MANAGEMENT SCIENCE
Vol. 61, No. 1, January 2015, pp. 19–38
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.2057

ICU Admission Control: An Empirical Study of Capacity Allocation and Its Implication for Patient Outcomes

Song-Hee Kim
Yale School of Management, Yale University, New Haven, Connecticut 06520, hailey.kim@yale.edu

Carri W. Chan
Columbia Business School, Columbia University, New York, New York 10027, cwchan@columbia.edu

Marcelo Olivares
Universidad de Chile, Santiago, Chile, molivares@u.uchile.cl

Gabriel Escobar
Division of Research, Kaiser Permanente, Oakland, California 94612, gabriel.escobar@kp.org

This work examines the process of admission to a hospital's intensive care unit (ICU). ICUs currently lack systematic admission criteria, largely because the impact of ICU admission on patient outcomes has not been well quantified. This makes evaluating the performance of candidate admission strategies difficult. Using a large patient-level data set of more than 190,000 hospitalizations across 15 hospitals, we first quantify the cost of denied ICU admission for a number of patient outcomes. We use hospital operational factors as instrumental variables to handle the endogeneity of the admission decisions and identify important specification issues that are required for this approach to be valid. Using the quantified cost estimates, we then provide a simulation framework for evaluating various admission strategies' performance. By simulating a hospital with 21 ICU beds, we find that we could save about $1.9 million per year by using an optimal policy based on observables designed to reduce readmissions and hospital length of stay. We also discuss the role of unobserved patient factors, which physicians may discretionarily account for when making admission decisions, and show that including these unobservables could result in a more than threefold increase in benefits compared to just optimizing the policy over the observable patient factors.

Keywords: healthcare delivery; empirical operations management; dynamic programming; capacity allocation; admission control; congestion; quality of service
History: Received May 17, 2012; accepted July 21, 2014, by Serguei Netessine, operations management. Published online in *Articles in Advance* November 20, 2014.

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 United States (Halpern et al. 1994, Rainey et al. 1994). Most hospital ICUs operate near full capacity (Green 2003, Pronovost et al. 2004), making ICU beds a limited resource that 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 to make these decisions, and how these decisions should vary under different scenarios.

The obvious ICU admission criteria are that very sick and unstable patients should be treated in the ICU, whereas stable patients do not require ICU care. However, identifying 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 (Fisher et al. 2004, Mullan 2004, O'Connor et al. 2004, Weinstein et al. 2004, Boumendil et al. 2012, Chen et al. 2012). Although ICU admission, discharge, and triage standards have been established by a critical care task force (Task Force of the American College of Critical Care Medicine, Society of Critical Care Medicine 1999), they are subjective in nature; the task force even admits that "[t]he criteria listed, while arrived at by consensus, are by necessity arbitrary" (p. 636). 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 metrics to characterize the benefits of different practices.

Our work takes an important step toward addressing this issue by estimating the cost of denied ICU care for *all* medical patients admitted to the hospital through the emergency department (ED). We focus on patients admitted through the ED because their care is the most likely to be affected by not only each patient's severity of illness but also hospital operational factors (due to the typical uncertainty in the volume and clinical severity of incoming patients in the ED). For ethical reasons, it is not possible to run a field experiment to randomize ICU treatment to patients to estimate this cost. Hence, we utilize observational data—as was done in prior research to measure the impact of ICU care 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)—from 15 hospitals covering more than 190,000 hospitalizations, of which we consider the admission decisions of over 70,000 patients.

Working with observational data presents an important econometric challenge: The decision to admit a patient to the ICU is endogenous. Specifically, there are factors related to patients' clinical severity that the deciding physicians take into account but that are unobserved in the data; such unobservables will be positively correlated with the ICU admission decision and adverse patient outcomes, generating a positive bias in the estimate of the causal effect of ICU admission. Prior studies by Shmueli et al. (2004) and Kc and Terwiesch (2012) propose using the ICU congestion level as an instrumental variable (IV), but we argue that using only this variable might violate the required exogeneity assumption of a valid IV. To be a valid IV, ICU congestion should affect patient outcomes only through its effect on the ICU admission decision, but since hospital resources are shared among patients, a congested ICU could *directly* impact the patient's recovery and therefore patient outcomes. Hence, in addition to using ICU congestion as an IV, we utilize our rich data to measure occupancy information on every unit that each patient visits and thereby separate the effect of ICU congestion on the ICU admission decision from its direct effect on patient outcomes. Many U.S. hospitals have started to collect data similar to what we use in this work, so the proposed methodology is applicable in other hospital settings.

Our analysis shows that ICU admission can reduce adverse patient outcomes in the range of 30%–75%, depending on the outcome. We also find that the impact of ICU admission is highly variable for different patients and outcomes, which supports the importance of understanding the varying impacts when making admission decisions. 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.

We use the estimated impact of ICU care on patient outcomes to compare the performance of various ICU admission strategies that use different types of information. We first estimate the current admission criteria of a single, representative hospital out of the 15 hospitals represented by our data. Note that the current admission criteria utilize both *observed* patient measures (i.e., recorded in our data set) and *unobserved* measures (i.e., not recorded in our data set but that physicians can potentially notice). These unobservables, such as the patient's cognitive state, have the potential to provide more information about the impact of ICU admission. Moreover, how these observable and unobservable factors are used to make an admission decision is subject to the physicians' discretion.

We also consider an objective and clearly defined policy that is optimized based on observed metrics alone. Via simulation, we find that using this *optimal policy* instead of the estimated current policy at the hospital that we selected can translate into savings of patient bed hours on the order of 2.2 years, equivalent to $1.9 million. We also show that this benefit is approximately five times larger than the benefit of adding an additional ICU bed, excluding the costs of maintaining the extra bed.

Even though our proposed optimal policy outperforms the estimated current policy in terms of certain patient outcomes that we consider, it does not do so for others; the proposed optimal policy does not account for valuable patient information that is unobserved in the data but that can be taken into account in the admission policy currently used by physicians. For this reason, we also simulate an optimal policy that incorporates both observed and unobserved information. We find that doing so improves patient outcomes unilaterally and results in a more than threefold increase in benefits compared to just optimizing the policy over the observables. We find similar results when we simulate the aforementioned policies at other hospitals represented by our data, but we emphasize that our findings might not be universal; to improve the ICU admission process at a specific hospital, one should utilize our estimation and simulation framework to assess the target hospital in question. Doing so provides evidence to help to convince managers and clinicians, who may often be reluctant to alter their current practices, of the potential benefits of one policy over the others.

In summary, we make the following key contributions:

• *Evaluation of patient outcomes*: To evaluate the performance of various admission policies, we must quantify the impact of ICU admission. Using a large patient-level data set of more than 190,000 hospitalizations across 15 hospitals, we quantify the cost of denied ICU admission for a number of patient outcomes, including hospital length of stay (LOS), hospital readmission, and patient transfers to higher levels of care. Although our estimation approach using instrumental variables has been used in previous work (Shmueli et al. 2003 and Kc and Terwiesch 2012), we make important methodological contributions by identifying key control variables that are required for the validity of this IV estimation.

• *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 specifically compare the derived optimal admission policies with the estimated current hospital admission policies and find that in some circumstances, it is useful to base admission decisions on observed metrics of patient risk alone, whereas in other scenarios, unobservables can provide valuable information. We are also able to quantify the benefit of unobserved information by examining how much patient outcomes improve when optimizing the admission decision based on both unobserved and observed measures versus observed measures alone.

The rest of this 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. Section 3 develops our econometric model to estimate the effect of ICU admission on patient outcomes, and Section 4 provides the estimation results. Section 5 uses the empirical results from §4 to develop a simulation study to compare the performance of hospitals' current ICU admission policies with alternative approaches. Section 6 summarizes our main contributions and provides guidelines for future research.

1.1. Literature Review

A number of works in healthcare operations management (OM) 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 transport time within the hospital, and Kuntz et al. (2014) examine the impact of hospital load on in-hospital mortality using the idea of safety tipping points. Jaeker and Tucker (2013) report that the LOS depends on current workload as well as the predictability and the pressure level of the incoming workload, and Batt and Terwiesch (2014) examine workload-dependent service times in the ED. Green et al. (2013) find that nurse absenteeism rates in an ED are correlated with anticipated future nurse workload levels, whereas Ramdas et al. (2012) and Kc and Staats (2012) study the impact of surgeon experience on outcomes. In contrast to these works, we specifically analyze the ICU admission decision.

A more specific area of interest within this broader space is studying adaptive mechanisms for managing ICU capacity. For instance, when a hospital does not have sufficient downstream bed capacity, surgical cases may be either delayed or canceled (Cady et al. 1995). In particular, when a new patient requires ICU care but there is no available bed, his or her care may be delayed, and the patient may be boarded 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) show that ED boarding caused by a congested ICU is an important factor driving ambulance diversion.

Speeding up the treatment of patients in a busy ICU is another mechanism that has received considerable attention from the OM and medical communities. 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 the ICU occupancy level on discharge practices in a cardiac surgical ICU and find that congested ICUs tend to speed up the treatment of their patients and that these affected patients are readmitted to the ICU more frequently. Admission and discharge decisions are fundamentally very different, utilizing different information and criteria. Hence, the detailed understanding of discharge decision making established by Kc and Terwiesch (2012) cannot provide insight into the admission decision.

Indeed, another method of managing ICU capacity is admission control, which is the topic of this paper. During periods of high congestion, some patients who may benefit from ICU care might be denied access because the ICU is full or because all available beds are being reserved for more severe incoming patients. Studies have confirmed that 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). This was observed not only in U.S. hospitals but also by researchers in many 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 United Kingdom.

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 age less than 39 was coded 1, 40–64 was coded 2, 65–74 was coded 3 (Medicare starts at 65), 75–84 was coded 4, and above 85 was coded 5
Gender	Females were coded 1 and males 0
Severity of illness 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 the 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, and 89
Severity of illness score 2: $\hat{P}$(*Mortality*)	An estimated probability of mortality (Escobar et al. 2008); predictors include the LAPS and Comorbidity Point Score (which measures the chronic illness burden and is based on 41 comorbidities; comorbidities are chronic diseases, such as diabetes, that may complicate patient care and recovery); coded as piecewise linear spline variables with knots at 0.004, 0.075, and 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

Closest to our work is that of Shmueli et al. (2003), which examines the impact of denied ICU admission on mortality among patients who have been referred for ICU admission. They use an IV approach to measure how ICU admission decreases mortality for patients of different levels of severity of illness and suggest possible ICU admission criteria. Their study cannot answer our research question for the following four reasons. First, Shmueli et al. (2003) use APACHE II—one of several ICU scoring systems whose scores are generally assigned based on data available within the first 24 hours of an ICU stay (Strand and Flaatten 2008)—to control for patient severity. APACHE II is not available for a typical ED patient and hence, as argued by Franklin et al. (1990), cannot be used to control for patients' severity of illness when deciding which (of all) ED patients should be admitted to the ICU.[1] In contrast, the hospitals that we analyze use uniform metrics of patients' severity of illness that are available for all admitted patients: the Laboratory-based Acute Physiology Score (LAPS) and the estimated probability of mortality. (See Table 1 for details. Previous work by Van Walraven et al. 2010 shows that the LAPS is a reasonable predictor of patient LOS and mortality.) We utilize these two measures to successfully control for patient severity in our study. Second, these authors' ICU admission criteria cannot be generalized to the (much larger) cohort of patients admitted from the ED. (In their study, 84% of patients were admitted to the ICU, whereas in our sample, only 9.9% were admitted.) In particular, the benefit of ICU care may be exaggerated in the work of Shmueli et al. (2003) because they only consider patients whose physicians have already determined that they require ICU care. Third, there is likely substantial variation in which patients will be recommended for ICU admission across hospitals and physicians because of heterogeneity in physicians' backgrounds, training, and opinions (Mullan 2004, Weinstein et al. 2004, Fisher et al. 2004, O'Connor et al. 2004). In a later study, Shmueli and Sprung (2005) explicitly state that the admission policy in the ICU that they are studying does not maximize the benefits of the ICU and that "the discrepancies actually originate from [an] inappropriate referral policy" (p. 71). In contrast, our study provides criteria to use *before* any subjectivity in the preselection 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 of ICU admission on mortality is similar across many patients, whereas the impact is highly variable for other outcomes, such as LOS and readmission. Accurately quantifying these effects is necessary when determining the optimal ICU admission decision.

When deriving the optimal admission policies that utilize our quantified estimates of the benefits of ICU admission, we draw upon the rich literature on the *stochastic knapsack problem*: see Miller (1969), Weber and Stidham (1987), Veatch and Wein (1992), Glasserman and Yao (1994), Papastavrou et al. (1996), and

[1] In fact, all of the aforementioned studies on ICU admission decisions control for patients' severity of illness using measures that are available only after patients are admitted to an ICU. Examples include the Acute Physiology and Chronic Health Evaluation II (APACHE II) score (Shmueli et al. 2004, Shmueli and Sprung 2005), the Simplified Acute Physiology Score II (SAPS II) (Iapichino et al. 2010, Simchen et al. 2004), the Simplified Therapeutic Intervention Scoring System (TISS) (Simchen et al. 2004), and the Mortality Prediction Model (MPM) (Louriz et al. 2012). See Strand and Flaatten (2008) for a review of these measures.

references therein. Specifically, we use a special case of the stochastic knapsack problem studied by Altman et al. (2001) and leverage some results from that work to characterize the optimal policy in our setting.

As we evaluate alternative admission policies in §5, we also discuss the role of unobserved metrics that are accounted for by the physician in the patient admission decision. The value of discretionary criteria (or experts' input) in decision making has started attracting interest in other areas of operations management (e.g., see Anand and Mendelson 1997, Osadchiy et al. 2013, Phillips et al. 2015). 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 data set comprising nearly 190,000 hospitalizations at 15 hospitals over the course of one and a half years, collected using a comprehensive electronic medical records system. 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, which allows us to ignore the potential impact that insurance status may have on the care pathways 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, the inpatient units can be broadly divided according to their varying nurse-to-patient ratios and treatment and monitoring levels. Generally, 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 a ratio of 1:3.5 to 1:4, and intermediate care units, with a ratio of 1:2.5 to 1:3, though not all hospitals have intermediate care units. Although there is some differentiation within each level of care, the units are relatively fungible, so if the medical ICU is very full, a patient may be admitted to the surgical ICU instead.

We focus on the ICU admission decision for patients who were admitted to a medical service at the hospital through the ED for the reasons discussed in the introduction. In our data set, about 55% (52%) of patients admitted to the hospital (ICU) were admitted via the ED to a medical service. The admission process works as follows. If an ED physician believes that a patient is eligible for ICU admission, an intensivist will be called to the ED for consultation. Although the intensivist makes the ultimate decision about whether to admit the patient from the ED, the decision is typically based on 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.

Table 2 Summary Statistics of Patient Characteristics, Grouped by Whether Their First Inpatient Unit Was an ICU vs. Non-ICU Bed

	Non-ICU	ICU	All
No. of obs.	63,197	6,936	70,133
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 deviations (in parentheses for continuous variables) are reported.

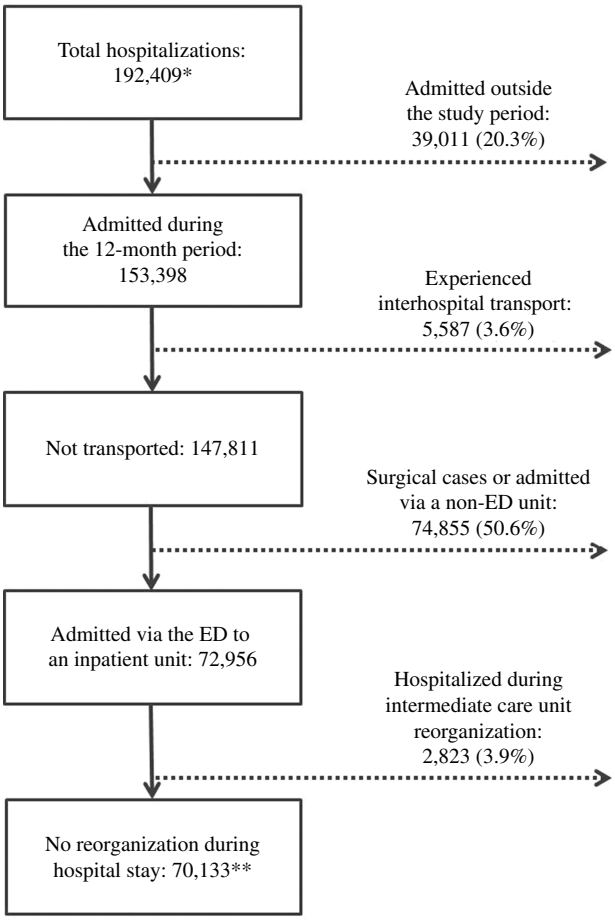
The patient-level information in our data set includes patient age, gender, admitting diagnosis, hospital, and two severity-of-illness scores. This information is described in detail in Table 1. Table 2 provides summary statistics for the covariates for all of the patients in our sample as well as for when patients are grouped by whether they were admitted to the ICU or not. In addition, we collect operational data that include every unit that each patient visits, along with unit admission and discharge dates and times. Since we have an inpatient data set, we do not have information on patients who were 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. The hospitals represented by our data set have heterogeneously sized inpatient units. Because defining congestion in a small ICU is challenging and because different mechanisms might be used to allocate beds in small ICUs, we consider only the patients who were treated in hospitals with ICUs of 10 or more beds. There are 15 such hospitals, and among them, the maximum ICU occupancy varied from 10 to 44. The average percentage of ICU beds among inpatient beds was 12.9%, with a minimum of 9.3% and a maximum of 21.5%.

We utilize patient flow data from all 192,409 patient visits at the selected 15 hospitals (indicated by one asterisk in Figure 1) to derive the capacity and instantaneous occupancy level of each inpatient unit. Because our data set consists of patients admitted and discharged within a 1.5-year time period, we restrict our study to the 12 months in the center of the period

Figure 1 Selection of the Patient Sample



to avoid censored estimation of capacity and occupancy. We exclude patients who experienced interhospital transport, as it is difficult to determine whether it was due to medical or personal needs. For the reasons explained in the introduction, we focus on the patients who were admitted via the ED to a medical service. The sizes of the inpatient units were quite stable over our study period. However, four hospitals underwent a small change in the capacity of the intermediate care unit, so we exclude patients who were hospitalized during these rare occurrences of intermediate care unit reorganization (such as reducing the number of beds). Thus, our final data set consists of 70,133 hospitalizations, as indicated by two asterisks 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 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 a lower level of care or will be discharged from the hospital as his health state improves. In contrast, 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 a lower level of care.[2] Note that patients who are admitted to and directly discharged from the ICU can never experience this event, so we study *TransferUp* over the subset of patients who visited the general ward at least once during their hospital stay.

Table 3 Summary Statistics for the Patient Outcomes

Outcome	*n*	Mean	Std. dev.	Median
Mortality	70,133	0.04	—	—
Readmit—two weeks	67,087	0.10	—	—
HospLOS (days)	70,133	3.9	4.9	3.0
TransferUp	68,200	0.03	—	—

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, based on 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 do not include patients with in-hospital death, as they could not be readmitted.

We use *HospLOS* as a measure of the time from admission to the first inpatient unit until hospital discharge, 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 were discharged between 10 A.M. and 3 P.M., whereas admission times were less concentrated and demonstrated a markedly different distribution (a similar issue was reported by Armony et al. 2010 and Shi et al. 2015 using data from other hospitals). To avoid this source of measurement error, we measure *HospLOS* as the number of nights that 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.

[2] Durbin and Kopel (1993) show that ICU readmission, which qualifies as a *TransferUp* event, leads to higher mortality and a longer LOS.

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 LOS. Section 3.1 develops an econometric model to measure the impact of ICU care on these outcomes, in which the main challenge is to account for the endogeneity in ICU admission decisions. Section 3.2 develops an estimation strategy using instrumental variables (IVs) to address the challenge. Section 3.3 describes our final estimation models.

3.1. Econometric Model for Patient Outcomes

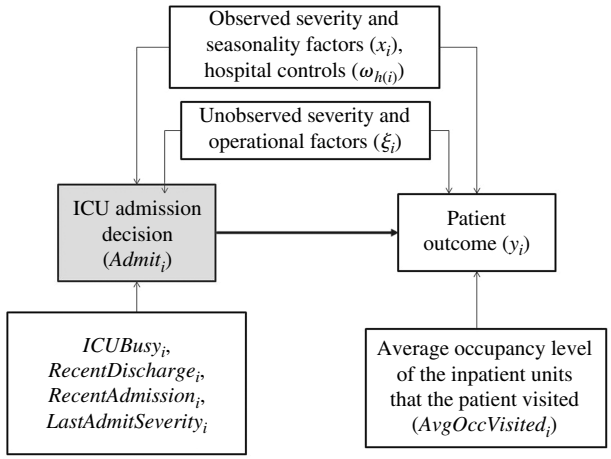
An ideal thought experiment to examine the implications of ICU admission for patient outcomes would be randomizing patient allocation to the ICU and non-ICU units, regardless of the severity of their condition. Of course, such an experiment would be impossible in practice because of ethical concerns.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

Our unit of observation is a hospital visit by 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's severity of illness measures that are useful in predicting patient outcomes. For example, Escobar et al. (2008) and Liu et al. (2010) illustrate how clinical severity measures based on automated laboratory and comorbidity measures can be used to successfully predict in-hospital mortality and hospital LOS. Let X_i denote those clinical severity factors as well as patient characteristic and seasonality controls that are observed in the data (see Table 1 for a detailed description of X_i). We also control for hospitals and let $\omega_{h(i)}$ denote the coefficients for a set of hospital indicator variables, where $h(i)$ is patient i's hospital. Our main hypothesis is that ICU treatment has a causal effect on patient outcomes. Accordingly, we let $Admit_i = 1$ if patient i is admitted to the ICU and zero otherwise. We model the patient outcome y_i as a random variable with distribution $f(y_i \mid \beta_1, \beta_2, Admit_i, X_i, \omega_{h(i)})$, where the parameter β_1 captures the effect of ICU admission and β_2 measures the effect of the observable characteristics X_i on the patient outcome. This distribution could be given by a model of the following form:

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

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).

Figure 2 Relationships Between the ICU Admission Decision, Patient Outcome, and Observed/Unobserved Severity of Illness



Note. The instrumental variables used to account for the endogeneity of the admission decision ($Admit_i$) are shown in the bottom left box.

The linear regression example in (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 squares (OLS) and to interpret the estimate of β_1 as the causal effect of ICU admission on the outcome. This approach ignores the fact that the admission decisions are endogenous; that is, aspects of patient severity that are unobservable in the data (e.g., the patient's cognitive state) are likely to affect admission decisions. Figure 2 illustrates this endogeneity issue in further detail. The term ξ_i represents clinical severity characteristics that are unobserved in the data but that are considered by 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 estimation to obtain consistent estimates of this linear regression model. We propose using hospital operational factors as IVs and describe and validate our choices in the next section.

3.2. Instrumental Variables

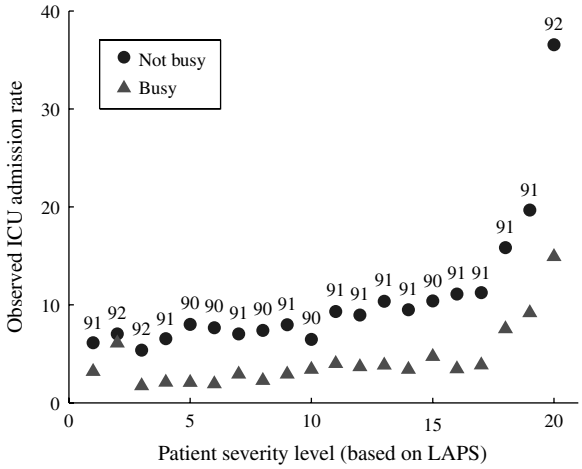
A valid instrumental variable, denoted by Z, needs to satisfy the following two conditions: (1) it has to influence the endogenous variable, which is the ICU admission decision, or $Admit_i$, in our case, and (2) it has to be exogenous, meaning that it cannot affect the patient outcome measure y_i other than through

the admission decision. We discuss several potential instruments in this section.

When deciding whether to admit an ED patient to the ICU, hospitals need 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 the 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 because the volume and clinical severity of incoming patients are 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.[3] Figure 3 graphs the admission rates for 20 different patient groups (classified by their LAPS on the horizontal axis) for two different occupancy levels: busy (marked with triangles) and not busy (marked with circles). Note that all 40 points in this graph represent enough observations (the smallest sample size was 144 patients) to give us meaningful rates on the y axis. The level of ICU occupancy associated with each patient was measured one hour prior to their ED discharge, which is a reasonable time period to cover the stage at which admission decisions are made. Above the circles, we also show the percentage (90%–92%) of the patients in each patient severity group who experienced an ICU that was not busy. That is, ICU admission decisions for patients at all clinical severity levels are affected by ICU occupancy; among patients in the same clinical severity group, a lower percentage of patients who experienced high ICU occupancy was sent to the ICU compared to the patients who experienced a low ICU occupancy level. We can repeat the same exercise for other cutoffs of ICU occupancy, including the 90th, 85th, and 75th percentiles, resulting in a change in admission rate that is much smaller or nonexistent for some groups of patients. Although other measures of ICU occupancy could be considered, Figure 3 visually supports the concept that $ICUBusy_i$ is a powerful instrument in the sense that it causes significant variation in the admission decision.

Figure 3 **Observed ICU Admission Rate for Patients at Different Severity Levels, as Characterized by Their Laps, Under High and Low ICU Occupancy (Busy and Not Busy, Respectively)**



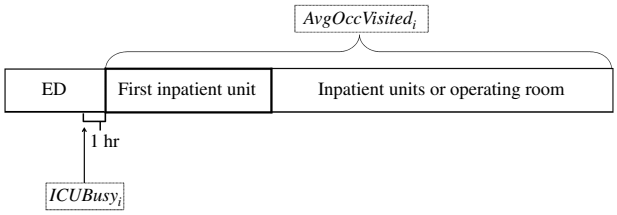
Note. The numbers above the circles indicate the fraction of patients (at a given severity of illness) who experienced a "not busy" ICU one hour before their discharge from the ED.

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, showing that readmission rates tend to be higher for patients who experienced a high ICU occupancy level during their ICU stay. One could imagine that this mechanism may apply to other patient outcomes and to other inpatient units. To overcome this issue, we take advantage of the fact that we have the complete care path of each patient in our data set, and we control for the congestion levels that a patient experienced in each of the visited inpatient units *during* his or her hospital stay. Specifically, let D_i be the set of days during which patient i stayed in the hospital (after leaving the ED) and $Occ_{i,d}$ be the occupancy of the inpatient unit where patient i stayed on day d. The average occupancy of the inpatient units visited by the patient during his or her hospital stay is defined as $AvgOccVisited_i = (1/|D_i|)\sum_{d\in D_i} Occ_{i,d}$ (see Figure 4 for a timeline that illustrates when this measure is calculated).[4] We include $AvgOccVisited_i$ as an additional control variable in the outcome model (1).

[3] For instance, suppose that an ICU has its occupancy at eight beds or below 94.5% of the time and at nine beds or below 95.8% of the time; this ICU is then considered busy when nine or more of its beds are occupied. Also note that we estimated the occupancy distribution by measuring the ICU bed occupancy every hour in the study period.

[4] We define the capacity of an inpatient unit as the 95th percentile of the bed occupancy distribution of that unit to compute $Occ_{i,d}$ because in many occasions, the maximal capacity is rarely observed, as hospitals may temporarily expand their standard capacity by a few beds in extreme circumstances (this was also pointed out by Armony et al. 2010 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 a median of 0.86, in our data set.

Figure 4 Timeline of the Process Flow for Patients Admitted Through the Emergency Department



Separating the effect of occupancy on the admission decision from its effect during the inpatient hospital stay is essential to have a proper IV identification strategy. However, 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. Note that $AvgOccVisited_i$ is not perfectly correlated with $ICUBusy_i$ because the latter is measured *before* patient i is physically moved to the inpatient unit and because the occupancy level typically varies during patient i's hospitalization period; the correlation between the two measures is 0.24 in our sample.

Another mechanism that could invalidate using $ICUBusy_i$ as an IV is when periods of high congestion coincide with the arrival of very severe patients, e.g., during an epidemic or a major accident affecting a large portion of the hospital's patient population. When we test this potential mechanism by analyzing the relationship between hospital occupancy and the LAPS (see Table 1 for a description), we find 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 clinical severity are not (this approach is 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 a comparison of differences in outcomes among patients who have similar observable characteristics captured by X_i but who received different treatments because of the different ICU occupancy levels at the time of their admission to an inpatient unit. Although this is not a perfectly randomized experiment, the 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 during our conversations with nurses, physicians, and hospital management. We refer to these variables 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. However, the intensivist in charge may have an incentive to "preserve the nurse hours" by demonstrating a continuous demand for those nurses, even after their patients are discharged,[5] 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 by 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 three-hour window before patient i's admission to the first inpatient unit. In our sample, 56% of the patients underwent no recent ICU discharges, 27% underwent one discharge, and 11% underwent 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 determine $RecentDischarge_i$.

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 or her negotiation with the intensivist. To measure $RecentAdmission_i$, we consider ICU admissions in the two-hour window before patient i's admission to the first inpatient unit but count an admission as a recent admission only if the patient is admitted via the ED to a medical service (excluding those who 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 set, 84% of the patients did not undergo recent admission, and 14% underwent one recent admission. Similar to how we calculated $RecentDischarge_i$, 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 clinical 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 or she might require a new patient to also be

[5] This behavior is related to supply sensitive demand that has been shown in the medical literature. For instance, see Wennberg et al. (2002) and Baker et al. (2008).

very sick before recommending ICU admission. We define $LastAdmitSeverity_i$ as a dummy variable indicating whether the last patient admitted to the ICU had a LAPS greater than or equal to the 66th percentile value of the observed LAPS distribution.

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's severity of illness 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 we implement the estimation using these IVs as instruments for the endogenous variable $Admit_i$.

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 approach to implement the IV estimation. However, because the ICU admission decision and all of our patient outcomes are discrete, a more efficient estimation approach is to develop nonlinear parametric models to characterize them and to jointly estimate the admission decision model and each of the patient outcome models. We describe this approach next.

The ICU admission decision is binary, and we model it through a probit model defined by

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

where X_i is observable patient characteristics, Z_i is 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 count 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, as follows:

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

where y_i^* is the latent variable. As previously discussed, the additional control $AvgOccVisited_i$ captures the effect of the congestion during the hospital stay of the patient. To account for the endogeneity in ICU admission decisions, represented by $Admit_i$, 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, which will 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 that can be estimated via the full maximum likelihood estimation (Cameron and Trivedi 1998, Wooldridge 2010). The endogeneity of the admission decision $Admit_i$ can be tested through a likelihood ratio test of the correlation coefficient ρ being nonzero.

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 overdispersion: Table 3 shows that the mean of $HospLOS_i$ is 3.9, and the variance is 24.0. Hence, we use the negative binomial regression, which can model overdispersion using the parametrization developed by Cameron and Trivedi (1986). We use the extension developed by Deb and Trivedi (2006) to include a binary endogenous variable—the ICU admission decision $Admit_i$—in the negative binomial regression and estimate it jointly with (2). The negative binomial regression includes the same covariates as in (3).

4. Estimation Results

As discussed in §3.3, we estimate the admission decision and patient outcome models 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% level).

Table 4 summarizes the results of the patient outcome models, and each row corresponds to a different patient outcome. Note that because of space limitations, we show 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. 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), defined as the average expected absolute change in the outcome (among all patients) when a patient is admitted to the ICU instead of to a ward. The average relative change (ARC) is also reported, which is the AME

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)

Notes. Each row corresponds to a different outcome (the dependent variable). AME, average marginal effect; ARC, average relative change. Standard errors in parentheses.

$^{+}p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

divided by the average outcome when patients are 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 in which the null hypothesis is that the ICU admission decision is exogenous. This test is equivalent to a likelihood ratio test against the model in which the correlation coefficient between the admission and the outcome models' errors, ρ, is restricted to be zero. The estimates of ρ are reported in the column "ρ (SE)." The null hypothesis is strongly rejected in all models; that is, 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 that affect the coefficients when ignoring the admission decision's endogeneity. This is consistent with the endogeneity problem discussed in Figure 2. ICU patients tend to be more severely ill, and because part of 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 strong that it leads to a positive correlation between being admitted to an ICU and experiencing adverse outcomes.

We do not find a significant effect of ICU admission on mortality rates, which is at first surprising, given the magnitude of the effect for other outcomes. A possible explanation of this finding 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 who are affected by the instrument*. In our context, this includes patients whose ICU admission decision is affected by the ICU congestion one hour prior to their ED discharge. Figure 3 shows that this set includes patients with a broad range of severity of illness; the ICU admission rate drops significantly when the ICU is congested for patients from all severity of illness 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 if ICU care and monitoring could substantially reduce this risk, ICU congestion is unlikely to have much effect on the patient's admission to the ICU.[6] 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

We now discuss alternative specifications to show that our main results are robust. Some of the controls for patient severity—*LAPS* and $\hat{P}$*(Mortality)*—were included with piece-wise linear functions to account for their possible nonlinear effects on admission decisions and patient outcomes (see Table 1 for details). We tried different specifications of these functions, and the results were similar.

We tested alternative measures to capture the ICU occupancy level in the ICU admission model. As discussed in §3.2, most of the adjustment to the ICU admission rate occurred when ICU occupancy went above the 95th percentile in our data set, so we defined $ICUBusy_i$ as a binary variable indicating occupancy levels above this threshold. We tested additional specifications in which several hospital characteristics—including hospital size (dividing hospitals into groups by size), the presence of an intermediate care unit at the hospital, and different shifts (7 A.M.–3 P.M., 3 P.M.–11 P.M., and 11 P.M.–7 A.M.)—interact with $ICUBusy_i$ to account for potential heterogeneous effects. In all cases, the estimated average

[6] This gets more complicated by the patients who are denied ICU admission because they are deemed "too sick for ICU treatment" or who have executed do-not-resuscitate orders (e.g., see Reignier et al. 2008).

effect of ICU occupancy on ICU admissions was similar to what was obtained in the main results.

In defining $RecentDischarge_i$ and $RecentAdmission_i$ in the ICU admission model, we used the three-hour and two-hour time windows, respectively. We previously experimented with shorter and longer time windows. For $RecentDischarge_i$, we observed that the effect persisted even when we considered an eight-hour time window (which we considered 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 observed 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 an IV, and the results were similar.

We also examined other factors that may affect the admission decision, such as the clinical severity of the patients currently in the ICU. Because our measures of clinical severity were implemented at the time of hospital admission (not at the time of ICU admission or at any time later in the hospital visit), this measure may not be very accurate, especially as we cannot account for how clinical severity improves or deteriorates during a patient's ICU stay. Nonetheless, when we controlled for the average clinical severity of patients in the ICU, we found that (1) a patient is less likely to be admitted to the ICU when there are many severe patients and that (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 alternative specifications.

In our model, we controlled for the admission month to capture potential seasonal effects and hospital fixed effects to account for variations in practice 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 alone. Thus, we also tried including hospital-month fixed effects and found that although these effects do seem to be statistically significant, accounting for them does not change our main results.

We used the full maximum likelihood estimation to estimate the patient outcome models. While being more efficient, the full maximum likelihood estimation imposes strong parametric assumptions on the distribution of outcomes. We performed some validation of these assumptions for the count variable *HospLOS* and observed overdispersion—the unconditional variance is 24.0, whereas the mean value is 3.9—and no evidence of zero inflation, as only 5.9% had a hospital LOS equal to zero. Hence, the negative binomial model seemed an appropriate model for this outcome.

All of the outcome models included 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: (1) the daily average occupancy of all of the inpatient units in the hospital during the patient's hospital stay, (2) the maximum occupancy level experienced by the patient in an inpatient unit during his or her hospital stay, (3) the average number of inpatients in the hospital during the patient's hospital stay over the maximum possible number of inpatients (without differentiating among different inpatient units), and (4) the average occupancy level of inpatient units at the time that the patient was discharged from the first inpatient unit that he or she visited. All of these alternative definitions gave results that were consistent with what we report for our main specification.

For *Readmit*, recall that we set a time window of two weeks based on discussions with doctors. We tested shorter and longer time windows, and the results for the two-week time window demonstrated higher statistical significance and a greater magnitude.

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. However, patients with 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 *HospLOS* based on the number of nights that a patient stayed in the hospital after being discharged from the ED. We tried defining *HospLOS* as the LOS rounded to the nearest day, and the results were similar. We also estimated the outcome models excluding patients with in-hospital death from 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 that 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 during which 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 the former patients, which is one of the patient outcomes that 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 may confound our estimation of the effect of ICU admission on patient outcomes.

The speed-up effect analyzed by Kc and Terwiesch (2009) is based on cardiac surgery patients, whereas our study is based on ED patients, which is a completely different patient population. We replicated their methodology using our patient sample (see §A.1 of the appendix for the details of this estimation). In particular, we found that we cannot reject the null hypothesis of no speed-up effect in our patient population (p-value of 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 by Kc and Terwiesch (2012). We also utilized our data on elective surgical patients admitted to the ICU. Our analysis of elective surgical patients alone strongly supported rejection of the null hypothesis of no speed-up effect (p-value of 0.001), and we found that a congested ICU reduces the ICU LOS of elective surgical patients by 12% on average. Therefore, our method correctly replicated the results of Kc and Terwiesch (2012) but, at the same time, showed no speed-up effect in the patients admitted to the ICU via the ED. We concluded that this mechanism is not relevant in our patient population and therefore cannot be confounding our main results.

We note that it is interesting to see how the mechanisms used to manage ICU capacity may vary across patient types (ED versus surgical patients). This is also reported by Chen et al. (2013), who show that in contrast to noncardiac patients, severity of illness scores have little impact on the admission decision for cardiac patients.

The second mechanism is ED boarding, which is defined as the time between the decision to admit the patient and when the patient is discharged from the ED and physically moved to the inpatient unit. A congested ICU can extend the ED boarding time; since the ED has less adequate resources to take care of the patient, a longer ED boarding time may have direct implications on the patient outcome.[7] This suggests that ICU congestion may influence patient outcomes through two different mechanisms: (1) the ICU admission decision, which is captured through model (2), and (2) the ED boarding time. Consequently, for ICU congestion to be a valid IV 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 included ED boarding time as a covariate in the outcome models (Equations (1) and (3)), but with special care because the ED boarding time can be endogenous. That is, a severely ill patient who requires urgent care is likely to have a shorter boarding time, and unobservable patient characteristics related to the patient's outcome can influence the ED boarding time. Section A.2 of the appendix provides a detailed description of the econometric model that we developed to handle this endogeneity problem using instrumental variables. This econometric model identifies the effects 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 §A.2 of the appendix) showed that for some outcomes, a longer ED boarding time led to worse patient outcomes, but for others, the effect was not significant. More importantly, the estimated effects of ICU admission were similar to those reported in §4. The main conclusion of this analysis was that our main results regarding the effect of ICU admission on patient outcomes were 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 because this quantification is an essential first step in comparing different ICU admission strategies. To examine how we can utilize the measures that we have just estimated, we consider a parsimonious model of patient flows into the ICU to examine the performance of various admission policies. We specifically 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 observable (i.e., recorded in our data set) 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 by Shmueli et al. (2003). This admission control problem can be viewed as a special case of the stochastic knapsack problem studied by Altman et al. (2001), and we leverage some results from that work to characterize its solution.

[7] California requires a 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.

Consider an ICU with B beds. To focus on the ICU admission decision, we assume that there is ample space in the other inpatient units to care for all patients. We use x to denote 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, or dt, indexed by t, where the periods are sufficiently short so that it is reasonable to assume that at most one patient arrives in a given period. A patient arrives in the ICU with probability λ in each period. Upon arrival, a decision must be made on whether to admit the patient or not. If admitted to the ICU, a patient's LOS is geometrically distributed, with a mean of $1/\mu$. We assume that patient discharge is exogenous, i.e., there is no speed-up in the ICU.[8]

If a patient is routed to the ward, an expected cost of ϕ_c is incurred, where c indexes the customer's class. Without loss of generality, classes are numbered $1, \ldots, C$, so that ϕ_c increases with c. Classes can be interpreted as the clinical severity of the patient, where the benefit of admitting a patient increases with his or her severity of illness. The objective is to choose an admission criterion that minimizes the total expected cost over a finite horizon.

An *admission policy* is defined as a decision rule for choosing whether to admit or reroute an incoming patient, with 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) show that the optimal admission policy is a threshold policy with the following structure: given an occupancy level x, admit a patient if and only if his or her class satisfies $\phi_c \geq \kappa_x$. The values $\{\kappa_1, \ldots, \kappa_B\}$ are referred to as the *optimal thresholds*. It is also shown that the thresholds κ_x increase with 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 on an ICU with $B = 21$ beds, which is the median ICU size in our data set. To simulate ICU admissions, we sample (with replacement) patient characteristics from a hospital whose 95th percentile of occupancy distribution was at 20 beds and whose 99th percentile was at 21 beds. This hospital treated 7,387 ED-medical patients during our study period. Each discrete time period lasts 10 minutes, and patients arrive in the ICU with probability λ, so that, on average, three patients arrive per hour. These parameters have been delicately chosen, so that the simulated setting is consistent with the regime of the hospitals in our study, which admit approximately 10% of the inpatients to the ICU under the current policy. The average patient's 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 measures to be considered: *HospLOS*, *TransferUp*, and *Readmit* (we do not study mortality since the estimates for that outcome are 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 clinical severity class. Each patient i is fully described by a set of observed characteristics X_i (recorded in our data set and described in Table 2) and the "error term" ξ_i, capturing other patient characteristics that are not observed in the data and that are taken into account by the physician when assessing the patient admission decision. We call X_i and ξ_i the observed and unobserved components, respectively, of the patient information. 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 of the information, or $I_i = (X_i, \xi_i)$, and policies that use only the observed component, or $I_i = X_i$.

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

$$\phi_i = E(y_i \mid Admit_i = 0, I_i) - E(y_i \mid Admit_i = 1, I_i), \quad (4)$$

where the expectation is taken with respect to ε_i, the error term in the corresponding outcome model. Here, we explain in detail how we estimate this cost for *Readmit* with information set $I_i = X_i$; the calculations for the other metrics are 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 we use only the observed component, ε_i follows a standard normal distribution. When the unobserved 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 a mean of $\rho\xi_i$ and a variance of $(1 - \rho)^2$. The parameters β_1, β_2, and ρ are the estimates of the readmission outcome reported in §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

[8] As discussed in §1.1, other mechanisms may be used, although we do not find that speed-up is used for the patient group that we study (see §A.1 of the appendix). Via numerical analysis, we found that the qualitative results extend when speed-ups are incorporated.

finite set of patient classes. To achieve this for each patient outcome, when $I_i = X_i$, we first calculate ϕ_i for all 7,387 patients treated in the hospital that we chose to simulate. When $I_i = (X_i, \xi_i)$—i.e., when 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. We then partition patients into 10 groups based on the deciles of this distribution; each patient class has lower and upper bounds on ϕ_i, which define patients that belong to the class. Class c's rerouting cost ϕ_c is set as the average rerouting costs for 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, or 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 indicates 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, 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. In particular, we compare four 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 observable policy* uses the observed component of patient information (i.e., $I_i = X_i$) to assess the expected rerouting cost and to derive the optimal threshold levels of admission. The *optimal full policy* uses the observed and unobserved 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 the 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 needs not be optimal, Figure 3 presents several patterns that are consistent with a threshold policy. First, admission rates tend to increase as clinical 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.[9]

We restrict the hospital that 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 that 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 instance, $f(x; \kappa)$ can change at every possible occupancy level, or it can change only once, such as when the ICU occupancy is 20 and above. For each model (5) with a 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 observable 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 or her class ϕ_c, which reduces the possible values of each threshold to $\{\phi_1 \ldots \phi_{10}\}$. This provides an upper bound on the performance of the optimal policies.

[9] Consider two patient classes, high (H) or low (L) severity of illness, and assume that patient severity for class $j \in \{L, H\}$ follows 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 that of the H group decreases more. These results are not specific to the normal distribution assumption for severity of illness; they hold for any distribution with a density function decreasing at the threshold κ (i.e., $f'(x) < 0$ for $x > \kappa$).

Table 5 Simulation Results of Alternative ICU Admission Control Policies

	Estimated current policy		For each outcome	
	BASE–21 beds	22 beds	Optimal observable	Optimal full
No. of readmissions	**2,550.4** (1.55)	−3.7 (0.06)	−26.6 (0.44)	−35.3 (0.41)
No. of transfer-ups	**762.9** (0.88)	−5.9 (0.08)	14.8 (0.50)	−38.6 (0.38)
Hospital LOS (years)	**245.6** (0.15)	−0.4 (0.01)	−2.0 (0.13)	−9.0 (0.12)
Total hospital LOS (years)	**272.9** (0.15)	−0.4 (0.01)	−2.2 (0.13)	−9.2 (0.12)
(Estimated savings in dollars)		(−$0.4 m)	(−$1.9 m)	(−$8.1 m)

Notes. 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. Standard errors in parentheses.

The estimated current policy may perform worse than the optimal observable policy for several reasons. First, the admission decision under the optimal observable policy is based on the rerouting cost ϕ_i, whereas in the estimated current policy described by Equation (5), the left-hand side of the inequality is not necessarily equal to ϕ_i. That is, the estimated current policy may not be appropriately weighting the observed metrics X_i. This is because the policy estimates how the physicians at the hospital weigh the available information to make a decision, which may be discretionary. Second, the threshold adjustment function $f(x;\kappa)$ may not set optimal threshold levels that properly account for the opportunity cost of using up a bed, which the optimized policy does. However, the estimated current policy has a richer information set than the optimal observable policy, so it is not known a priori which will perform better. Because the optimal full policy utilizes the same information as the estimated current policy, accurately weights both the observed and the unobserved information ($I_i = (X_i, \xi_i)$), and optimizes the thresholds, the optimal full policy will perform better than the estimated current policy.

5.4. Results and Discussion

Table 5 summarizes the simulated patient outcomes—*HospLOS*, *TransferUp*, and *Readmit*—under the alternative policies that we consider. Noting that the ICU admission decision is an inherently multiobjective problem, we also consider a combined outcome that considers the impact of ICU admission on total hospital days for the current inpatient stay as well as any potential subsequent hospital stay due to readmission. In particular, we convert each readmission into an 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.[10] Finally, for comparison purposes, we convert hospital days into dollar amounts by utilizing an estimate of $2,419 per hospital day, as given by Kaiser Family Foundation (2012).

The column labeled "BASE–21 beds" lists the performance of the estimated current policy in a 21-bed ICU. Under the current policy (estimated as described above), on average, there were 2,550 hospital readmissions, 762.9 transfer-ups to the ICU from the general wards, and a total of 245.6 hospital bed years spent by patients 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 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 an expense of $3,164 per ICU day (Aloe et al. 2009). Note that the cost of an extra ICU bed ($0.8 m) is double the estimated savings achieved by reducing readmissions and hospital LOS ($0.4 m). 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 column labeled "Optimal observable" provides the performance of the optimal policy based on observable measures alone ($I_i = X_i$). Each row corresponds to a different policy optimized to minimize the corresponding outcome. Because the optimal observable policy optimizes the admission thresholds while also utilizing the direct relationship between the available information (X_i) and patient outcomes, it can sometimes perform better than the estimated current policy. This is the case when we use the optimal occupancy-dependent thresholds derived from the cost function for readmissions ($\phi_c^{Readmit}$) and hospital LOS ($\phi_c^{HospLOS}$); we observe 26.6 fewer readmissions and two fewer years of hospital LOS on average than for the estimated current policy. However, the estimated current policy may outperform the optimal

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

observable policy because it utilizes more information (ξ_i) in addition to X_i, which appears to be useful in predicting patient outcomes (as indicated by the correlation coefficient ρ; see Table 4). Indeed, the optimal observable policy aimed at minimizing *TransferUp* has *more* transfer-ups (15 more on average) compared to the estimated current policy. That said, this is not a systematic effect; we find that the optimal observable policy can outperform the current policy across all patient outcomes when we examine other hospitals. These results suggest that the unobserved information can be useful but that optimizing the admission decision based solely on observed criteria can often result in better patient outcomes.

Finally, we further explore the benefits of incorporating unobserved information in the admission decision. The column labeled "Optimal full" uses both the observed and the unobserved information ($I_i = (X_i, \xi_i)$) and further optimizes the admission thresholds. We see that by optimizing the thresholds, patient outcomes can be universally improved compared to those resulting from the estimated current policy. The difference between "Optimal full" and "Optimal observable" measures the value of capturing currently unobserved metrics in the admission decision (i.e., incorporating ξ_i in the information set) in terms of improving patient outcomes. We can see that the benefit 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, forming 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 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 for various patient outcomes. Furthermore, we provided a detailed characterization of the optimal ICU admission policy based on observed measures of clinical severity and showed how to compute these policies for different patient outcomes using empirical data, dynamic programming, and simulation methods. Via simulation experiments, we were able to compare the performance of admission policies based purely on observed 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 observed metrics alone can outperform current hospital policies. For other outcome measures, we found that the unobserved criteria used by doctors are useful and can help to improve system performance relative to a decision based solely on observed criteria. We believe that 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 to address related questions. We believe that the present work can be easily applied to study capacity allocation and the impact of the occupancy level on available resources in many other healthcare settings. For instance, the level of care can differ among different ICU units. In particular, rather than having only one type of ICU, many hospitals have specialized ICUs, such as cardiac, surgical, and medical ICUs, so the nurse-to-patient ratios and levels of treatment might differ. However, resources 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 in these different types of ICUs is performed 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 data set 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 nonpaying patients. Second, in §5.1, we introduce a stylized model of ICU admission, with a constant arrival rate for inpatients and a constant departure rate for ICU patients. We believe that it serves its role of giving us insights into 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 clinical 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, Yom-Tov and Mandelbaum 2014). Third, a limitation of the instrumental variable estimation strategy is that it provides an estimate of the average effect of ICU admission over the subset of patients whose admission decision depends on ICU occupancy (known as the latent average treatment effect, or LATE). This excludes two sets of patients: (1) Patients who are never admitted to the ICU, even if there is ample space in the ICU. This set of patients is probably the one that benefits the least from ICU treatment. (2) Patients who are severely ill enough to be admitted to the ICU no matter how busy it is. These are usually the most severe patients, who include those patients with a higher risk of dying. Hence, the effect estimated through our IV approach 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 §§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 and patient outcomes might prove useful in developing decision support tools for ICU admission, discharge, and capacity planning.

Acknowledgments

The authors gratefully acknowledge the editors and reviewers for their many helpful suggestions and comments, which have greatly improved this paper. The authors also thank the participants at numerous seminars; the Wharton Empirical Workshop in Operations Management; and the INFORMS, MSOM, and POMS meetings for very helpful comments. The authors thank Marla Gardner and John Greene for their help in preparing the ICU data, along with the staff in the Division of Research and hospitals in Kaiser Permanente Northern California for their time and invaluable contributions to this research. Song-Hee Kim acknowledges support from the INFORMS MSOM Society in the form of a best student paper award for a preliminary version of this work. The research by Carri W. Chan was partially supported by the National Science Foundation [CAREER Award CMMI-1350059]. Marcelo Olivares thanks the Instituto Sistemas Complejos de Ingenieria for financial support.

Appendix. Accounting for Speed-Up and ED Boarding Effects

This appendix provides a detailed description of the econometric models used in the analysis in §4.2.

Table A.1 Estimation Results of Model (6)

	Busy coefficient (standard error)	No. of observations	R^2
ED, medical	−0.02 (0.03)	10,521	0.16
Non-ED, surgical	−0.13** (0.04)	4,524	0.14

**$p < 0.01$.

A.1. Speed-Up in the ICU

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

Define $FirstICU\ LOS_i$ as the ICU LOS during patient i's first ICU visit and $BUSY_i$ as the bed utilization of the ICU at the time that patient i was discharged from this ICU visit. Because our data set does not include information on the number of scheduled arrivals, our definition of $BUSY_i$ is not the same as that of Kc and Terwiesch (2012). Instead, we let $BUSY_i$ be 1 if the number of existing ICU patients at the time that patient i is discharged from the ICU exceeds the 95th percentile of occupancy.[11] We estimate the effect of ICU occupancy on ICU LOS through the following regression:

$$\log(FirstICU\ LOS_i) = \gamma BUSY_i + \beta X_i + u_i, \qquad (6)$$

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, reflecting a speed-up effect.

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

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 affects ED boarding time but is unrelated to the clinical severity of the patient. The instrument that we use is the "average level of bed occupancy of the inpatient unit that the patient goes to after the ED," labeled *FirstInpatientOcc*, whose average is taken during the time that the patient is boarding in the ED. The logic is similar to that of our *ICUBusy* instrument: if the patient was routed to an inpatient unit but this unit was busy

[11] 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, or the Weibull and Cox proportional hazard models, with *BUSY* included as either time invariant or varying over time, and the results were consistent.

Table A.2 Estimation Results of the Patient Outcome Model Including ED Boarding Time as an Endogenous Covariate

Outcome	ICU admission	Log(*EDBoard*)
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.
***$p < 0.001$.

when the patient was in the ED, 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 to ED discharge, whereas *FirstInpatientOcc* measures the occupancy of ICU or the ward, depending on where the patient is routed 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; in fact, 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 are available from the authors upon request.

The estimation of the model is 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(*EDBoard*) as the dependent variable and the IVs and controls as covariates, and (2) we then calculate the residuals of this regression and include the residuals and log(*EDBoard*) as additional covariates in the outcome model. See Wooldridge (2010) for more details on the control function approach. Table A.2 reports the estimated coefficients for ICU admission and log(*EDBoard*) for the different outcome models.

References

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

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

Altman E, Jiménez T, Koole G (2001) On optimal call admission control in resource-sharing system. *IEEE Trans. Comm.* 49(9): 1659–1668.

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

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

Armony M, Israelit S, Mandelbaum A, Marmor Y, Tseytlin Y, Yom-Tov G (2010) Patient flow in hospitals: A data-based queueing-science perspective. Working paper, New York University, New York.

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

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

Batt R, Terwiesch C (2014) Doctors under load: An empirical study of state-dependent service times in emergency care. Working paper, University of Wisconsin–Madison, Madison.

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

Brilli R, Spevetz A, Branson R, Campbell G, Cohen H, Dasta J, Harvey M, 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, Mattes M, Burton S (1995) Reducing intensive care unit length of stay: A stepdown unit for first-day heart surgery patients. *J. Nursing Administration* 25(12):29–30.

Cameron A, Trivedi P (1986) Econometric models based on count data. Comparisons and applications of some estimators and tests. *J. Appl. Econometrics* 1(1):29–53.

Cameron A, Trivedi P (1998) *Regression Analysis of Count Data* (Cambridge University Press, Cambridge, UK).

Chalfin D, Trzeciak S, Likourezos A, Baumann B, Dellinger R (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, Kennedy EH, Sales A, Hofer T (2013) Use of health IT for higher-value critical care. *New England J. Medicine* 368(7): 594–597.

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

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

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

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

Escobar G, Greene J, Gardner M, Marelich G, Quick B, Kipnis P (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). *J. Hospital Medicine* 6(2):74–80.

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

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

Fisher E, Wennberg D, Stukel T, Gottlieb D (2004) Variations in the longitudinal efficiency of academic medical centers. *Health Affairs* 2004:VAR-19–VAR-32.

Franklin C, Rackow E, Mamdani B, Burke G, Weil M (1990) Triage considerations in medical intensive care. *Arch. Internal Medicine* 150(7):1455–1459.

Glasserman P, Yao D (1994) Monotone optimal control of permutable GSMPS. *Math. Oper. Res.* 19(2):449–476.

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

Green L, Savin S, Savva N (2013) Nursevendor problem: Personnel staffing in the presence of endogenous absenteeism. *Management Sci.* 59(10):2237–2256.

Halpern NA, Bettes L, Greenstein R (1994) Federal and nationwide intensive care units and healthcare costs: 1986–1992. *Critical Care Medicine* 22(12):2001–2007.

Halpern N, Pastores S, Thaler H, Greenstein R (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, Corbella D, Minelli C, Mills GH, Artigas A, Edbooke DL, Pezzi A, 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, MI).

Jaeker JB, Tucker AL (2013) An empirical study of the spillover effects of workload on patient length of stay. Working paper, Boston University, Boston.

Kaiser Family Foundation (2012) Hospital adjusted expenses per inpatient day: California in year 2009. Accessed November 14, 2014, http://kaiserf.am/1ukoJPW.

Kaplan R, Porter M (2011) How to solve the cost crisis in health care. *Harvard Bus. Rev.* 89(9):46–52.

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

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

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

Kuntz L, Mennicken R, Scholtes S (2014) Stress on the ward: Evidence of safety tipping points in hospitals. *Management Sci.*, ePub ahead of print May 19, 2014, http://dx.doi.org/10.1287/mnsc.2014.1917.

Liu V, Kipnis P, Gould M, Escobar G (2010) Length of stay predictions: Improvements through the use of automated laboratory and comorbidity variables. *Medical Care* 48(8):739–744.

Louriz M, Abidi K, Akkaoui M, Madani N, Chater K, Belayachi J, Dendane T, Zeggwagh AA, Abouqal R (2012) Determinants and outcomes associated with decisions to deny or to delay intensive care unit admission in Morocco. *Intensive Care Medicine* 38(5):830–837.

Luyt C, Combes A, Aegerter P, Guidet B, Trouillet J, Gibert C, Chastre J (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 Sci.* 16(3):234–245.

Mullan F (2004) Wrestling with variation: An interview with Jack Wennberg. *Health Affairs* 2004:VAR-73–VAR-80.

O'Connor A, Llewellyn-Thomas H, Flood A (2004) Modifying unwarranted variations in health care: Shared decision making using patient decision aids. *Health Affairs* 2004:VAR-63–VAR-72.

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

Papastavrou JD, Rajagopalan S, Kleywegt AJ (1996) The dynamic and stochastic knapsack problem with deadlines. *Management Sci.* 42(12):1706–1718.

Phillips R, Şimşek AS, van Ryzin G (2015) The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Sci.* Forthcoming.

Pronovost P, Needham D, Waters H, Birkmeyer C, Calinawan J, Birkmeyer J, Dorman T (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. *Sociol. Methodology* 25:111–164.

Rainey TG, Raphaely RC, Chalfin DB, Parks LK, Fitzpatrick MA, Connor SJ, Johanso WL, Kalowes PG (1994) Joint Position Statement: Essential provisions for critical care in health system reform. *Critical Care Medicine* 22(12):2017–2019.

Ramdas K, Saleh K, Stern S, Liu H (2012) New joints more hip? Learning in the use of new components. Working paper, London Business School, London.

Reignier J, Dumont R, Katsahian S, Martin-Lefevre L, Renard B, Fiancette M, Lebert C, Clementi E, Bontemps F (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, Jegers M (2012) Monitoring costs in the ICU: A search for a pertinent methodology. *Acta Anaesthesiologica Scandinavica* 56(9):1104–1113.

Robert R, Reignier J, Tournoux-Facon C, Boulain T, Lesieur O, Gissot V, Souday V, Hamrouni M, Chapon C, Gouello J (2012) Refusal of intensive care unit admission due to a full unit impact on mortality. *Amer. J. Respiratory Critical Care Medicine* 185(10):1081–1087.

Shi P, Chou MC, Dai JG, Ding D, Sim J (2015) Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Sci.* Forthcoming.

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

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

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

Simchen E, Sprung C, Galai N, Zitser-Gurevich Y, Bar-Lavi Y, Gurman G, Klein M, 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, Clancy M, Goldfrad C, Rowan K (2005) Admissions to intensive care units from emergency departments: A descriptive study. *Emergency Medicine J.* 22(6):423–428.

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

Sprung C, Geber D, Eidelman L, Baras M, Pizov R, Nimrod A, Oppenheim A, Epstein L, Cotev S (1999) Evaluation of triage decisions for intensive care admission. *Critical Care Medicine* 27(6):1073–1079.

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

Strauss M, LoGerfo J, Yeltatzie J, Temkin N, Hudson L (1986) Rationing of intensive care unit services. *J. Amer. Medical Assoc.* 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(3):633–638.

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

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

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

Weber R, Stidham S Jr (1987) Optimal control of service rates in networks of queues. *Adv. Appl. Probab.* 19(1):202–218.

Weinstein JN, Bronner KK, Morgan TS, Wennberg JE (2004) Trends and geographic variations in major surgery for degenerative diseases of the hip, knee, and spine. *Health Affairs* 2004:VAR-81–VAR-89.

Wennberg JE, Fisher ES, Skinner JS (2002) Geography and the debate over Medicare reform. *Health Affairs* 2002:W96–W114.

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

Yom-Tov G, Mandelbaum A (2014) Erlang-*R*: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing Service Oper. Management* 16(2):283–299.

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