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 [§]

April 23, 2014

Abstract

This work examines the admission process to hospitals' Intensive Care Units (ICUs), which currently lacks well-defined admission criteria. A major challenge that has impeded the progress of developing ICU admission standards is that the impact of ICU admission on patient outcomes has not been well quantified, making it difficult to evaluate the performance of candidate admission strategies. Using a large patient-level dataset of over 190,000 hospitalizations across 15 hospitals, we first quantify the cost of denied ICU admission for a number of patient outcomes. We make methodological contributions in this context, improving upon previously developed instrumental variable approaches. Using the estimates from our econometric analysis, we provide a framework to evaluate the performance of various admission strategies. By simulating a hospital with 21 ICU beds, we then show that we could save about 1.9 million dollars per year by using our optimal objective policy designed to reduce readmissions and hospital length-of-stay. We also discuss the role of physicians' discretion on the performance of alternative admission strategies.

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

[*]Columbia University, Email: sk3116@columbia.edu

[†]Columbia Business School

[‡]Dpto. Ingeneria Industrial, Universidad de Chile / Columbia Business School

[§]Division of Research, Kaiser Permanente

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 and Greenstein 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 examine what could be changed to improve the ICU admission decision 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 (Mullan 2004, Boumendil et al. 2012, Chen et al. 2012). 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 (Kaplan and Porter 2011, Chen et al. 2013), claiming that a primary reason for this gap is the general lack of objective criteria to characterize the benefit of different practices.

Our work takes an important step towards addressing this issue in that aims 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 exhibits 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. Our analysis shows that ICU care can reduce patient adverse health outcomes in the range of 30% to 75% (depending on the outcome). 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.

Equipped with these estimates, we compare the performance of various ICU admission strategies. We examine how the information available for decision making can impact performance. In particular, we examine the current admission criteria used by hospitals, which utilizes objective patient measures as well as doctors' discretionary information. 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. As such we compare the performance of the current policy to an 'optimal policy' which is based on objective metrics alone. 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 policy that uses 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 examine the benefit of the doctors' discretionary assessment of patient risk by examining an optimal policy which incorporates both objective and discretionary information. We find that in doing so, patient outcomes unilaterally improve, and we are able to capture the value of the doctors' discretionary information. Considering all the patient outcome measures that we study, a conservative estimate of the benefit of using the suggested policies at a single hospital translates into savings of patient bed hours on the order of 2.2 years, equivalent to US$1.9 million. This is approximately 5 times larger than the benefit that would be obtained by adding an additional bed to the ICU, excluding the costs of maintaining the extra bed.

In summary, we make the following key contributions:

- **Patient Outcomes:** In order to evaluate the performance of various admission policies, we require a quantification of the impact of ICU admission. Using a large patient-level dataset of over 190,000 hospitalizations across 15 hospitals, we quantify the cost of denied ICU admission for 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.

- **Evaluation and Comparison of ICU admissions:** Based on the estimates from our econometric analysis, we are able to calibrate a simulation model, which we use to compare the performance of various admission strategies. We compare the derived optimal admission policies with the current hospital admission policies

and find that in some circumstances it is useful to base admission decisions on objective metrics of patient risk alone where as in others the manner in which physicians incorporate objective and discretionary criteria in the admission decision can be beneficial. We are also able to quantify the benefit of discretionary information by examining how much patient outcomes improve when optimizing the admission decision based on both discretionary and objective criteria versus objective criteria alone.

The rest of the paper is organized as follows. We conclude this section with a brief literature review. Section 2 describes the context of the problem and the data used in this empirical study. Section 3 develops the econometric model to estimate the effect of admission decisions on various patient outcomes. Section 4 provides our estimation results. Section 5 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 6 summarizes our main contributions and provides guidelines for future research.

1.1 Literature Review

There have been a number of works in healthcare Operations Management (OM) that study the effect of workload and congestion on healthcare productivity. On the empirical side, Kc and Terwiesch (2009) show that hospital congestion can accelerate patient transportation time within the hospital; Kuntz et al. (2014) examine the impact of hospital load on in-hospital mortality using the ides of safety tipping points; Green et al. (2013) find that nurse absenteeism rates in an ED are correlated with anticipated future nurse workload levels; Ramdas et al. (2012) and Kc and Staats (2012) study the impact of surgeon experience on 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.

A more specific area of interest within this broader space is the study of mechanisms to manage ICU capacity. 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. 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.

Indeed, another alternative to manage ICU capacity that has been considered in the past and is considered here 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 the UK.

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.

Closest to our work is Shmueli et al. (2003) that examines 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 preselected 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.

We have seen that a number of mechanisms —including, but not limited to, ICU admission control— are used to manage ICU capacity in various settings. However, it is hard to find standards for when and 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 patient 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).

Our work takes an important step towards quantifying the 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 4.2.

We draw upon the rich literature on the *stochastic knapsack problem* when considering optimal policies utilize our estimated quantified benefits of ICU admission: see Miller (1969), Weber and Stidham Jr (1987), Veatch and Wein (1992), Glasserman and Yao (1994), Papastavrou et al. (1996) and references there in. Specifically, we use a special case of the stochastic knapsack problem studied in Altman et al. (2001), and leverage some results from that work to characterize our optimal policy.

As we evaluate alternative admission policies in Section 5, we also discuss the role of physicians' discretion on the performance of alternative admission strategies. The value of discretionary criteria (or experts' input) in decision making has started receiving interest in other areas of Operations Management; e.g., see Anand and Mendelson (1997), Phillips et al. (2013) and Osadchiy et al. (2013). To the best of our knowledge, our study is the first to address this issue in the healthcare operations literature.

2 Setting and Data

We employ a large patient dataset collected from 15 hospitals, comprising of nearly 190,000 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, though 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[1] and the other a predictor for in-hospital death [2]. 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.

2.1 Data Selection

We now describe the sample selection procedure for the data used in this study as depicted in Figure 1. 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 maximum ICU occupancy varied from 10 to 44. 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 1.1, 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

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

[2] These multiple severity of illness scores reflect the complexity in defining objective severity of illness measures. Table 1 explains patient characteristics in detail.

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.

2.2 Measuring Patient Outcomes

To quantify the benefit of ICU care, we focus on four types of patient outcomes whose summary statistics are provided in Table 3: (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 (Iezzoni et al. 2003, 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.[3] Note that patients who were admitted to and directly discharged from the ICU can never experience 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 define 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.

3 Measuring the Impact of ICU Admission on Patient Outcomes

In this section, we study how access to ICU care affects patient mortality, readmissions, transfer-up events and hospital length of stay. Section 3.1 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. Section 3.2 develops an

[3] 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).

estimation strategy using Instrumental Variables (IVs) to address this endogeneity problem and Section 3.3 describes our final estimation models.

3.1 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.

Our unit of observation is a hospital visit of a patient, indexed by i. Let y_i denote a measure capturing a patient outcome of interest during this visit (e.g., $HospLOS_i$). 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. Let X_i denote those patient severity factors as well as seasonality controls that are observed in the data. We also control for hospitals, where $\omega_{h(i)}$ denotes the coefficients for a set of hospital indicator variables and $h(i)$ is patient i's hospital. Our main hypothesis is that ICU treatment has a causal effect on patient outcomes. Accordingly, let $Admit_i = 1$ if patient i is admitted to the ICU and zero otherwise. We model patient outcome y_i as a random variable with distribution $f(y_i|\beta_1, \beta_2, Admit_i, X_i, \omega_{h(i)})$, where the parameter β_1 captures the affect of ICU admission and β_2 measures the effect of the observable characteristics X_i on the patient outcome, respectively. For example, this distribution could be given by a model of the form:

$$\log(y_i) = \beta_1 Admit_i + X_i\beta_2 + \omega_{h(i)} + \varepsilon_i, \tag{1}$$

with the error term ε_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 (1) is useful to illustrate the main estimation challenge. A naive approach to estimate the effect of ICU admission on y_i is to estimate the regression model (1) via Ordinary Least Square (OLS) and interpret the estimate of β_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 (e.g. the cognitive state of the patient) are likely to affect admission decisions. Figure 4 illustrates this endogeneity issue in further detail. The term ξ_i represents patient severity characteristics that are unobserved in the data but that are considered by the physicians when making the ICU admission decision. As such, both admission decisions and patient outcomes are affected by X_i and ξ_i. Since ξ_i is absorbed as part of the error term of model (1), the covariate $Admit_i$ is positively correlated with ε_i, therefore violating the strict exogeneity assumption required for consistent estimation through OLS. This endogeneity problem 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 β_1 to be negative).

An alternative is to use Instrumental Variables (IVs) estimation to obtain consistent estimates of this linear regression model. A valid instrument should be correlated with the admission decision $Admit_i$ but unrelated to the unobserved patient severity factors ε_i determining the outcome y_i. We propose using hospital operational factors that affect the ICU admission decision but are otherwise unrelated to patient severity. We describe and validate these IVs in the next section.

3.2 Instrumental Variables

A valid instrumental variable, denoted by Z, needs to satisfy the following two conditions: (1) it has to influence the endogenous variable, in our case, the ICU admission decision $Admit_i$; and (2) it has to be exogenous, that is, it cannot affect the patient outcome measure y_i other than through the admission decision. In this section, we discuss several potential instruments and empirically validates them here and in Section 4.

When deciding the ICU admission of an ED patient, the hospitals needs to evaluate the benefit of ICU treatment for this focal patient versus the opportunity cost of reserving the bed for a future, potentially more severe, incoming patient. This trade-off is particularly relevant when bed occupancy in the ICU is high – with only few beds left, admitting a patient now increases the probability that a future severe patient will be denied admission because the ICU is full. Because the number of beds is limited and the volume and severity of incoming patients is stochastic, the problem resembles an *admission control problem*: Altman et al. (2001) show that, for problems of this kind under various system conditions, the optimal admission control policy exhibits a reduction in the admission rate as the system occupancy increases.

We examine the data to identify differences in ICU admission rates due to occupancy. An ICU is labeled as "Busy" ($ICUBusy = 1$) if the bed occupancy is above the 95th percentile of its occupancy distribution, estimated by measuring the ICU bed occupancy every hour in the study period.[4] Figure 2 graphs the admission rates for 20 different patient groups (classified by their LAPS score on the horizontal axis) for two different occupancy levels: busy (marked with triangles) and not busy (marked with circles). The level of ICU occupancy associated to each patient is measured one hour prior to their ED discharge, which is a reasonable time period to cover the stage at which admission decisions are made. On top of the circles, we also show the percentage of the patients in each patient severity level group that saw an ICU that is "Not Busy". Across all groups, 90 to 92% patients are such patients, suggesting that there is no association between incoming patient severity level and the ICU occupancy level. 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 including the 90^{th}, 85^{th} and 75^{th} percentiles. The change in admission rate was much smaller and non-existent for some groups of patients. Although other measures of ICU occupancy could be considered, Figure 2 suggests that $ICUBusy_i$ is a statistically powerful

[4]For instance, an ICU with 10 beds is considered busy when 9 or 10 of its beds are occupied if its occupancy is 8 beds or below 94.5% of the time and 9 beds or below more than 95% of the time.

instrument, in the sense that it explains significant variation in the admission decision.

However, for $ICUBusy_i$ to be a valid instrument it also has to be uncorrelated with the unobservable factors ε_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 each patient 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).[5] We include $AvgOccVisited_i$ as an additional control variable in the outcome model (1) (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; the correlation between the two measures is 0.24 in our sample. Separating the effect of occupancy on 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 or a major accident affecting a large portion of the hospital's patient population. 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.

In addition to $ICUBusy$, we consider other instrumental variables that were suggested as potential factors affecting ICU admission decisions from our conversations with nurses, physicians and hospital management. We refer

[5] We define capacity of an inpatient unit as the 95^{th} 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

to them as the set of *behavioral factors*. The first factor, $RecentDischarge_i$, accounts for recent discharges from the ICU and is motivated by the following mechanism. 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[6], leading 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 factor, $RecentAdmission_i$, 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 factor, $LastAdmitSeverity_i$, 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 the ED physician just treated a very severe patient, he might require a new patient to be also very sick to recommend ICU admission. We define $LastAdmitSeverity_i$ as a dummy variable indicating whether the last patient admitted to the ICU had a LAPS score greater than or equal to the 66^{th} percentile value of the observed LAPS distribution. Table 2 provides summary statistics of the covariates for all the patients in our sample as well as patient grouped by whether they were admitted to the ICU or not.

The behavioral factors – $RecentDischarge_i$, $RecentAdmission_i$ and $LastAdmitSeverity_i$ – exhibit no correlation with the LAPS score of the incoming patient, suggesting that they are unrelated to patient severity and therefore appear to be exogenous. This is expected given the randomness in the arrival process of new incoming ED patients.

We define the vector of IVs, labeled Z, as these three behavioral factors plus $ICUBusy$. The next section describes how to implement the estimation using these IVs to instrument for the endogenous variable $Admit_i$.

[6]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).

3.3 Estimation

When the patient outcome is modeled via a linear regression as in (1), 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 develop non-linear parametric models to characterize the admission decision and the patient outcome and estimate these two models jointly via Full Maximum Likelihood Estimation (FMLE) (Wooldridge 2010). We describe this approach next.

The ICU admission decision is binary and is modeled through a Probit model defined by:

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

where X_i are observable patient characteristics, Z_i are the IVs and ξ_i is an error term following a Standard Normal distribution.

Patient outcomes are modeled using two different approaches depending on whether the 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:

$$\begin{aligned} y_i^* &= \beta_1 Admit_i + X_i\beta_2 + \omega_{h(i)} + \beta_3 AvgOccVisited_i + \varepsilon_i \\ y_i &= \mathbb{1}\{y_i^* > 0\}, \end{aligned} \tag{3}$$

where y_i^* is the latent variable. The additional control $AvgOccVisited_i$ 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 $Admit_i$, we allow for the error term ε_i to be correlated with the unobservable factors affecting admission (ξ_i in equation (2)) by assuming that the random vector (ξ_i, ε_i) follows a Standard Bivariate Normal distribution with correlation coefficient ρ (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 (3). 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 $Admit_i$ can be tested through a likelihood ratio test of the correlation coefficient ρ being different from zero.

The patient outcome defined by $HospLOS_i$ 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 $HospLOS_i$ reveals over-dispersion (Table 3 shows the mean of $HospLOS_i$ 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 $Admit_i$ – into the negative binomial regression, which is estimated jointly with model (2). The negative binomial regression includes the same covariates as in (3). The next section describes the estimation results of all the outcome models.

4 Estimation Results

In this section, we discuss the results of the patient outcome models, which are summarized in Table 4. As discussed in Section 3.3, we estimate the admission decision and patient outcome model jointly to account for the endogeneity of the admission decisions. We find that all of our instruments have an impact on whether a patient is admitted to the ICU. For example, we find that when the ICU is busy, the likelihood of being admitted to the ICU decreases by 53% on average (statistically significant at the 0.1%). For space limitations, Table 4 shows only the coefficient and the marginal effects of $Admit_i$ (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 4, the coefficients of $Admit_i$ 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 $\rho = 0$" shows 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 between the admission and outcome models' errors, ρ, is restricted to be zero. The estimates of ρ are reported in the column "ρ (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 4 ('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.

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

from all severity classes. However, anecdotal evidence from our conversations with physicians in this hospital network suggest that 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's admission to the ICU.[7] This suggests that ICU congestion plays no significant role in determining the admission decisions of patients with true risk of dying. Therefore, our estimation approach cannot be used to measure the benefit of ICU admission for this subset of patients as they do not comply with the instrumental variable.

4.1 Robustness Analysis and Alternative Model Specifications

This section describes analyses using alternative specifications that support the robustness of our main results. Some of the controls of patient severity–LAPS and $\hat{P}$(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, $ICUBusy_i$ 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 $ICUBusy_i$ 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 $RecentDischarge_i$ and $RecentAdmission_i$ 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 $RecentDischarge_i$, 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 $RecentAdmission_i$, 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.

Furthermore, we observe that the behavioral factors are less powerful IVs 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

[7] This gets more complicated by the patients who are denied ICU admission because they are deemed "too sick for ICU treatment" or have executed Do-Not-Resuscitate (DNR) orders; e.g., see Reignier et al. (2008).

an IV and the results were similar.

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 at the time of ICU admission or any time later in their hospital visit), 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 the 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–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 the two week time window demonstrated higher statistical significance and magnitude.

All the outcome models include the covariate $AvgOccVisited_i$ to control for the average occupancy level during each patient's stay in the hospital. 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 what we report for 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 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 for the $HospLOS$ model, and the results were again similar.

4.2 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. 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 with our main IV and can therefore confound our estimation of the effect of ICU admission on patient outcomes.

The speed-up effect analyzed in Kc and Terwiesch (2009) was based on cardiac surgery patients, whereas our study is based on ED patients, a completely different patient population. Therefore, we replicate their methodology in our patient sample to measure the magnitude of the effect. The details of this estimation is described in Appendix A.1. The results of this analysis cannot reject the null hypothesis of no speed-up effect in our patient population (p-value 0.47).

To further validate the replication of this methodology, we estimated the same model using a sample of patients comparable to the one studied in Kc and Terwiesch (2012): we utilized our data that also include elective surgical patients admitted to the ICU. These results strongly reject the null hypothesis of no speed-up effect (p-value 0.001): a congested ICU reduces length of stay in the ICU by 12%. Therefore, our method correctly replicates the results of Kc and Terwiesch (2012), but at the same time shows no speed-up effect in the patients admitted to the ICU via the ED. We conclude that this mechanism is not relevant in our patient population and therefore cannot be confounding our main results regarding the effect of ICU admission on patient outcomes.

It is also interesting to see how the mechanisms to manage ICU capacity may vary across patient types (ED versus surgical patients). This was also reported in Chen et al. (2013), showing that in contrast to non-cardiac patients, severity scores have little impact on the admission decision for cardiac patients.

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. 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.[8] 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.

To account for this mechanism, we include a measure of ED boarding time as a covariate in the outcome models (equations (1) and (3)). 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, which is measured in the data. If a patient's ICU admission has been delayed (shown by long ED boarding time), the patient's outcomes might

[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.

be adversely affected by not receiving timely care. Therefore, the effect of ED boarding time should be negative. However, ED boarding time is endogenous and can be affected by unobservable patient characteristics related to the patient's outcome. A severe patient that requires urgent care is likely to have a shorter boarding time. Appendix A.2 describes in detail the econometric model developed to handle this endogeneity problem using instrumental variables. This econometric model identifies the effect of ED boarding and ICU admissions on patient outcomes, 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.

The results of this estimation (reported in Appendix A.2) show that, for some outcomes, a longer ED boarding time leads to worse patient outcomes; for other outcomes the effect is not significant. More importantly, the estimated effects of ICU admission are similar to those reported in Section 4. The main conclusion of this analysis is that our main results regarding the effect ICU admission on patient outcomes are not confounded by the effect of ED boarding time.

5 Evaluating Alternative Admission Policies

A primary objective in our study is to quantify the benefits of ICU care. This is an essential first step in comparing different ICU admission strategies. To examine how we can utilize the measures we have just estimated, we consider a parsimonious model of patient flows into the ICU to examine the performance of various admission policies. We leverage our estimation results to calibrate a simulation model, which allows us to compare patient outcomes across different 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.

5.1 Model of Admission Control

We model the ICU admission control problem as a discrete version of the Erlang loss model, similar to the one used in Shmueli et al. (2003). This admission control problem can be viewed as a special case of the stochastic knapsack problem studied in Altman et al. (2001), and we leverage some results from that work to characterize its solution.

Consider an ICU with B beds. In order to focus on the ICU admission decision, we assume there is ample space in the other inpatient units to care for all patients. We denote by x the number of occupied ICU beds at any given point in time. When $x = B$, arriving patients must be routed to the general Ward. Time is discretized into periods of fixed length dt, indexed by t, where the periods are sufficiently short so that it is reasonable to assume at most one patient arrives in a given period. In each period, a patient arrives to the ICU with probability λ. Upon arrival, a decision must be made on whether to admit the patient or not. If admitted to the ICU, a patient's length of stay is geometrically distributed with mean $1/\mu$. We assume that patient discharge is exogenous, i.e. there is no speed-up in the ICU.[9]

If a patient is routed to the Ward an expected cost of ϕ_c is incurred, where c indexes the customer's class. Without

[9] As discussed in Section 1.1, other mechanisms may be used; although we do not find that speedup is used for the patient group we study; see Appendix A.1. Via numerical analysis, we found that the qualitative results extend when speed-ups are incorporated.

loss of generality, classes are numbered $1 \dots C$ so that ϕ_c increases in c; therefore, classes can be interpreted as the severity of the patient, where the benefit of admitting a patient increases with his severity. The objective is to choose an admission criteria that minimizes the total expected cost over a finite-horizon.

An *admission policy* is defined as a decision rule that chooses whether to admit or reroute an incoming patient, for each possible state characterized by the class of the incoming patient (c) and the number of occupied ICU beds $x \in [0, B]$. Altman et al. (2001) shows that the optimal admission policy is a threshold policy with the following structure. Given a occupancy level x, admit a patient if and only if his class satisfies $\phi_c \geq \kappa_x$. The values $\{\kappa_1 \dots \kappa_B\}$ are referred to as the *optimal thresholds*. It is also shown that the thresholds κ_x are increasing in x.

Next, we describe how we set the primitives of this admission control problem in order to run a simulation.

5.2 Model Calibration and Simulation

The simulation analysis focuses in an ICU with $B = 21$ beds, which is the median ICU size in our data. To simulate ICU admissions, we sample (with replacement) patient characteristics from a hospital whose 95^{th} percentile of occupancy distribution was at 20 beds and 99^{th} percentile at 21 beds. This hospital treated 7,387 ED-medical patients during our study period. Each discrete time period lasts 10 minutes and patients arrives to the ICU with probability λ so that on average 3 patients arrive per hour. This has been 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. The average patient LOS in the ICU is $1/\mu = 60$ hours, which corresponds to the average duration of ICU stay in our sample.

Next, we describe how to estimate the expected rerouting costs ϕ_c for each patient class c. This requires defining the health outcome measure to be considered–$HospLOS$, $TransferUp$, and $Readmit$ (we do not study mortality since the estimates for that outcome were imprecise and not statistically significant). Let y be the outcome of interest. Recall that ϕ_c represents the difference in this expected health outcome if a patient is admitted to the ICU versus not admitted.

Information about the incoming patient is essential to assess his severity class. Each patient i is fully described by a set of objective characteristics X_i (recorded in our data and described in Table 2) and the "error term" ξ_i capturing other patient characteristics, not observed in the data, that are taken into account by the physician when assessing the patient admission. We therefore call X_i and ξ_i the objective and discretionary component of the patient information, respectively. Defining an admission policy requires specifying what kind of information is considered when making a decision, which we define as the information set I_i. We focus on studying policies that use all the information, $I_i = (X_i, \xi_i)$, and policies that use only the objective component, $I_i = X_i$.

For a given patient with information set I_i, the expected rerouting cost is calculated as:

$$\phi_i = E(y_i | Admit_i = 0, I_i) - E(y_i | Admit_i = 1, I_i), \tag{4}$$

where the expectation is taken with respect to ε_i, the error term in the corresponding outcome model. We explain in detail how we estimate this cost for $Readmit$ with information set $I_i = X_i$; the calculation for the other metrics is

similar. For readmissions, Equation (4) becomes:

$$\phi_i^{Readmit} = \Pr(\varepsilon_i \geq -\beta_2 X_i) - \Pr(\varepsilon_i \geq -\beta_2 X_i - \beta_1),$$

which is positive when $\beta_1 < 0$. When only the objective information component is observed, ε_i follows a Standard Normal distribution. When the discretionary component ξ_i is also included in the information set (i.e., $I_i = (X_i, \xi_i)$), ε_i follows a Normal distribution with mean $\rho\xi_i$ and variance $(1-\rho)^2$. The parameters β_1, β_2, ρ are the estimates of the Readmission outcome reported in Section 4 and therefore the probabilities can be calculated numerically.

Equation (4) calculates the rerouting cost for a specific patient. In practice, deriving the optimal admission policy via dynamic programming requires a finite set of patient classes. To achieve this for each health outcome, we first calculate ϕ_i for all 7,387 patients treated in the hospital that we chose to simulate. (For the case where the discretionary component ξ_i is included in the information set and, hence, the value of (4) depends on ξ_i, we generate 1,000 realizations of ξ_i and compute 7,387,000 values of ϕ_i. Then we partition patients into 10 groups based on the deciles of this distribution; each patient class has lower and upper bound on ϕ_i which defines patients that belong to the class group. Class c's rerouting cost ϕ_c is set as the average rerouting costs of the patients in that class.

A *policy* is specified by a function that maps patient information set I_i and the number of occupied beds (x) to an admission decision. The following procedure describes how we carry out our discrete-time simulation of a given policy. At $t = 0$, occupancy is set to zero. In every period, with probability λ, a patient is sampled from the population of patients, characterized by X_i and a random vector (ξ_i, ε_i) from a Bivariate Standard Normal with correlation coefficient ρ. A patient is admitted to the ICU if $x < B$ and the policy evaluates to do so. This will result in an increase in ICU occupancy to $x + 1$. Otherwise, the patient is not admitted. At the end of the period, each ICU patient leaves with probability μ. We simulate a full year, with one month of warmup after which the system status reaches stationarity, over 1,000 iterations.

5.3 Admission Control Policies

We use the simulation model described above to examine how different ICU admission strategies impact aggregate patient outcomes. We compare 4 different policies. The **Estimated Current Policy** corresponds to an empirical model of the admission policy used at the hospitals in our study, which we estimate from the data. The **Optimal Objective Policy** uses the objective component of patient information (i.e. $I_i = X_i$) to assess the expected rerouting cost, and derive the optimal threshold levels of admission. The **Optimal Full Policy** uses the objective and discretionary components ($I_i = (X_i, \xi_i)$) in assessing the expected rerouting cost. The fourth policy is similar to the Estimated Current Policy, but with $B = 22$ as bed capacity. We now describe each of these policies in more detail.

Estimated Current Policy: The structural results of Altman et al. (2001) establish that the optimal policy is of threshold form. Although the policy currently used by the hospital need not be optimal, Figure 2 presents several patterns that are consistent with a threshold policy. First, admission rates tend to increase as patient severity increases. Second, admission rates decrease at higher levels of occupancy, consistent with threshold levels that increase with the number of occupied beds. Third, the drop in admission rate due to an increase in occupancy is higher for more severe

patients, which can be shown to be in line with a threshold policy.[10]

We restrict the hospital we choose to simulate to follow a threshold policy that uses an information set $I_i = (X_i, \xi_i)$ and develop an empirical model to estimate the parameters of this policy. The model is given by:

$$Admit_i(I_i, x) = 1\{X_i\theta + \xi_i \geq f(x;\kappa)\}, \tag{5}$$

where $f(x;\kappa)$ is a function that parameterizes the thresholds as a function of the occupancy x. Assuming $\xi_i \sim N(0,1)$, the model can be estimated via a Probit model. We experiment (and hence fit the Probit model) with all possible combinations of the way the occupancy x can affect the admission policy; that is, we vary the number of thresholds and the locations of the thresholds that the occupancy x can have. For instances, $f(x;\kappa)$ can change at every possible occupancy level or it can change only once, say when the ICU occupancy is 20 and above. For each model (5) with different combination for $f(x;\kappa)$, we compute the Bayesian Information Criterion (BIC), which is a commonly used metric to select the most parsimonious model that best describes data; it is computed based on the likelihood and has a penalty term for the number of parameters in the model; see Raftery (1995). We then choose the model that has the smallest BIC value to be our estimated current policy.

Optimal Policies: Since the optimal policy is of threshold form, a patient i is admitted if:

$$Admit_i(I_i, x) = 1\{\phi_i > \kappa_x\},$$

where ϕ_i is calculated by equation (4). We use dynamic programming to determine the threshold values $\{\kappa_x\}_{x=0}^{B}$ that minimize total costs. Notice that the calculation of ϕ_i depends on the information set I_i, therefore the optimal policy depends on I_i, which leads to the **Optimal Objective Policy** ($I_i = X_i$) and the **Optimal Full Policy** ($I_i = (X_i, \xi_i)$). To facilitate the dynamic programming recursion, we assign patient i the rerouting cost of his class ϕ_c, which reduces the possible values of each threshold to $\{\phi_1 \dots \phi_{10}\}$. This provides a lower bound on the performance of the optimal policies.

The Estimated Current Policy may perform worse than the Optimal Objective Policy. This could occur for several reasons. First, the admission decision under the optimal objective policy is based on the rerouting cost ϕ_i, while, in the current admission decision described by equation (5), the left hand side of the inequality is not necessarily equal to ϕ_i. More specifically, this implies that the estimated current policy may not be appropriately weighting the objective metrics X_i. Second, the threshold adjustment function $f(x;\kappa)$ may not set the optimal threshold levels that properly accounts for the opportunity cost of using up a bed, which the optimized policy can. On the other hand, the Estimated Current Policy has a richer information set than the Optimal Objective Policy, so it is not a priori known which will perform better. Because the Optimal Full Policy utilizes the same information as the Estimated Current Policy, accurately weights both the objective and discretionary information ($I_i = (X_i, \xi_i)$), and optimizes the thresholds, the Estimated Current Policy will have worse performance.

[10] Consider two patient classes, high (H) and low (L) severity, and assume that patient severity for class $j \in \{L, H\}$ follows a Normal(μ_j, σ^2), where $\mu_L < \mu_H$. Given a threshold κ, the admission probability for patient class j is given by $\Pr(N(\mu_j, \sigma^2) > \kappa)$; assume $\mu_L < \mu_L < \kappa$ (less than 50% of patients in all classes are admitted). An increase in occupancy raises the threshold to $\kappa + \Delta$, which decreases the admission rates of all groups, but the H group decreases by more. These results are not specific to the Normal distribution assumption on severity – they hold for any distribution with density function decreasing at the threshold κ (i.e. $f'(x) < 0$ for $x > \kappa$).

5.4 Results and Discussion

Table 5 summarizes the simulated patient outcomes–$HospLOS$, $TransferUp$, and $Readmit$–under the alternative policies we consider. Noting that the ICU admission decision is an inherently multi-objective problem, we also consider a combined outcome which considers the impact of ICU admission on total hospital days in the current inpatient stay as well as any potential subsequent hospital stay due to readmission. In particular, we convert each readmission into the average stay of 3.9 hospital days (see Table 3) and add this to $HospLOS$; we note that this is a conservative measure as readmitted patients are likely to stay longer in the hospital[11]. Finally, for comparison purposes, we convert hospital days into dollar amounts utilizing an estimate of $2,419 per hospital day as given in The Kaiser Family Foundation, statehealthfacts.org (2012).

The column labeled "BASE-21 Beds" under "Estimated Current Policy" lists the performance of the estimated current policy in a 21-bed ICU. Under the current policy (estimated as described above), there were on average 2550 hospital readmissions, 762.9 transfer-ups to the ICU from the general wards, and patients spent a total of 245.6 hospital bed years 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 3): 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 column, labeled "22 Beds", we also report the change in performance of the estimated 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 column, labeled "Optimal Objective", provides the performance of the optimal policy based on objective measures alone, so that the information set $I_i = X_i$. Each row corresponds to a different policy optimized to minimize the corresponding outcome. Because the optimal objective policy optimizes the admission thresholds while also utilizing the direct relationship between the available information (X_i) and patient outcomes, it can sometimes do better than the estimated current policy. This is the case when we use the optimal occupancy-dependent threshold derived from the cost function for readmissions ($\phi_c^{Readmit}$) and hospital LOS ($\phi_c^{HospLOS}$); we observe 26.6 fewer readmissions and 2 fewer years of hospital LOS on average compared to the current policy. On the other hand, the estimated current policy may outperform the optimal objective policy since it utilizes more information ($I_i = (X_i, \xi_i)$) and as we saw in Table 4, ξ_i captures a significant portion of unobserved factors that affect patient outcomes (as indicated by the correlation coefficient ρ). Indeed, the optimal objective policy aimed at minimizing $Transferup$ has *more* transfer-ups (15 more on average) compared to the estimated current policy. That said, this was not a systematic effect. We examined other hospitals and found that the optimal objective policy can outperform the current policy across all patient outcomes. These results suggest that the discretionary information can be useful, but optimizing the admission decision based solely on objective criteria can often result in better patient outcomes.

[11] We do not include (convert) $TransferUp$ into total hospital days because, while patients who are transferred up tend to have longer LOS, this is captured in the effect of $HospLOS$. To avoid double-counting, we only combine $Readmit$ and $HospLOS$

Finally, we further explore the benefits of incorporating the doctors' discretionary information in the admission decision. The last column, labeled "Optimal Full", uses the objective and discretionary information ($I_i = (X_i, \xi_i)$) available to the physicians, and further optimizes the admission thresholds. We see that by optimizing the thresholds, patient outcomes can be universally improved compared to the estimated current policy. The difference between "Optimal Full" and "Optimal Objective" captures the value of the physicians' discretion with respect to how much patient outcomes can be improved when also incorporating ξ_i in the information set. We can see the benefit of the physicians' assessment can be quite substantial, resulting in 8.4 fewer readmissions, 52.7 fewer transfer-ups, and 6.8 fewer patient years spent in the hospital. Moreover, these gains are orders of magnitude greater than what we achieve by adding an additional ICU bed, without incurring the costs of finding space and paying for such a structural change.

6 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 each hospital 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 can help 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 5.1, 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. Possible extensions of this simulation model could incorporate 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 (2014)). Third, there is a limitation of our empirical strategy due to the fact that an estimation based on instrumental variables provides an estimate of the average effect of the endogenous variable on the population that is affected by the instrument. In our context, our approach measures the average effect of ICU treatment on health outcomes for those patients whose ICU admission decision depends on the congestion of the ICU. This excludes two sets of patients: (1) Patients that are never admitted to the ICU, even if there is ample space in the ICU. This set of patients is probably the ones that benefit the least from ICU treatment. (2) Patients that are severe enough to be admitted to the ICU no matter how busy it is. These are usually the most severe patients and includes those patients with higher risk of dying. Hence, a limitation of our IV strategy is that we cannot estimate the benefit of ICU admission for all patients, but only for those for which the hospital's admission decision was affected by the level of congestion of the ICU, which probably excludes the most severe and the more healthy patients. Estimating the effect for these extreme cases would probably require a randomized experiment, which would be ethically questionable especially for the high severity group. Lastly, we hope to tease out and quantify the impact of the different adaptive mechanisms discussed in Sections 1.1 and 4.2–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

Allon, G., S. Deo, W. Lin. 2013. The impact of hospital size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research* **61**(3) 544–562.

Aloe, K., M. Ryan, L. Raffaniello, L. Williams. 2009. Creation of an intermediate respiratory care unit to decrease intensive care utilization. *Journal of Nursing Administration* **39**(11) 494–498.

Altman, E., T. Jiménez, G. Koole. 2001. On optimal call admission control in resource-sharing system. *Communications, IEEE*

Transactions on **49**(9) 1659–1668.

Anand, K., H. Mendelson. 1997. Information and organization for horizontal multimarket coordination. *Management Science* **43**(12) 1609–1627.

Anderson, D., C. Price, B. Golden, W. Jank, E. Wasil. 2011. Examining the discharge practices of surgeons at a large medical center. *Health Care Management Science* **14**(4) 1–10.

Armony, M., S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, G. Yom-Tov. 2011. Patient flow in hospitals: a data-based queueing-science perspective. *Working paper, New York University*.

Azoulay, E., F. Pochard, S. Chevret, C. Vinsonneau, M. Garrouste, Y. Cohen, M. Thuong, C. Paugam, C. Apperre, B. De Cagny, et al. 2001. Compliance with triage to intensive care recommendations. *Critical Care Medicine* **29**(11) 2132–2136.

Baker, L., S. Atlas, C. Afendulis. 2008. Expanded use of imaging technology and the challenge of measuring value. *Health Affairs* **27**(6) 1467–1478.

Batt, R., C. Terwiesch. 2012. Doctors under load: An empirical study of state-dependent service times in emergency care. *Working Paper, The Wharton School*.

Boumendil, A., D. Angus, A. Guitonneau, A. Menn, C. Ginsburg, K. Takun, A. Davido, R. Masmoudi, B. Doumenc, D. Pateron, et al. 2012. Variability of intensive care admission decisions for the very elderly. *PloS ONE* **7**(4) e34387.

Brilli, R., A. Spevetz, R. Branson, G. Campbell, H. Cohen, J. Dasta, M. Harvey, M. Kelley, K. Kelly, M. Rudis, et al. 2001. Critical care delivery in the intensive care unit: defining clinical roles and the best practice model. *Critical Care Medicine* **29**(10) 2007–2019.

Cady, N., M. Mattes, S. Burton. 1995. Reducing intensive care unit length of stay: A stepdown unit for first-day heart surgery patients. *Journal of Nursing Administration* **25**(12) 29–30.

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

Cameron, A., P. Trivedi. 1998. *Regression analysis of count data*. Cambridge University Press.

Chalfin, D., S. Trzeciak, A. Likourezos, B. Baumann, R. Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* **35**(6) 1477–1483.

Chen, L., Edward H. Kennedy, A. Sales, T. Hofer. 2013. Use of health IT for higher-value critical care. *New England Journal of Medicine* **368** 594–597.

Chen, M., M. Render, A. Sales, E. Kennedy, W. Wiitala, T. Hofer. 2012. Intensive care unit admitting patterns in the veterans affairs health care system. *Archives of Internal Medicine* **172**(16) 1220–1226.

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

Durbin Jr, C., R. Kopel. 1993. A case-control study of patients readmitted to the intensive care unit. *Critical Care Medicine* **21**(10) 1547–1553.

Escher, M., T. Perneger, J. Chevrolet. 2004. National questionnaire survey on what influences doctors' decisions about admission to intensive care. *BMJ* **329**(7463) 425–429.

Escobar, G., J. Greene, M. Gardner, G. Marelich, B. Quick, P. Kipnis. 2011. Intra-hospital transfers to a higher level of care:

Contribution to total hospital and intensive care unit (icu) mortality and length of stay (los). *Journal of Hospital Medicine* **6**(2) 74–80.

Escobar, G., J. Greene, P. Scheirer, M. Gardner, D. Draper, P. Kipnis. 2008. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Medical Care* **46**(3) 232–239.

Feldman, Z., A. Mandelbaum, W. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Science* **54**(2) 324–338.

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

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

Glasserman, P., D. Yao. 1994. Monotone optimal control of permutable gsmps. *Mathematics of Operations Research* **19**(2) 449–476.

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

Green, L., S. Savin, N. Savva. 2013. Nursevendor problem: Personnel staffing in the presence of endogenous absenteeism. *Management Science, published online.* doi:10.1287/mnsc.2013.1713.

Halpern, L., N.and Bettes, R. Greenstein. 1994. Federal and nationwide intensive care units and healthcare costs: 1986–1992. *Critical Care Medicine* **22** 2001–2007.

Halpern, N., S. Pastores, H. Thaler, R. Greenstein. 2007. Critical care medicine use and cost among medicare beneficiaries 1995-2000: Major discrepancies between two united states federal medicare databases*. *Critical Care Medicine* **35**(3) 692–699.

Iapichino, G., D. Corbella, C. Minelli, G.H. Mills, A. Artigas, D.L. Edbooke, A. Pezzi, J. Kesecioglu, N. Patroniti, M. Baras, et al. 2010. Reasons for refusal of admission to intensive care and impact on mortality. *Intensive Care Medicine* **36**(10) 1772–1779.

Iezzoni, L., et al. 2003. *Risk adjustment for measuring health care outcomes*, vol. 3. Health Administration Press, Ann Arbor.

Jaeker, J.B., A.L. Tucker. 2013. An empirical study of the spillover effects of workload on patient length of stay. *Working Paper, Harvard Business School*.

Joint Position Statement. 1994. Essential provisions for critical care in health system reform. *Critical Care Medicine* **22** 2017–2019.

Kaplan, R., M. Porter. 2011. How to solve the cost crisis in health care. *Harvard Business Review* **89**(9) 10.

Kc, D., B. Staats. 2012. Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. *Manufacturing and Service Operations Management* **14**(4) 618–633.

Kc, D., C. Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55**(9) 1486–1498.

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

Kuntz, L., R. Mennicken, S. Scholtes. 2014. Stress on the ward: Evidence of safety tipping points in hospitals. *Forthcoming in Management Science*.

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

Louriz, M., K. Abidi, M. Akkaoui, N. Madani, K. Chater, J. Belayachi, T. Dendane, A.A. Zeggwagh, R. Abouqal. 2012. Determinants and outcomes associated with decisions to deny or to delay intensive care unit admission in morocco. *Intensive Care Medicine* 1–8.

Luyt, C., A. Combes, P. Aegerter, B. Guidet, J. Trouillet, C. Gibert, J. Chastre. 2007. Mortality among patients admitted to intensive care units during weekday day shifts compared with off hours. *Critical Care Medicine* **35**(1) 3–11.

Miller, B. 1969. A queueing reward system with several customer classes. *Management Science* **16**(3) 234–245.

Mullan, F. 2004. Wrestling With Variation: An Interview With Jack Wennberg. *Health Affairs* **Suppl Variation** VAR73–80.

O'Connor, A., H. Llewellyn-Thomas, A. Flood. 2004. Modifying Unwarranted Variations In Health Care: Shared Decision Making Using Patient Decision Aids. *Health Affairs* **Suppl Variation** VAR63–72.

Osadchiy, N., V. Gaur, S. Seshadri. 2013. Sales forecasting with financial indicators and experts' input. *Production and Operations Management* **22**(5) 1056–1076.

Papastavrou, J., S. Rajagopalan, A. Kleywegt. 1996. Discrete dynamic programming and capital allocation. *Management Science* **42** 1706–1718.

Phillips, R., A. Simsek, G. Van Ryzin. 2013. Does field price discretion improve profits? evidence from auto lending. *Working Paper, Columbia Business School*.

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

Raftery, A. 1995. Bayesian model selection in social research. *Sociological methodology* **25** 111–164.

Ramdas, K., K. Saleh, S. Stern, H. Liu. 2012. New joints more hip? learning in the use of new components. *Working Paper, London Business School*.

Reignier, J., Romain, S. Katsahian, L. Martin-Lefevre, B. Renard, M. Fiancette, C. Lebert, E. Clementi, F. Bontemps. 2008. Patient-related factors and circumstances surrounding decisions to forego life-sustaining treatment, including intensive care unit admission refusal*. *Critical care medicine* **36**(7) 2076–2083.

Reis Miranda, D., M. Jegers. 2012. Monitoring costs in the ICU: a search for a pertinent methodology. *Acta Anaesthesiologica Scandinavica* **56**(9) 1104–13.

Robert, R., J. Reignier, C. Tournoux-Facon, T. Boulain, O. Lesieur, V. Gissot, V. Souday, M. Hamrouni, C. Chapon, J. Gouello. 2012. Refusal of intensive care unit admission due to a full unit impact on mortality. *American journal of respiratory and critical care medicine* **185**(10) 1081–1087.

Shi, P., M. Chou, J. Dai, D. Ding, J. Sim. 2012. Hospital inpatient operations: Mathematical models and managerial insights. *Working paper, Georgia Institute of Technology*.

Shmueli, A., M. Baras, C. Sprung. 2004. The effect of intensive care on in-hospital survival. *Health Services and Outcomes Research Methodology* **5**(3) 163–174.

Shmueli, A., C. Sprung. 2005. Assessing the in-hospital survival benefits of intensive care. *International Journal of Technology Assessment in Health Care* **21**(01) 66–72.

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

Simchen, E., C. Sprung, N. Galai, Y. Zitser-Gurevich, Y. Bar-Lavi, G. Gurman, M. Klein, A. Lev, L. Levi, F. Zveibil, et al. 2004. Survival of critically ill patients hospitalized in and out of intensive care units under paucity of intensive care unit beds. *Critical Care Medicine* **32**(8) 1654–1661.

Simpson, H., M. Clancy, C. Goldfrad, K. Rowan. 2005. Admissions to intensive care units from emergency departments: a descriptive study. *Emergency Medicine Journal* **22**(6) 423–428.

Singer, D., P. Carr, A. Mulley, G. Thibault. 1983. Rationing intensive care physician responses to a resource shortage. *New England Journal of Medicine* **309**(19) 1155–1160.

Sprung, C., D. Geber, L. Eidelman, M. Baras, R. Pizov, A. Nimrod, A. Oppenheim, L. Epstein, S. Cotev. 1999. Evaluation of triage decisions for intensive care admission. *Critical Care Medicine* **27**(6) 1073.

Strand, K., H. Flaatten. 2008. Severity scoring in the icu: a review. *Acta Anaesthesiologica Scandinavica* **52**(4) 467–478.

Strauss, M., J. LoGerfo, J. Yeltatzie, N. Temkin, L. Hudson. 1986. Rationing of intensive care unit services. *Journal of the American Medical Association* **255**(9) 1143–1146.

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

The Kaiser Family Foundation, statehealthfacts.org. 2012. Hospital adjusted expenses per inpatient day, 2009. URL http://www.statehealthfacts.org/.

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

Vanhecke, T., M. Gandhi, P. McCullough, M. Lazar, K. Ravikrishnan, P. Kadaj, R. Begle. 2008. Outcomes of patients considered for, but not admitted to, the intensive care unit*. *Critical Care Medicine* **36**(3) 812–817.

Veatch, M., L. Wein. 1992. Monotone control of queueing networks. *Queueing Systems* **12**(3-4) 391–408.

Weber, R., S. Stidham Jr. 1987. Optimal control of service rates in networks of queues. *Advances in Applied Probability* **19**(1) 202–218.

Weinstein, J., K. Bronner, T. Morgan, J. Wennberg. 2004. Variations in Major Surgery for Degenerative Diseases of the Hip, Knee, and Spine. *Health Affairs* **Suppl Variation** VAR81–89.

Wennberg, J., E. Fisher, J. Skinner, et al. 2002. Geography and the debate over medicare reform. *Health Affairs* **21**(2) 10–10.

Wooldridge, J. 2010. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, MA.

Yom-Tov, G., A. Mandelbaum. 2014. The Erlang-R queue: time-varying QED queues with re-entrant customers in support of healthcare staffing. *Forthcoming in Manufacturing and Operations Management* .

Ziser, A., M. Alkobi, R. Markovits, B. Rozenberg. 2002. The postanaesthesia care unit as a temporary admission location due to intensive care and ward overflow. *British Journal of Anaesthesia* **88**(4) 577–579.

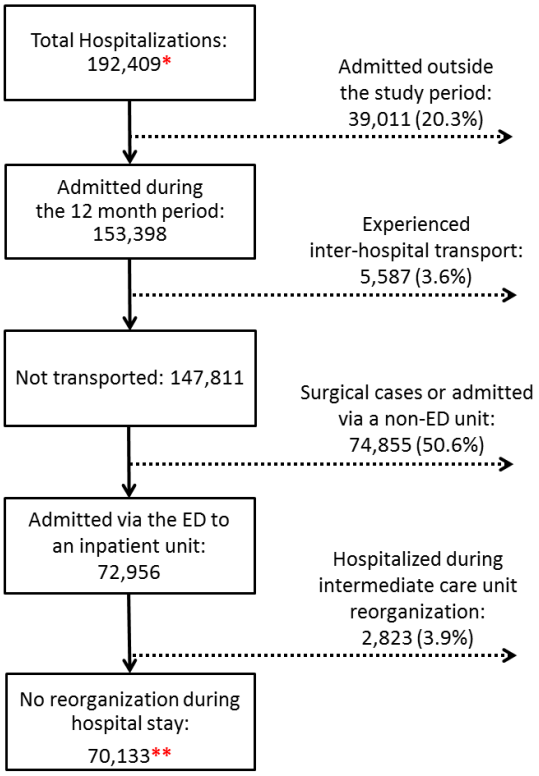

Figure 1: Selection of the patient sample.

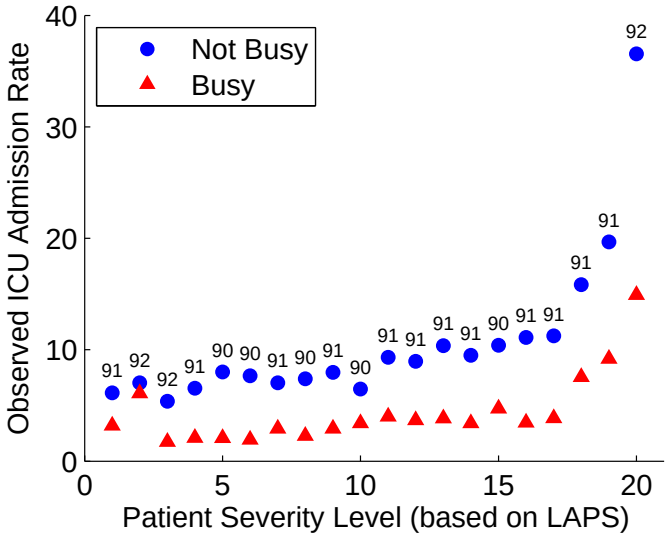

Figure 2: Observed ICU admission rate for patients with different severity levels characterized by LAPS, under high and low ICU occupancy (Busy and Not Busy, respectively). Numbers above circles indicate the fraction of patients (with given severity) that observed a "Not Busy ICU" one hour before their discharge from the ED.

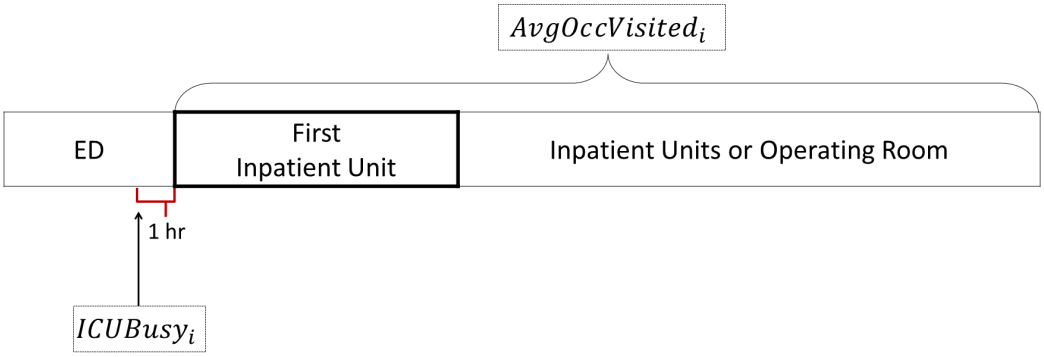

Figure 3: Time-Line of the process flow for patients admitted through the emergency department.

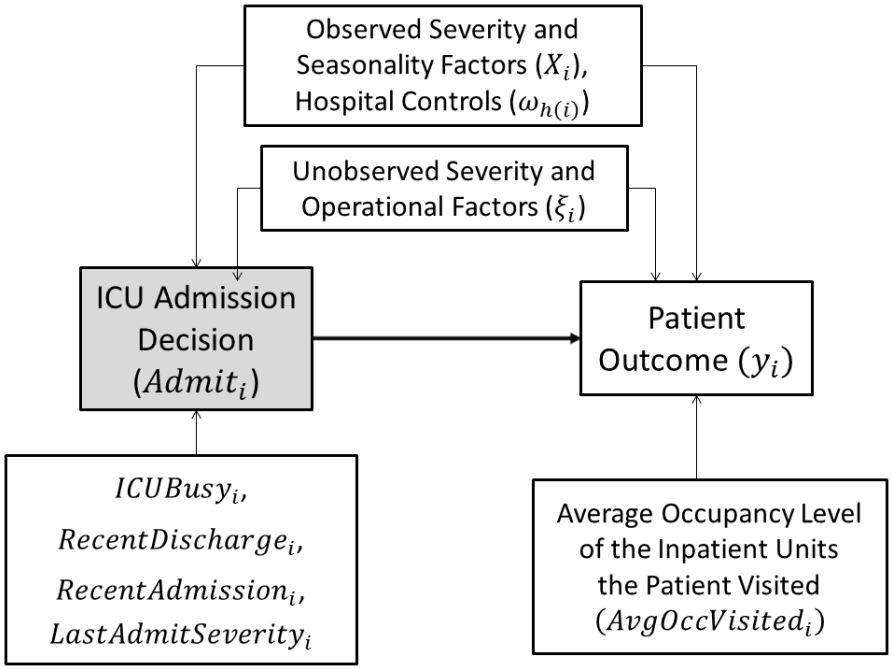

Figure 4: Relationship between ICU admission decision, patient outcome and observed/unobserved patient severity. The instrumental variables used to account for the endogeneity of the admission decision ($Admit_i$) are shown in the bottom-left box.

Table 1: Description of the patient characteristics and seasonality control variables (labeled X_i in our econometric models) used to predict patient outcomes.

Variable	Description and Coding
Age	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
Gender	Females were coded 1 and males 0
Severity score 1: LAPS	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
Severity score 2: $\hat{P}$(Mortality)	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
Admitting diagnosis	grouped into one of 44 broad diagnostic categories such as pneumonia; categorical variable to denote each diagnosis
Month/Time/Day	Month/Time/Day of week of ED admission; categorical variables

Table 2: Summary statistics of patient characteristics, grouped by whether their first inpatient unit was an ICU versus non-ICU bed.

	Non-ICU	ICU	ALL
Num. of obs.	63197	6936	70133
Selected X Covariates			
Age	67.3 (17.8)	64.0 (18.0)	67.0 (17.8)
LAPS	23.5 (18.1)	36.1 (25.2)	24.7 (19.3)
$\hat{P}(Mortality)$	0.044 (0.067)	0.095 (0.131)	0.049 (0.077)
Female	0.546	0.495	0.541
Z Covariates			
$ICUBusy$	0.096	0.039	0.091
$RecentDischarge$	0.033 (0.048)	0.040 (0.052)	0.034 (0.049)
$RecentAdmission$	0.009 (0.022)	0.009 (0.021)	0.009 (0.022)
$LastAdmitSeverity$	0.341	0.311	0.338

Note. Average and standard deviation (in parentheses for continuous variables) are reported.

Table 3: Summary statistics of the patient outcomes

Outcome	n	Mean	Standard deviation	Median
Mortality	70,133	0.04	-	-
TransferUp	68,200	0.03	-	-
Readmission - 2 weeks	67,087	0.10	-	-
Hospital LOS (days)	70,133	3.9	4.9	3.0

Table 4: Estimation results of the effect of ICU admission on patient outcomes.

	With IV					Without IV
Outcome	Estimate (SE)	AME	ARC	ρ (SE)	Test $\rho = 0$	Estimate (SE)
Mortality	0.01 (0.13)	0.001	+1.6%	0.20** (0.07)	0.00	0.42*** (0.03)
Readmit	-0.22^{+}(0.13)	-0.034	-32.2%	0.15* (0.07)	0.03	0.05* (0.02)
TransferUp	-0.65*** (0.16)	-0.028	-77.3%	0.32** (0.10)	0.00	-0.08* (0.04)
HospLOS (days)	-0.44*** (0.01)	-1.2	-33.0%	0.56*** (0.01)	0.00	0.28*** (0.01)

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 5: Simulation Results of Alternative ICU Admission Control Policies

	Estimated Current Policy		Optimal Objective	Optimal Full
	BASE - 21 Beds	22 Beds	for Each Outcome	for Each Outcome
# Readmissions	**2550**	-3.7	-26.6	-35.3
# Transfer-ups	**762.9**	-5.9	14.8	-38.6
Hospital LOS in years	**245.6**	-0.4	-2.0	-9.0
Total Hospital LOS in years	**272.9**	-0.4	-2.2	-9.2
(Estimated savings in dollars)		(-$ 0.4 m)	(-$ 1.9 m)	(-$ 8.1 m)

Note. The performance measures of the estimated current policy are denoted in bold; all other results are changes from the performance of the estimated current policy.

A Appendix: Accounting for Speed-up and ED Boarding effects

This appendix describes in detail the econometric models used in the analysis of Section 4.2.

A.1 Speed-up in the ICU

We describe the methodology used to measure the effect of ICU congestion on patient length of stay in the ICU. The methodology replicates the approach developed IN Kc and Terwiesch (2012); see that article for further details.

Define *firstICU LOS*$_i$ as the ICU length of stay during patient i's first ICU visit and $BUSY_i$ 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 $BUSY_i$ is not the same as in Kc and Terwiesch (2012). Instead, we let $BUSY_i$ 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.[12] We estimate the effect of ICU occupancy on ICU length of stay through the following regression:

$$log(firstICU\ LOS_i) = \gamma BUSY_i + \beta X_i + u_i, \tag{6}$$

where X_i is a vector of observable patient characteristics that describe the patient's severity of illness. A negative γ suggests that high ICU congestion leads to a shorter ICU LOS – a speed-up effect.

The regression model (6) is estimated with two samples of patients that were admitted to the ICU: (1) Surgical patients and (2) ED patients. The estimation results are reported in Table 6.

Table 6: Estimation results of model (6)

	$Busy$ Coefficient (Standard Error)	# Observations	R^2
ED, Medical	-0.02 (0.03)	10521	0.16
Non-ED, Surgical	-0.13** (0.04)	4524	0.14

Note. $^{+}(p < 0.1), *(p < 0.05), **(p < 0.01), ***(p < 0.001)$.

A.2 ED Boarding Time

In this section, we describe how we estimate an alternative specification of the outcome models that accounts for the effect of the endogenous variable ED boarding time.

ED boarding time ($EDboard$), defined as the time between the decision to hospitalize the patient until the patient is discharged from the ED and physically moved to the inpatient unit, is added as an additional covariate in the outcome models (1) and (3). This new specification has two endogenous covariates: $Admit$ and $EDboard$; the former is instrumented by $ICUBusy$ so we need additional instruments for the latter. A valid exogenous instrumental variable

[12] We have tried various specifications for defining $BUSY$, 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 $BUSY$ measure included as both time-invariant and time-varying, and the results were consistent.

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 $FirstInpatientOcc$, where the average is taken during the time the patient is boarding in the ED. The logic is similar to our $ICUBusy$ 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. Recall that $ICUBusy$ is based on the level of occupancy of the ICU one hour prior ED discharge, whereas $FirstInpatientOcc$ measures the occupancy of ICU or the ward, depending on where the patient is routed to after the ED. Hence, the two instrumental variables are not perfectly correlated. A regression of the logarithm of ED boarding time on $FirstInpatientOcc$ shows a positive and highly significant effect; a 10% increase in the inpatient occupancy increases ED boarding time by 18%. For this model, we use similar controls as in our earlier specification. Details of the regression output is available from the authors upon request.

The estimation of the model goes as follows. 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 *log(ED Board)* 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 *log(ED Board)* 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 *log(ED Boarding Time)* for the different outcome models.

Table 7: Estimation results of the patient outcome model including ED Boarding Time as an endogenous covariate.

Outcome	ICU admission	Log(ED board)
Mortality	0.03 (0.13)	0.05 (0.04)
Readmit	-0.21 (0.13)	-0.01 (0.03)
TransferUp	-0.61*** (0.16)	0.16*** (0.04)
HospLOS (days)	-0.40*** (0.01)	0.01 (0.01)

*Note.*Standard errors in parentheses. $^{+}(p < 0.1), *(p < 0.05), **(p < 0.01), ***(p < 0.001)$.