2, 65-74 coded 3 (Medicare starts at 65), 75-84 coded 4 and above 85 coded 5</td>
</tr>
<tr>
<td>Gender</td>
<td>Females were coded 1 and males 0</td>
</tr>
<tr>
<td>Severity score 1: LAPS</td>
<td>Laboratory-based Acute Physiology Score (Escobar et al. 2008); measures physiologic derangement at admission and is mapped from 14 laboratory test results, such as arterial pH and white blood cell count, obtained in the 24 hours preceding hospitalization to an integer value that can range from 0 to a theoretical maximum of 256 (the maximum LAPS value in our data set was 166); coded as piecewise linear spline variables with knots at 39, 69, 89</td>
</tr>
<tr>
<td>Severity score 2: $\hat{P}(\text{Mortality})$</td>
<td>an estimated probability of mortality (Escobar et al. 2008); predictors include LAPS and Comorbidity Point Score (measures the chronic illness burden and is based on 41 comorbidities); coded as piecewise linear spline variables with knots at 0.004, 0.075, 0.2</td>
</tr>
<tr>
<td>Admitting diagnosis</td>
<td>grouped into one of 44 broad diagnostic categories such as pneumonia; categorical variable to denote each diagnosis</td>
</tr>
<tr>
<td>Month/Time/Day</td>
<td>Month/Time/Day of week of ED admission; categorical variables</td>
</tr>
</tbody>
</table>

Table 2: Patient Characteristics by First Inpatient Units - Non-ICU versus ICU

<table>
<thead>
<tr>
<th></th>
<th>Non-ICU</th>
<th>ICU</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num. of obs.</td>
<td>63197</td>
<td>6936</td>
<td>70133</td>
</tr>
</tbody>
</table>

**Selected X Covariates**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Non-ICU</th>
<th>ICU</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>67.3 (17.8)</td>
<td>64.0 (18.0)</td>
<td>67.0 (17.8)</td>
</tr>
<tr>
<td>LAPS</td>
<td>23.5 (18.1)</td>
<td>36.1 (25.2)</td>
<td>24.7 (19.3)</td>
</tr>
<tr>
<td>$\hat{P}(\text{Mortality})$</td>
<td>0.044 (0.067)</td>
<td>0.095 (0.131)</td>
<td>0.049 (0.077)</td>
</tr>
<tr>
<td>Female</td>
<td>0.546</td>
<td>0.495</td>
<td>0.541</td>
</tr>
</tbody>
</table>

**Z Covariates**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Non-ICU</th>
<th>ICU</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICU Busy</td>
<td>0.096</td>
<td>0.039</td>
<td>0.091</td>
</tr>
<tr>
<td>Recent Discharge</td>
<td>0.033 (0.048)</td>
<td>0.040 (0.052)</td>
<td>0.034 (0.049)</td>
</tr>
<tr>
<td>Recent Admission</td>
<td>0.009 (0.022)</td>
<td>0.009 (0.021)</td>
<td>0.009 (0.022)</td>
</tr>
<tr>
<td>Last Admit Severity</td>
<td>0.341</td>
<td>0.311</td>
<td>0.338</td>
</tr>
</tbody>
</table>

*Note. Average and standard deviation (in parentheses for continuous variables) are reported.*
### Table 3: ICU Admission Model Estimation Result

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Estimate (Standard Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICU Busy</td>
<td>-0.48*** (0.03)</td>
</tr>
<tr>
<td>Recent Discharge</td>
<td>+1.39*** (0.15)</td>
</tr>
<tr>
<td>Recent Admission</td>
<td>-0.91** (0.34)</td>
</tr>
<tr>
<td>Last Admit Severity</td>
<td>-0.04* (0.02)</td>
</tr>
<tr>
<td>Hospital Indicators</td>
<td>Included</td>
</tr>
<tr>
<td>(\hat{P}(\text{Mortality}) [0,0.004])</td>
<td>28.68** (10.61)</td>
</tr>
<tr>
<td>(\hat{P}(\text{Mortality}) [0.004,0.075])</td>
<td>-1.28* (0.53)</td>
</tr>
<tr>
<td>(\hat{P}(\text{Mortality}) [0.75,0.2])</td>
<td>-0.96** (0.34)</td>
</tr>
<tr>
<td>(\hat{P}(\text{Mortality}) [0.2,])</td>
<td>-0.59** (0.22)</td>
</tr>
<tr>
<td>LAPS [0,39]</td>
<td>0.01*** (0.00)</td>
</tr>
<tr>
<td>LAPS [39,69]</td>
<td>0.02*** (0.00)</td>
</tr>
<tr>
<td>LAPS [69,89]</td>
<td>0.03*** (0.00)</td>
</tr>
<tr>
<td>LAPS [89,]</td>
<td>0.01* (0.01)</td>
</tr>
<tr>
<td>Pseudo (R^2)</td>
<td>0.181</td>
</tr>
</tbody>
</table>

*Note.* Full estimation results are in the online appendix. *\((p < 0.05)\), **\((p < 0.01)\), ***\((p < 0.001)\).*

### Table 4: Patient Outcome Variable Summary Statistics

<table>
<thead>
<tr>
<th>Outcome</th>
<th>n</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality</td>
<td>70,133</td>
<td>0.04</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TransferUp</td>
<td>68,200</td>
<td>0.03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Readmission - 2 weeks</td>
<td>67,087</td>
<td>0.10</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hospital LOS (days)</td>
<td>70,133</td>
<td>3.9</td>
<td>4.9</td>
<td>3.0</td>
</tr>
</tbody>
</table>

### Table 5: Effect of ICU Admission on Patient Outcomes and Average Estimated Marginal Effects

<table>
<thead>
<tr>
<th>Outcome</th>
<th>With IV</th>
<th>Without IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate (SE)</td>
<td>AME</td>
</tr>
<tr>
<td>Mortality</td>
<td>0.01 (0.13)</td>
<td>0.001</td>
</tr>
<tr>
<td>Readmit</td>
<td>-0.22* (0.13)</td>
<td>-0.034</td>
</tr>
<tr>
<td>TransferUp</td>
<td>-0.65*** (0.16)</td>
<td>-0.028</td>
</tr>
<tr>
<td>HospLOS (days)</td>
<td>-0.44*** (0.01)</td>
<td>-1.2</td>
</tr>
</tbody>
</table>

*Note.* Each row corresponds to a different outcome (the dependent variable); AME - Average Marginal Effect; ARC - Average Relative Change; Standard errors in parentheses. *\((p < 0.1)\), *\((p < 0.05)\), **\((p < 0.01)\), ***\((p < 0.001)\).*
### Table 6: Average Estimated Effect of First Inpatient Unit on Patients with Different Severity of Illness

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Low Patient Severity</th>
<th>Medium Patient Severity</th>
<th>High Patient Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Admit to Ward</td>
<td>Admit to ICU</td>
<td>Benefit</td>
</tr>
<tr>
<td>Readmit (pr)</td>
<td>0.077</td>
<td>0.050</td>
<td>-0.026</td>
</tr>
<tr>
<td>TransferUp (pr)</td>
<td>0.017</td>
<td>0.003</td>
<td>-0.014</td>
</tr>
<tr>
<td>HospLOS (days)</td>
<td>2.6</td>
<td>1.7</td>
<td>-0.8</td>
</tr>
</tbody>
</table>

*Note.* Patients are divided into three severity of illness groups (low, medium, and high) by the 30th and 70th percentile values of LAPS; pr - probability.

### Table 7: Estimation Results of the Patient Outcome Model Accounting for ED Boarding Time

<table>
<thead>
<tr>
<th>Outcome</th>
<th>ICU admission</th>
<th>Log(ED boarding time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Readmit</td>
<td>-0.21</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>TransferUp</td>
<td>-0.61***</td>
<td>0.16***</td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>HospLOS (days)</td>
<td>-0.40***</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
</tbody>
</table>

*Note.* Standard errors in parentheses. ∗ ∗ ∗ (p < 0.001).

### Table 8: Simulation Results of Alternative ICU Admission Control Policies

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Readmit</td>
<td>2640.8</td>
<td>-4.4</td>
<td>-24.3</td>
<td>-3.48 (79%)</td>
<td>-3.56 (81%)</td>
</tr>
<tr>
<td>Transferup</td>
<td>744.5</td>
<td>-6.7</td>
<td>14.0</td>
<td>-1.11 (17%)</td>
<td>-1.06 (16%)</td>
</tr>
<tr>
<td>HospLOS (days)</td>
<td>87545.2</td>
<td>-233.6</td>
<td>1515.6</td>
<td>-52.11 (22%)</td>
<td>-58.60 (25%)</td>
</tr>
</tbody>
</table>

*Note.* The performance measures of the current policy are denoted in bold; all other results are changes from the performance of the current policy. The percentage values in parentheses denote how much each alternative policy captures in relation to the benefit of adding one more bed.