Assessing the Impact of Service Level when Customer Needs are Uncertain: An Empirical Investigation of Hospital Step-Down Units

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

Linda V. Green
Decision, Risk, and Operations, Columbia Business School, lvg1@columbia.edu

Suparerk Lekwijit
Operations, Information and Decisions, The Wharton School, slek@wharton.upenn.edu

Lijian Lu
Decision, Risk, and Operations, Columbia Business School, ll2755@columbia.edu

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

Many service systems have servers with different capabilities and customers with varying needs. One common way this occurs is when servers are hierarchical in their skills or in the level of service they can provide. Much of the literature studying such systems relies on an understanding of the relative costs and benefits associated with serving different customer types by the different levels of service. In this work, we focus on estimating these costs and benefits in a complex healthcare setting where the major differentiation among server types is the intensity of service provided. Step Down Units (SDUs) were initially introduced in hospitals to provide an intermediate level of care for semi-critically ill patients who are not sick enough to require intensive care but not stable enough to be treated in the general medical/surgical ward. One complicating factor is that the needs of customers is sometimes uncertain – specifically, it is difficult to know a priori which level of care a particular patient needs. Using data from 10 hospitals from a single hospital network, we take a data-driven approach to classify patients based on severity and empirically estimate the clinical and operational outcomes associated with routing these patients to the SDU. Our findings suggest that an SDU may be a cost-effective way to treat patients when used for patients who are post-ICU. However, the impact of SDU care is more nuanced for patients admitted from the emergency department (ED) and may result in increased mortality risk and hospital LOS for patients who should be treated in the ICU. Our results imply that more study is needed when using SDU care this way.

Key words: healthcare, empirical operations management, congestion, quality of service

1. Introduction

Hospitals are responsible for the largest component of national health care expenditures and are therefore under pressure from government and private payers to become more cost efficient (Centers for Medicare & Medicaid Services 2016). Traditionally, inpatient care at hospitals had been defined by two levels of care: Intensive care units (ICUs) and general medical/surgical wards (wards). With one nurse per one or two patients, ICUs provide the highest level of care and are very costly to operate, with annual costs in the

U.S. between $121 and $263 billion (i.e., 17.4%-39% of total hospital costs (Coopersmith et al. 2012)). In an effort to mitigate critical care costs, Step-down units (SDUs), sometimes called transitional care or intermediate care units, have been used to provide an intermediate, third level of care for semi-critically ill patients who are not severe enough to require intensive care but not stable enough to be treated in the ward. SDUs typically have one nurse per three to four patients and are generally less expensive to operate than ICUs primarily due to lower nurse-to-patient ratios. On the other hand, SDUs are more expensive than general wards where there are, generally, about 6 patients per nurse. With the use of SDUs becoming more widespread, it is of growing importance for hospital administrators and healthcare providers to have a better understanding of the benefits and best practices associated with using this intermediate level of care.

At a conceptual level, the hospital and ICU/SDU/ward system can be thought of as a general service system with three levels of service and heterogenous customers. The levels are nested, in the sense that the lowest level (ward) has the least capabilities and can only provide service to a subset of customers (patients); the second level (SDU) can provide service to the lowest level customers plus additional customers with greater needs; and the highest level (ICU) can provide service (theoretically) to *all* customers. Due to higher staffing levels as well as specialized equipment, higher levels of care are more costly to provide. It is of interest to understand whether such a structure is beneficial and, if so, how to best utilize the different levels of service. This is more challenging when there is uncertainty concerning which customers are best served at each level, making it very difficult to evaluate the cost-benefit tradeoffs. The ultimate goal is to understand effective management of such a service system, including capacity management of each level of service, when and how to route customers, as well as how to classify customers and identify their needs for the different levels of service.

There has been a considerable amount of research into capacity management of service systems and the development of routing policies to different service types (e.g. Wallace and Whitt (2005), Gurvich et al. (2008) among many others). Such issues have been studied in various service settings including call-centers (e.g. Gans et al. (2003)), hospitals (e.g. Armony et al. (2017), Best et al. (2015)), cloud-computing (e.g. Maguluri et al. (2012)), among many others. A common assumption in these works is a general understanding of the relative costs and benefits associated with different customer groups receiving service from the various server types. Yet, in some contexts these relative costs and benefits may not be known. Specifically, the needs of customers may be uncertain prior to starting service. In this work, our goal is to gain an understanding of how best to use different levels of service to serve customers with uncertain needs by empirically examining how different customer groups are impacted by being served at differing levels. We examine this question in a healthcare context–the SDU.

There is a lack of consensus in the medical community surrounding the use of SDUs as well as a lack of substantive evidence concerning their effectiveness. Still, many hospitals have SDUs and others are considering introducing these units. Even within a single hospital, the use of SDUs is generally not standardized.

Therefore, it is very important to understand their value and how they can best be used. This paper examines whether or not SDUs are associated with improved operational and/or clinical outcomes for different types of patients. In this context, the aforementioned costs and benefits are not necessarily financial in nature. For instance, they can correspond to deteriorations or improvements in patient outcomes. Such analysis can provide insights into how the nested levels of care structure could be used to treat patients with differentiated service requirements and potentially lower hospital operating costs without sacrificing patient outcomes. Given the increasing pressures for hospitals to reduce costs and improve quality, such insights can be very valuable to hospital administrators. More broadly, this analysis may also provide insights into the analysis and management of other service systems with different levels of care (e.g., call-centers).

To the best of our knowledge, our work is the first to conduct a multi-hospital study to empirically examine the role of an SDU for patients who are discharged from the ICU as well as those who are admitted from the Emergency Department (ED). Our analyses are based on recent data from Kaiser Permanente Northern California, an integrated health care delivery system serving 3.6 million members that operates 21 hospitals, some of which do and some of which do not have SDUs. The cohort and type of data we employ have been described in previous studies (see Escobar et al. (2013), Kim et al. (2015) among others). Our data source is based on nearly 170,000 hospitalizations in a total of 10 hospitals over a course of one and half years. Each of the 10 hospitals in our study has an ICU and SDU, though the number of beds in each of the units varies across hospitals.

There are a number of challenges which arise when trying to understand the impact of SDU care on patient outcomes. One challenge is that there are limited studies regarding its efficacy and, more specifically, which patients can be safely admitted to the SDU (Nasraway et al. 1998). While there is some evidence that some ICU patients who are at low risk of needing life support could be given less intensive care in an SDU with no impact on outcomes (e.g. Zimmerman et al. (1995)), there is also evidence that some critical care patients who are treated in SDUs or general wards instead of the ICU are worse off (e.g. Simchen et al. (2004)). As such, it seems that there are patients who may benefit from being cared for in an SDU rather than in a general ward, while others who are treated in an SDU rather than an ICU may suffer adverse consequences. An important empirical challenge is to be able to classify patients in order to accurately assess the impact of SDU admission on patient outcomes. To that end, we initially segregate patients who are candidates for SDU care into two broad groups: those who are discharged from the ICU and those who are admitted to an inpatient unit from the ED. Taking a data-driven approach, we then stratify patients from the ED into high and low severity groups.

In developing an understanding of SDUs, we face an important estimation challenge. The SDU admission decision may be affected by health factors which are known to the physician at the time of the decision, but are unobservable in the data. For instance, a patient's physical appearance (i.e. whether he/she appears ashen or pale) may provide evidence of early shock. Thus, a physician may determine that, despite relatively

stable vital signs and lab scores, a patient who is pale and sweating will benefit from SDU care relative to being sent to the general medical ward. But because the patient is more critical than the average ward patient, he/she is also more likely to have worse outcomes. Similarly, it may be more appropriate to admit a patient to the ICU if he is cognitively impaired and not lucid. Thus, patients who are admitted to the SDU instead of the ICU may be healthier by unobservable measures. Ignoring this potential endogeneity could result in biased estimates. To address this challenge, we utilize an instrumental variable approach to identify the desired effects.

Our empirical findings suggest that SDU care is associated with substantial improvements in various patient outcomes for patients discharged from the ICU as well as low severity patients being admitted from the ED. However, we find that SDU admission is associated with worse outcomes for high severity patients coming from the ED. Our results suggest that when SDUs are used as originally intended, as intermediary units for post-ICU care, they may result in improved outcomes relative to ward care. However, if hospital administrators wish to expand the use of SDUs beyond post-ICU care, it is important to be able to classify which patients should or should not be treated in the SDU. More generally, our findings highlight the importance of being able to accurately classify customers and to quantify the (dis)utility associated with different service capabilities when considering routing decisions.

The rest of the paper is organized as follows. We conclude this section with a brief summary of related papers in the literature. In Section 2, we introduce our study setting and describe our data, including the two patient cohorts we study. In Section 3, we describe our econometric model for our first cohort of patients–those being discharged from the ICU. The estimation results for this cohort is provided in Section 4. Section 5 describes how we partition patients who are admitted from the ED into high and low severity patients and then discusses the econometric model we use for these patients. Results for these patient types are provided in Section 6. Section 7 provides concluding remarks as well as discussions for future research.

1.1. Literature Review

Our work is related to existing literature in both the operations management and medical communities. Within the operations literature, our work is related to three streams of research: 1) management of general service systems, 2) management of healthcare operations, and 3) empirical analysis of healthcare operations.

There has been a large body of literature examining how to route customers to servers with different skill sets (see the survey article Gans et al. (2003) and the references therein). Research in this area has considered customer prioritization (e.g., Mandelbaum and Stolyar (2004) and Gurvich and Whitt (2009)), customer routing (e.g., Bell and Williams (2001) and Tezcan and Dai (2010)), and staffing (e.g., Wallace and Whitt (2005) and Gurvich and Whitt (2010)). Additionally, there have been a number of works studying service settings with different levels of service. In call-centers, one can consider human servers as providing more intense and costly service than chat-room or automated response systems (e.g. Gans et al. (2003),

Tezcan and Behzad (2012), Luo and Zhang (2013), Tezcan and Zhang (2014)). Maglaras and Zeevi (2005) considers pricing, admission control, and the design of a mechanism to relay congestion information in a system where servers can provide either a guaranteed service rate or a best-effort service rate. In call center settings, VIP customers often require a higher level of service than the typical customer, raising questions on how to route customers to various servers (e.g. Gans et al. (2003)). Such features also arise in healthcare settings including the SDU we study in this paper. Chan et al. (2013) considers how to prioritize burn-injured patients for treatment in hospitals with burn-units which provide the specialized, intense therapies (e.g. skin grafting surgeries) required for severely burned patients versus other hospitals with less intense treatment capabilities.

The nested structure of the different levels of care we examine in the hospital setting bears similarities to the gate-keeper literature (e.g. Shumsky and Pinker (2003), Hasija et al. (2005), Lee et al. (2012)) where the specialist is able to provide services the gate-keeper is not able to. However, in contrast to this literature, in our setting, the lowest level of service does not make the decision to route customers to higher levels of service as in the gate-keeper literature. The nested structure is also related to the classic toll-booth problem considered in Edie (1954) as certain lanes can serve all types of vehicles, while others can only serve a subset of them (e.g. Green (1985)). Rather than having a central planner making routing decisions for customers whose needs may be unknown to him/her, in the toll-booth problem, the customers know their needs and self-direct to servers.

There are a number of papers which utilize stochastic modeling and queueing approaches to study resource allocation in hospital settings (e.g. Mandelbaum et al. (2012), Shi et al. (2014), Huang et al. (2015), Huh et al. (2013), Barz and Rajaram (2015)). In all of these works, the focus is on admitting patients with heterogeneous needs to different units *within the same level of care*. That is, servers are interchangeable. In contrast, our work considers the impact of admitting patients to *different levels of care*. In doing so, we are able to capture heterogeneous service requirements of customers (patients) as well as the various levels of service (care).

There has been a growing body of work in healthcare operations management using mathematical models to manage heterogeneous patients in systems with differentiated server types. Best et al. (2015) examines how to determine the amount of flexibility allowed in hospital wings in order to minimize costs associated with lack of access to care. Dai and Shi (2017) uses an approximate dynamic programming approach to determine how to allocate patients to primary and non-primary units. Armony et al. (2017) uses fluid and diffusion models to determine allocation among expensive resources (ICU beds) that can be used to treat all patient types rather than cheaper resources (SDU beds) that can only treat a subset of patients. An underlying assumption in all of these works is that, in addition to a patient's type, the relative costs (i.e. degradation of patient outcomes) to treat that patient in different types of units are known. Our aim is to provide a

framework to classify patients as well as to provide rigorous, quantitative estimates of the outcomes for patients treated in an SDU.

As we take an empirical approach to quantify the costs/benefits of treating patients in the SDU, our work is closely related to papers in the empirical operations management literature, especially those focused on healthcare settings. Jerath et al. (2015) empirically estimates how customers' service needs impact their preferences to use different types of service channels when interacting with a health insurance call center. In hospital settings, Stowell et al. (2013), Kim et al. (2015), Kuntz et al. (2016) take an empirical approach to explore the impact of admitting patients to different types of hospital units on patient outcomes. While these works highlight the undesirability of 'off-placement', Wang et al. (2016) explicitly considers how information on hospital (server) quality needs to be patient-specific. As such, while hospitals are capable of treating all different types of patients, which is similar to the SDU, the costs/benefits associated with being treated at a specific hospital are quite varied. Unfortunately, it is not always possible to treat patients at the most appropriate hospital or hospital unit. Congestion is a common reason for this lack of access to care. There have been a number of studies examining the impact of congestion and lack of access to care on patient outcomes (e.g. Kc and Terwiesch (2012), Kuntz et al. (2015), Berry Jaeker and Tucker (2016), among others). Batt and Terwiesch (2017) and Freeman et al. (2016) empirically examine how less or more skilled servers can be used to treat some patients during congested periods. In a similar vein, we examine how treating different patient types in an SDU, which is a higher level of care than the ward, but lower than the ICU, impacts their outcomes.

There is a lack of consensus within the medical community about the role of the SDU. Those who advocate the use of SDUs see them as an alternative to either maintaining larger ICUs or jeopardizing patient care due to premature, demand-driven, discharge of patients from ICUs to general care units. As the name suggests, the initial role of SDUs was to serve as a transition for patients after being discharged from the ICU. In practice, SDUs are often used to treat other patients, for example, those who might have gone to an ICU but were blocked because the ICU was full. In general, the use of SDUs has evolved without substantial evidence as to their benefits and what their role should be. On one hand, some studies argue that SDUs are a cost-effective approach to treat patients by providing a safe and less expensive environment for patients who are not quite sick enough to require treatment in the ICU, but not quite stable enough to be treated in the ward. Without an SDU, most of these patients end up being cared for in the ICU. Byrick et al. (1986) suggests that the use of the SDU could alleviate ICU congestion by reducing ICU length-of-stay (LOS) without increasing mortality rates. This reduction is possible because patients do not have to reach as high a level of stability to be discharged to an SDU rather than to a general medical-surgical ward. Other studies that have shown the cost-effectiveness of an SDU include Harding (2009), Stacy (2011), and Tosteson et al. (1996). On the other hand, a survey of studies on SDUs raises doubts about these benefits and argues that there is not enough evidence of cost-effectiveness (Keenan et al. 1998). While we do not explicitly consider

the cost-effectiveness of SDUs (due to lack of detailed financial data), our study provides some insight into these questions by providing rigorous and robust estimates to the effectiveness of SDUs for patients of varying types. At a high-level, one can project ordinal cost estimates due to the lower (higher) staffing levels in the SDU versus the ICU (ward). From a methodological standpoint, our study differentiates itself in that the majority of these studies are conducted exclusively within a single hospital, whereas our study utilizes data from 10 different hospitals. Additionally, rather than conducting a before-and-after study, which may be limited by the inability to control for temporal changes such as staffing changes or closures of nearby hospitals, we utilize an instrumental variable approach to identify the impact of different care pathways (going to the SDU versus ward following ICU discharge as well as going to the SDU versus ward or ICU upon hospital admission from the ED). Our multi-center study provides compelling evidence that there are some patients for whom SDU care is associated with improved clinical outcomes, while there are others for whom SDU care is associated with worse clinical outcomes. As such, our results suggest that it would be of value for the medical community to focus more attention on developing an understanding of which patients would or would not benefit from SDU care at hospitals of varying patient mix and resource availability. More broadly, our results suggest that one must be prudent when introducing multiple levels of service in service systems with highly heterogeneous customers as there can be substantial variation in the costs and benefits associated with (incorrectly) routing customers to these servers.

Our estimation approach utilizes an instrumental variable which is based on an operational measure–congestion in an inpatient unit–as has been done in Kim et al. (2015) and Kc and Terwiesch (2012), among others. While the general methodology is similar, the question we are considering is wholly different. The aforementioned works focus on the ICU, while our focus is on the SDU. From an operational standpoint, it is of value to develop an understanding of how servers with lower costs due to lower staffing levels (SDUs) may be used to serve heterogeneous customers. Additionally, from the viewpoint of clinicians and hospital administrators, these units are fundamentally different in their use and role. As a customer's type and, subsequently, his service requirements are not always observable to managers of the service system, it can be challenging to estimate the costs and benefits associated with being served by particular server types. This challenge arises in the SDU setting because they serve as the site of intermediate care between the ICU and the ward; that is, there are risks of adverse consequences in admitting a patient to the SDU who actually needs ICU care, as well as benefits to admitting patients who might be too sick for the ward. As such, we first take a data-driven approach to help classify customers (patients) before estimating the impact of SDU care on patient outcomes.

2. Setting and Data

We utilize patient data from 10 hospitals from Kaiser Permanente Northern California[1], containing 165,948 hospitalizations over a course of one and a half years. We note that even within the Kaiser Permanente Northern California system, there is no consensus on how to use SDUs. Thus, some hospitals have SDUs, while others do not.

Our data contains operational and patient level information. Operational level information includes every unit to which a patient is admitted during his hospital stay along with the date and time of admission and discharge for each unit. Our objective in this work is to understand the impact of service by flexible servers (SDU care) on heterogeneous customer (patient) types. Table 1 summarizes the distribution of where patients come from immediately preceding their SDU visit. Over 78% of patients in the SDU come from the ED or ICU. As such, our analysis will focus on these two patient cohorts. Specifically, we will focus on how transfer to the SDU impacts patients who are admitted to an inpatient unit from the ED as well as patients who are discharged from the ICU to lower levels of care. Figure 1 depicts these two transfer decisions that will be the heart of our empirical investigation. Given the contrasting routes to the SDU of these patients, it is reasonable to assume the impact of SDU care may differ substantially and our objective is to rigorously estimate the treatment effect of SDU care for these heterogeneous patient types.

Table 1 Distribution of Units Preceding the SDU

Unit Preceding SDU	Percentage
ED	60.93%
ICU	17.11%
Ward	13.88%
Post-Anesthesia Recovery Unit (PAR)	4.25%
Operating Room (OR)	3.58%
Other/Unknown	0.25%

For each inpatient unit in each hospital, we use these patient flow data to derive hourly occupancy levels and we define its capacity as the maximum occupancy level over the time horizon of our study. Table 2 summarizes the capacity for each of the different levels of inpatient care in each hospital. While each level of care may have further divisions based on specific services, e.g. medical versus surgical ICU, clinicians and administrators at the study hospitals indicate that it is widely accepted practice at their hospitals to consider the boundaries as somewhat fluid in the sense that if a medical service patient requires ICU care, but there are no medical ICU beds available, he will likely be cared for in the surgical ICU. We observe substantial heterogeneity across these hospitals; the SDU capacity varies from 11 to 32 beds and the number of ICU beds in a given hospital ranges from one half to twice the number in the SDU.

[1] This project was approved by the Kaiser Permanente Northern California Institutional Review Board for the Protection of Human Subjects, which has jurisdiction over all study hospitals, and the Columbia University Institutional Review Board for the Protection of Human Subjects.

Figure 1 Types of Admission Decisions

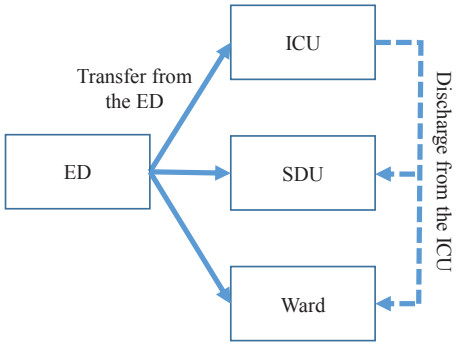


Table 2 Capacity of Various Inpatient Units in terms of number of beds

Hosp	ICU	SDU	Ward
1	11	24	61
2	11	25	76
3	16	14	77
4	16	19	76
5	16	24	78
6	23	19	124
7	24	20	145
8	26	27	110
9	31	11	188
10	32	32	100

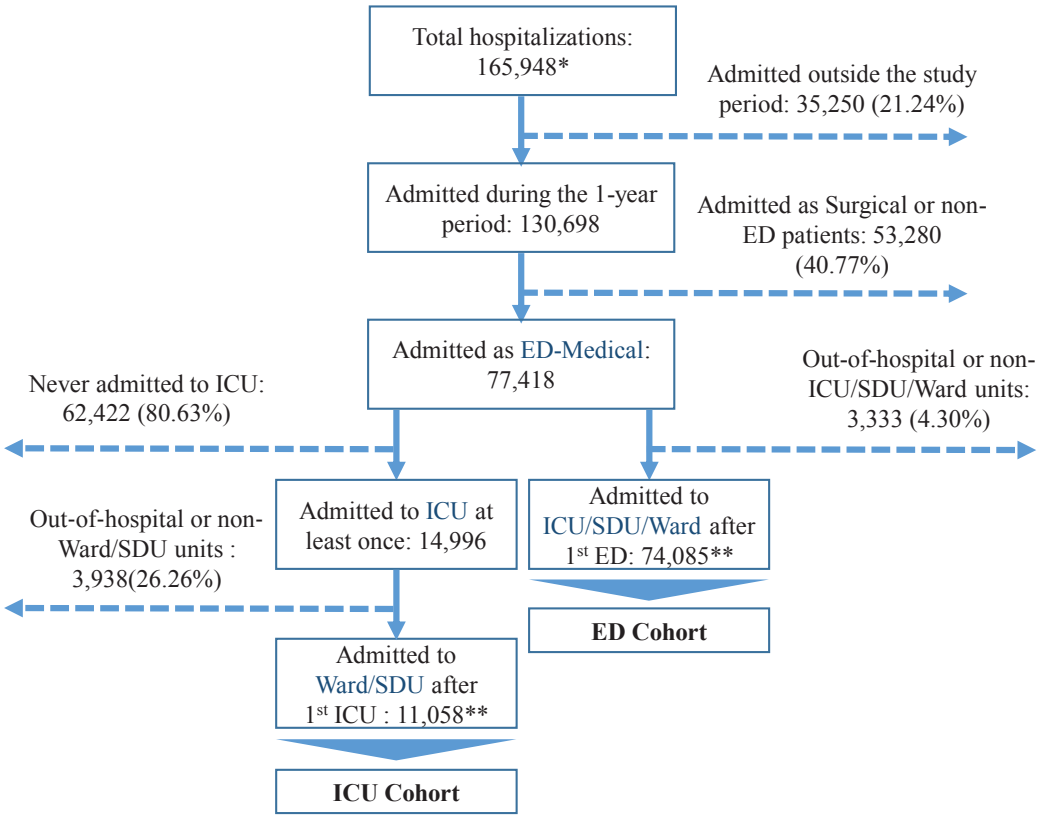
Our dataset also contains information about patient characteristics such as age, gender, admitting diagnosis and three different severity scores. One score (LAPS2) is based on lab results taken 72 hours preceding hospital admission and the second (COPS2) is based on comorbidities, such as diabetes, that may complicate patient recovery. These severity scores are assigned at hospital admission and are not updated during the hospital stay (more details on these scores can be found in Escobar et al. (2008, 2013)). The third severity score is the simplified acute physiology score 3 (SAPS3), which is a common severity score used exclusively for ICU patients (see, e.g, Strand and Flaatte (2008), Mbongo et al. (2009), Christensen et al. (2011)).

2.1. Data Selection

Since we study two different transfer decisions (from the ED and from the ICU), we form two separate patient cohorts: an ICU Cohort and an ED Cohort. Our data selection process is depicted in Figure 2. Because we use the patient flow data to determine the occupancy level (and capacity) for each unit, we first restrict both of our cohorts to the 12 months in the center of the 1.5 year time period in order to avoid censored estimates. A patient's admission category is defined as a combination of whether or not they were admitted through the ED, and whether they were admitted to a medical or surgical service resulting in 4 categories: ED-medical, ED-surgical, non-ED-medical, or non-ED-surgical. We primarily focus on patients

who are admitted via the ED to a medical service for two major reasons. First, this group is the largest, consisting of about 60% of the patients treated in these hospitals, and is similar to the cohort considered in Kim et al. (2015). Second, the care pathways of surgical patients tend to be fairly standardized (e.g. Gustafsson et al. (2011), Lassen et al. (2013), Miller et al. (2014), Thiele et al. (2015) among many other), especially for non-ED-surgical patients, which is the larger of the two surgical groups. In contrast, the care pathways of ED-medical patients are more variable. It is this variability we will leverage in our identification strategy (see Sections 3 and 5).

Figure 2 Data Selection



* to determine capacity and occupancy
** patient cohorts used in our econometric model

2.1.1. ICU Cohort Many SDUs are designed as true 'step-down units', where patients can only be admitted following ICU discharges (e.g. Eachempati et al. (2004)). Moreover, the ICU is the second most frequent unit from which SDU patients are transferred. Thus, our first cohort considers patients discharged from the ICU to either the SDU or ward. To form the ICU Cohort, we consider patients who are admitted to the ICU at least once during their hospital stay. For each patient, we focus on the initial ICU admission

within each hospitalization. We exclude patients who die in the ICU or are discharged directly home from the ICU, since there is no decision about whether to route these patients to the SDU or ward following ICU discharge[2].

2.1.2. ED Cohort Over 60% of SDU patients are admitted from the ED. For these patients, we consider the ED to inpatient unit admission decision. The three possible units a patient can be admitted to are the ICU, the SDU, or the Ward. We exclude the less than 5% of ED-medical patients who go directly to the Operating Room (OR) or Post-Anesthesia Recovery unit (PAR) from the ED.

Table 3 provides some summary statistics of these two cohorts. The SDU introduces a third level of care that, ideally, will be used to treat moderate to low severity patients, but not high severity patients. Our goal is to understand how service in this unit impacts quality of service, as measured by patient outcomes across different patient types. In doing so, we can gain a better understanding of the costs and benefits associated with utilizing a three levels of care structure to provide service to heterogenous customers.

Table 3 Summary Statistics of Patient Demographics

	ED Cohort				ICU Cohort			
Variable	mean	std	min	max	mean	std	min	max
Age	67.68	17.53	18	111	68.13	15.91	18	105
Male	0.47	0.50	0	1	0.55	0.50	0	1
LAPS2	74.70	37.35	0	272	75.13	49.10	0	262
COPS2	46.18	44.21	0	290	46.63	44.61	0	267
SAPS3	N/A				45.41	11.79	15	100
ED LOS (hrs)	1.46	2.20	0.02	118.68	1.57	2.73	0.02	118.68
Total LOS (hrs)	108.89	162.71	0.90	13138.50	235.53	362.38	6.20	13138.50
ICU LOS (hrs)	N/A				61.87	84.75	0.02	2279.17
LOS before ICU (hrs)	N/A				32.99	108.88	0	4877.58

Note: LAPS2 is a severity score based on lab results taken 72 hours preceding hospital admission. COPS2 is a severity score based on comorbidities. SAPS3 is a severity score used for ICU patients.

2.2. Patient Outcomes

We consider four patient outcomes: (1) in-hospital death ($Mortality$), (2) remaining hospital length-of-stay ($HospRemLOS$), (3) hospital readmission ($HospReadm$), and (4) ICU readmission ($ICUReadm$) for ICU patients.

The outcome $HospRemLOS$ is defined as the remaining time spent in the hospital following the transfer decision. Thus, for patients in the ED Cohort, this will be their total inpatient LOS; for patients in the ICU Cohort, this will be the remaining time spent in the hospital following ICU discharge.

$HospReadm_{2w}$ is defined as hospital readmission within two weeks after leaving the hospital (e.g., see Doran et al. (2013) and Ouanes et al. (2012) which use these durations). In calculating hospital readmission

[2] We consider analysis including these patients in our robustness checks.

rates, we exclude patients with in-hospital death. We also do robustness checks for different time windows for hospital readmission.

Following Brown et al. (2013) which aims to define reasonable time windows for ICU readmission, we consider $ICUReadm_{2d}$ ($ICUReadm_{5d}$) which indicate ICU readmission within two (five) days following ICU discharge. This measure is studied only for the ICU Cohort. We also do robustness checks for different time windows for ICU readmission.

Table 4 summarizes these patient outcomes for the two cohorts.

Table 4 Summary Statistics of Patient Outcomes: Mean (Number of observations or standard deviation for continuous variables)

	ED Cohort			ICU Cohort	
	ICU	SDU	Ward	SDU	Ward
Outcome	mean (N/std)	mean (N/std)	mean (N/std)	mean (N/std)	mean (N/std)
Mortality	0.12 (8,630)	0.04 (14,832)	0.03 (50,623)	0.06 (3,832)	0.07 (7,226)
HospRemLOS (days)	6.67 (11.51)	4.23 (5.89)	4.05 (5.79)	7.24 (14.76)	5.13 (10.91)
HospReadm - 2 weeks	0.12 (7,629)	0.11 (14,269)	0.10 (49,206)	0.14 (3,585)	0.13 (6,685)
ICUReadm - 2 days	N/A			0.04 (3,832)	0.05 (7,226)
ICUReadm - 5 days	N/A			0.08 (3,832)	0.06 (7,226)

2.3. Hypotheses

As there are various flows of patients into the SDU, we expect the impact of admission to the SDU to vary across different patient types. In particular, there is evidence that SDU care may improve or degrade patient outcomes (e.g. Zimmerman et al. (1995), Simchen et al. (2004)). Thus, we hypothesize that the SDU is beneficial or detrimental depending on patient type and severity–it will help moderate to low severity patients, but hurt high severity patients. More formally, we outline our hypotheses below.

As SDUs were initially developed with the intent to provide a 'step-down' from the ICU, we expect that ICU clinicians use SDUs appropriately so that:

Hypothesis 1 (ICU patients) *Patients discharged from the ICU will have better outcomes (lower mortality and readmission rates and shorter LOS) if admitted to the SDU rather than the ward.*

For patients admitted from the ED, the impact of SDU care is likely to be more nuanced. Specifically, this is a highly heterogenous group. We will describe how we partition patients into low, medium, and high severity groups in Section 5. The majority of patients admitted to the hospital from the ED do not go to the ICU (Kim et al. 2015). Thus, we expect that for most patients (i.e. low and medium severity patients), being treated in the SDU will either improve or have no impact on their outcomes. On the other hand, the sickest patients should be admitted to the highest level of care, so being admitted to the SDU is likely to result in worse outcomes. Note that in the following, we assume that low severity patients are rarely admitted to the ICU while high severity patients are rarely admitted to the ward.

Hypothesis 2 (Low Severity ED patients) *Low severity patients admitted from the ED will have no worse, and possibly better outcomes (lower mortality and readmission rates and shorter LOS), if admitted to the SDU rather than the ward.*

Hypothesis 3 (Medium Severity ED patients) *Medium severity patients admitted from the ED will have no worse, and possibly better outcomes (lower mortality and readmission rates and shorter LOS), if admitted to the SDU rather than the ward. On the other hand, they will have have no better, and possibly worse outcomes, if admitted to the SDU rather than the ICU.*

Hypothesis 4 (High Severity ED patients) *High severity patients admitted from the ED will have worse outcomes (higher mortality and readmission rates and long LOS) if admitted to the SDU rather than the ICU.*

3. ICU Cohort: Econometric Approach

We begin by explicitly stating our fundamental research question for the ICU cohort: Following ICU discharge, is SDU care associated with better patient outcomes than those for patients receiving ward care and, if so, what is the magnitude of the improvement? By exploring these questions, we will develop some insight into the value of differentiated levels of service (i.e. SDU versus ward) for one customer type (ICU patients). In Section 5, we expand our analysis to understand the impact of this level of service on additional patient types, providing insights into the role of customer differentiation.

3.1. Econometric Challenge: Endogeneity

Our objective is to utilize retrospective patient data to determine if ICU patients who are transferred to the SDU have better outcomes than those transferred to the ward. Because we are using retrospective data, an estimation challenge arises due to the fact that the routing decision following ICU discharge is likely correlated with patient outcomes. To highlight this challenge, we start with the following reduced form model for hospital LOS:

$$\log(HospRemLOS_i) = \beta X_i + \gamma ADMITSDU_i + \nu_{h(i)} + \epsilon_i \tag{1}$$

where X_i is a vector of control variables including patient characteristics (e.g. age) and seasonal factors (e.g, admission time of day), $ADMITSDU_i$ is an indicator variable that equals 1 if patient i is transferred directly to the SDU following ICU discharge, $h(i)$ is the hospital where patient i is treated, $\nu_{h(i)}$ is the hospital fixed effect and ϵ_i denotes the error term. See Table 14 in Appendix A for more details on control variables. While we include controls for patient severity, unobservable patient severity measures may be correlated with both $HospRemLOS$ and $ADMITSDU$. That is, sicker patients are more likely to be transferred to the SDU than the ward, but are also more likely to have bad outcomes. As such, our estimates for γ may be biased and we may erroneously conclude that going to the SDU hurts patients. To overcome this potential endogeneity bias, we utilize an identification strategy using Instrumental Variables (IVs).

3.1.1. Instrumental Variable A valid instrument should be 1) correlated with the endogenous variable, $ADMITSDU_i$, and 2) unrelated to the unobservable factors captured in ϵ_i which affect patient outcomes. We propose to use congestion in the SDU one hour before the ICU discharge as an IV. In particular, we define $SDUBusy_i$ as an indicator variable that equals one when the number of available beds in the SDU one hour prior to patient i's discharge from the ICU is less than or equal to two, and zero otherwise[3]. On average, about 11% patients are discharged from the ICU when the SDU is busy ($SDUBusy = 1$), though this varies quite a bit across hospitals (see Table 15).

When controlling for various patient characteristics in a Probit regression model, we also find at the 0.1% significance level that when the SDU is busy, patients are less likely to go to the SDU. In particular, we estimate that, on average, 21.14% percent of patients are routed to the SDU if $SDUBusy = 1$ and this percentage increases to 35.91% if $SDUBusy = 0$. Namely, a congested SDU is predicted to result in a 47% reduction in the likelihood of the SDU admission. Hence, condition 1 is satisfied.

We now consider Condition 2 and consider whether $SDUBusy_i$ is uncorrelated with unobservable factors in patient outcomes captured in ϵ_i. Since we cannot examine unobservable measures, we use patient severity, $SAPS3$, as a proxy for those unobservable factors. In particular, we perform a two-sample Kolmogorov-Smirnov test (see Gibbons and Chakraborti 2011 for details) to test the hypothesis that the distribution of SAPS3 for patients who are discharged from ICU when $SDUBusy = 1$ is not statistically different to that when $SDUBusy = 0$. The p-value for the combined Kolmogorov-Smirnov test is 0.136. Thus, we cannot reject the null hypothesis and believe that patients who are discharged from the ICU when $SDUBusy = 1$ are statistically similar to patients who are discharged from the ICU when $SDUBusy = 0$. For completeness, we also check this for the LAPS2 score, which is assigned at the time of hospital admission. The p-value of the combined Kolmogorov-Smirnov test is 0.334.

Kc and Terwiesch 2012 demonstrates that ICU congestion could result in early discharge, which could, in turn, affect the routing decision of ICU patients. While ICU congestion has been used as an IV in a number of hospital studies (e.g. Kc and Terwiesch 2012, Kim et al. 2015), we find that ICU congestion is not a valid IV. This is because the impact of ICU congestion does not exhibit a consistent effect on routing post-ICU patients, i.e., a congested ICU could result in both a higher and a lower percentage of patients being admitted to the SDU depending on a patient's severity score. Moreover, we find that the ICU congestion is correlated with a patient's SAPS3 and LAPS2 score.

We also considered using a number of additional instrumental variables. Specifically, we considered a measure of the average severity of other patients in the ICU, a measure of how the discharged patient compares to the severity of other patients in the ICU, and a measure of severity for the most recently discharged ICU patient. We find that all of these measures are correlated with the SAPS3 and LAPS2 scores, suggesting they may also be correlated with unobservable measures of severity, thereby invalidating these variables as potential instruments.

[3] We also do a number of robustness checks by considering different specifications of $SDUBusy_i$.

3.2. Econometric Model

3.2.1. Continuous outcome models We now present our estimation model for our continuous outcome, $HospRemLOS$. Since the ICU to SDU routing decision, $ADMITSDU_i$, is a binary variable, we model the ICU discharge decision via a latent variable model.

$$\begin{aligned} ADMITSDU_i^* &= X_i\theta + \alpha SDUBusy_i + \omega_{h(i)} + \xi_i, \\ ADMITSDU_i &= \mathbf{1}\left\{ADMITSDU_i^* > 0\right\}, \\ \log(HospRemLOS_i) &= X_i\beta + \gamma \cdot ADMITSDU_i + \delta \cdot AvgOccVisited_i + \nu_{h(i)} + \varepsilon_i, \end{aligned} \tag{2}$$

where $ADMITSDU_i^*$ is a latent variable which represents the propensity towards SDU admission; X_i is a vector of control variables for patient information; $\omega_{h(i)}$ is the hospital fixed effect; and, ξ_i represents unobservable factors that affect the routing at ICU discharge. For the outcome equation, $\nu_{h(i)}$ is the hospital fixed effect; and ε_i captures unobservable factors that affect patient outcomes.

Because congestion during a patient's hospital stay could impact the patient's outcomes (see Kuntz et al. (2015) and Kc and Terwiesch (2012)), we also control for the daily average occupancy level, denoted as $AvgOccVisited_i$, patient i experiences for all inpatient units s/he is admitted to *after leaving the ICU and before leaving hospital*. We also conduct robustness checks for different specifications of occupancy during the stay, as well as with such a control excluded. Kim et al. (2015) provides additional discussion regarding the necessity of such a control.

The error terms (ξ_i, ε_i) in (2) may be correlated to model the endogeneity between the routing decision at ICU discharge and the patient outcome. We assume that (ξ_i, ε_i) follows a Standard Bivariate Normal distribution with correlation coefficient ρ. This model can be jointly estimated using a treatment effect model via Full Maximum Likelihood Estimation (FMLE) (Greene 2012). A likelihood ratio test of null $\rho = 0$ can be used to test the presence of endogeneity.

3.2.2. Discrete outcome models For the binary outcomes ($Mortality, HospReadm, ICUReadm$), we modify Eq. (2) by replacing the continuous patient outcome with a probit model. Specifically, we have:

$$\begin{aligned} ADMITSDU_i^* &= X_i\theta + \alpha SDUBusy_i + \omega_{h(i)} + \xi_i, \\ ADMITSDU_i &= \mathbf{1}\left\{ADMITSDU_i^* > 0\right\}, \\ y_i^* &= X_i\beta + \gamma \cdot ADMITSDU_i + \delta \cdot AvgOccVisited_i + \nu_{h(i)} + \varepsilon_i, \\ y_i &= \mathbf{1}\left\{y_i^* > 0\right\} \end{aligned} \tag{3}$$

where y_i^* is a latent variable which represents the propensity for the outcome. Similar to before, we assume that (ξ_i, ε_i) follows a Standard Bivariate Normal distribution with correlation coefficient ρ. This Bivariate Probit model can be jointly estimated via FMLE (see Cameron and Trivedi 1998, Greene 2012). The presence of endogeneity can be tested through a likelihood ratio test of null $\rho = 0$.

For ICU readmission, we modified $AvgOccVisited_i$ to be the daily average occupancy level that patient i experiences in all inpatient units s/he is admitted to *between two consecutive ICU admissions*.

3.3. Impact of Congestion on ICU LOS

Kc and Terwiesch (2012) found evidence that when ICUs are highly congested, current ICU patients may be demand-driven discharged, in order to accommodate incoming demand of more severe patients. Kim et al. (2015) found that patients admitted to a medical service from the ED do not seem to be susceptible to such demand-driven discharges. While we look at a similar group of patients to Kim et al. (2015), one potential concern is that we only consider patients treated in hospitals with SDUs, while Kim et al. (2015) includes hospitals with SDUs as well as those without. Thus, it is possible that the presence of an SDU makes it more likely for medical patients who were admitted to the hospital via the ED and are being treated in the ICU to be demand-driven discharged; thus, making it possible that these types of discharges occur in our dataset. A patient who is demand-driven discharged is by definition, discharged earlier than under ordinary circumstances and therefore more critical than if he were discharged later at a more appropriate time. So such a patient is more likely to be admitted to the SDU, but also more likely to have bad outcomes. If this were the case, this could cause a downward bias of our results.

To check this, we estimated the following reduced form model:

$$\log(ICULOS_i) = \eta X_i + \kappa ICUBUSY_i + \upsilon_i \tag{4}$$

to explore whether ICU LOS is reduced when the ICU is busy. We estimate κ to be -0.05 with standard error 0.04. Thus, consistent with Keenan et al. (1998) and Kim et al. (2015), we do not find evidence that patients are demand-driven discharged. To dig a little deeper, we examined whether the SDU congestion had an impact on whether patients are demand-driven discharged. To do this, we enhance our regression model to include a measure of SDU congestion:

$$\log(ICULOS_i) = \eta X_i + \kappa ICUBUSY_i + \phi SDUBUSY_i + \psi\left(ICUBUSY_i \times SDUBUSY_i\right) + \upsilon_i \tag{5}$$

In particular, we would expect demand-driven discharges to be most common when the ICU is busy and the SDU is not. Table 5 summarizes these results with the base case of both the ICU and SDU not being busy (81.5% of time). We find that the coefficients have very large standard errors and are not statistically significant. While it is possible that lack of statistical power is the reason we do not find evidence to support the hypothesis that a busy ICU may result in demand-driven discharges, we find that our sample size would need to be larger than 350,000 for the estimated coefficients to be statistically significant when using the approach in Gelman and Hill (2006).

Our IV analysis is based on the evidence that a busy SDU decreases the likelihood of SDU admission. However, it is also possible that patients may stay longer in the ICU when the SDU is busy, making them more stable upon discharge from the ICU and potentially biasing our results. To test this hypothesis, we ran the reduced form model in Equation (4), but with $SDUbusy_i$ as an explanatory variable. We find the

Table 5 Effect of $ICUBusy$ and $SDUBusy$ on ICU LOS

Parameter	ICU Busy	SDU Busy	Estimate (SE)	# Observations: Total = 11,058
κ	1	0	-0.057 (0.040)	855
ϕ	0	1	-0.039 (0.039)	1,056
ψ	1	1	-0.034 (0.096)	136

Note. Standard error in parentheses. $^{+}(p<10\%),^{*}(p<5\%),^{**}(p<1\%),^{***}(p<0.1\%)$.

coefficient for $SDUBusy_i$ to be -0.02 with standard error 0.03. This is consistent with the results in Table 5, which suggests that the relationship between a busy SDU and ICU LOS is not statistically significant. As an additional check, we ran a hazard rate model to examine the impact of $SDUBusy$ after controlling for patient characteristics, seasonality, and hospital fixed effects. Again, we see that a busy SDU does not have a statistically significant effect on the likelihood of ICU discharge. Thus, we do not find evidence to support that the busy-ness of the SDU impacts ICU LOS.

4. ICU Cohort: Results

We start by exploring the impact of SDU care on patients being discharged from the ICU. Because we jointly estimate the SDU admission decision and patient outcomes, using FMLE, the impact of $SDUBusy_i$ may vary slightly for different outcomes. That said, we observe that the differences are very minor. For illustrative purposes, we note that the coefficient for the impact of $SDUBusy_i$ in the Mortality model is -0.5110 with standard error 0.0503 and p-value $< 0.1\%$.

Table 6 Estimated Effect of SDU Admission Following ICU discharge (γ) on Patient Outcomes and Correlation between error terms (ρ) for the admission decision and patient outcomes: $N = 11,058$

	With IV					Without IV
Outcome	γ (SE)	Predicted Outcome $\hat{P}_{SDUBusy=0}$	Predicted Outcome $\hat{P}_{SDUBusy=1}$	ρ (SE)	Test $\rho=0$	γ (SE)
$Mortality$	-0.60** (0.22)	8.24%	9.93%	0.26^{+} (0.14)	0.07	-0.18*** (0.05)
$\log(HospRemLOS)$	-0.35*** (0.10)	3.77	4.05	0.44*** (0.05)	0.00	0.38*** (0.02)
$ICUReadm_{2d}$	-0.51** (0.20)	5.22%	6.38%	0.32* (0.12)	0.02	0.01 (0.05)
$ICUReadm_{5d}$	-0.51** (0.18)	8.18%	9.83%	0.36** (0.11)	0.05	0.09* (0.04)
$HospReadm_{2w}$	-0.43* (0.21)	14.02%	15.26%	0.21^{+} (0.12)	0.09	0.05 (0.04)

Note. Standard error in parentheses. $^{+}(p<10\%),^{*}(p<5\%),^{**}(p<1\%),^{***}(p<0.1\%)$.
Predicted outcome: $\hat{P}_{SDUBusy=0}$ - Average predicted outcome if the SDU was never busy
$\hat{P}_{SDUBusy=1}$ - Average predicted outcome if the SDU was always busy.
Predicted $HospRemLOS$ (days) is shown instead of $\log(HospRemLOS)$

As we are primarily interested in estimating the causal effects of SDU admission on patient outcomes, we report only the coefficient of SDU admission on the patient outcomes, i.e., γ in (2) and (3). Table 6 summarizes the relationship between SDU admission right after ICU discharge and patient outcomes. The sign of SDU admission is negative and statistically significant in all outcome measures, suggesting that routing an ICU discharge to the SDU is associated with improved patient outcomes. To get a rough

estimate of the magnitude of the effects we've estimated, we also use our estimation results to predict patient outcomes under two extreme scenarios: (i) the SDU has ample capacity and is never busy (referred to as $\hat{P}_{SDUBUSY=0}$) versus (ii) the SDU is always busy ($\hat{P}_{SDUBusy=1}$). There are some patients who are stable enough that even if there were ample capacity in the SDU, they would not be admitted to the SDU. Instead they would be sent to the ward, irrespective of the SDU bed availability. Using the first stage of our models, we estimate the likelihood of SDU admission depending on whether the SDU is busy or not, and then use our second stage results to estimate the likely patient outcome. We find that, on average, availability of SDU care is associated with significant improvements in patient outcomes: the relative reduction is 17% in the likelihood of in-hospital death, 0.28 fewer remaining hospital days, 18% (17%) in the likelihood of ICU readmission within 2 (5) days, and 8% in the likelihood of hospital readmission within 2 weeks. Based on the size of our cohort, we estimate eliminating SDU busy-ness would translate into annual savings of 187 lives, 3,096 hospital days, 128 (182) ICU readmissions in 2 (5) days, and 137 hospital readmissions in 2 weeks (weekly savings of 3.6 lives, 59.5 hospital days, 2.5 (3.5) ICU readmissions, and 2.6 hospital readmissions) aggregated across the 10 hospitals.

Our empirical findings also suggest strong evidence of an endogeneity bias between the routing following ICU discharge and patient outcomes. The p-value of the likelihood ratio test with null hypothesis that the correlation between the two error terms in our model $\rho = 0$ is small, as seen in Table 6, implying a strong correlation between the routing at ICU discharge and patient outcomes. Ignoring this endogeneity tends to result in underestimates of the benefit of SDU care and could result in a qualitatively different insight; see the column titled with "Without IV".

4.1. Robustness Checks

We now describe a number of robustness checks for our main results. First, we tried different specifications of control variables. Recall that, some of our control variables – age, severity scores (LAPS2, COPS2, SAPS3), length-of-stay at ICU, and length-of-stay before ICU admission – are modeled as spline variables to account for their possible non-linear effects on the ICU to SDU routings and patient outcomes. We repeated the analysis with different specifications, including changing the number of cutoffs and values of these cutoffs. Our results are qualitatively similar to these changes.

The second robustness check we did is with respect to specifications for the congestion experienced by a patient during the hospital visit ($AvgOccVisited$ in Equations (2) and (3)). We considered specifications which exclude this control as well as ones that examine the maximum occupancy in any unit during a patient's hospital stay. All specifications yield similar results to those reported in Table 6.

Another factor which could be impacting our results is "do not resuscitate (DNR)" orders, which are patients' end-of life wishes not to undergo Cardiopulmonary resuscitation (CPR) or advanced cardiac life support if their heart were to stop or they were to stop breathing. In speaking with intensivists, we learned

it is possible that patients with DNRs are more likely to be sent to the ward, but also may be more likely to die, resulting in an overestimate of the effect of SDU care. Unfortunately, we do not have access to patients' DNR status, so cannot control for this. That said, DNR orders only represent 9% of ICU patients (Jayes et al. 1993), so this is likely to affect only a small percentage of patients. Additionally, there is evidence that DNR orders do not change the quality of care (Baker et al. 2003). We do not expect DNR orders to impact our results for hospital readmission since we exclude patients who died in hospital in this model. For the LOS models, we also considered the robustness of our results to including patients with in-hospital death. We find that our results are very robust.

We also considered alternative specifications for the length of time window for readmission. For ICU readmission, we varied the time window of the ICU readmission from time of ICU discharge from 2 to 7 days and also during any time frame during the same hospital stay. Only the results for ICU readmission within 2 days were statistically significant, though the sign of the coefficient was negative in all models. For hospital readmission, we consider hospital readmission within 1 week, 2 weeks, and 30 days after a patient is discharged from the hospital. We found that while SDU admission is associated with lower hospital readmission risk, the effect is *weaker* when the elapsed time between two consecutive hospital stays is longer.

4.1.1. Definition of our IV We also consider various definitions of a busy SDU. First, we considered different cutoffs for the number of available beds, ranging from one bed to four. On average, the percentage of patients, who are discharged from the ICU when the SDU is congested, varies from 34% to 3% when the cutoff is decreased from four beds to one (Table 15). The capacity of the SDU was defined as the maximum occupancy level over the 12-month time horizon in our study. While capacity changes in the hospitals we study are very rare, we also allow for the bed capacity (defined as the maximum occupancy level) to change over time. Specifically, we define a time-varying capacity as the maximum occupancy level over three non-overlapping 4-month periods during the total 12-month time horizon.

Note that while we find our IV to be statistically significant based on various definitions of bed capacity, it can be very challenging to accurately determine the number of beds available in a unit. This is because capacity depends on multiple factors including the number of physical beds, but also the number of nurses and physicians available to staff them. As such, we also considered alternative measures of SDU congestion based on percentiles of the SDU occupancy level. We did this using a binary variable indicating whether the occupancy level exceeds a threshold percentile as well as a piece-wise linear spline to potentially model non-linear effects of SDU congestion on the SDU admission decision. Next, we considered different time lines for when SDU congestion was measured: 1 hour (main specification), 2 hours and 6 hours before ICU discharge.

We find that the results for the mortality and LOS models to be very robust to the various specifications with the coefficients all negative and all with the same order of magnitude. All coefficient estimates have

a p-value < 0.05, with most having a p-value $p < 0.001$. Interestingly, the coefficient estimates for the ICU and hospital readmission models are all negative; however, the statistical significance of the coefficient estimates varies substantially, with some specifications indicating a p-value < 0.001 and others not being significant, even at the level of p-value $< .1$. Thus, while our mortality and LOS results are quite robust, the readmission results do not seem to be.

4.1.2. Patients discharged out of the hospital In all of our analysis for the ICU cohort, we focused on patients who were discharged from the ICU to the SDU or ward. While the majority of patients (83.82%) go to one of these units, a number of patients are actually discharged directly out of the hospital from the ICU (see Table 16 in the Appendix). Not surprisingly, patients who are discharged out of the hospital directly from the ICU appear to be healthier (lower severity scores and younger) than those admitted to an inpatient unit following ICU discharge. We find that if we include all patients who are discharged alive from the ICU (to the SDU, ward or out of the hospital) instead of just those discharged to the SDU or ward, the busy-ness of the SDU still has a statistically significant effect on the likelihood of SDU admission following ICU discharge. In this analysis, we found evidence that patients are 1.77% more likely to be discharged home alive when the SDU is busy ($p < 0.05$). As such, excluding these patients from our analysis may bias our estimates to make the SDU seem more beneficial than it is because the patients who end up staying in the hospital are sicker and SDU treatment likely benefits them more. With this in mind, we re-ran our ICU and hospital readmission models including patients discharged out of the hospital alive. Note that we do not examine our mortality and LOS models as these are inpatient outcomes, and a patient who is discharged out of the hospital alive will, by definition, have $death_i = 0$ and $HospRemLOS = 0$. We find that the coefficient estimates for our readmission models are negative, but not statistically significant. This is consistent with our other specifications which suggest that the readmission results are not very robust.

5. ED Cohort: Econometric Approach

In this section, we study the routing decision regarding the ED Cohort. We aim to empirically estimate how SDU admission immediately following transfer from the ED affects patient outcomes, comparing to ED patients who are transferred to the ICU or ward. Here, a similar estimation challenge arises. Routing decisions are associated with patient severity and, thus, with patient outcomes.

Kim et al. (2015) examined this problem in the context of admitting patients to the ICU from the ED. In that paper, the goal was to estimate the impact of admitting a patient to the highest level of care, i.e. the ICU versus elsewhere. In contrast, our objective is to understand the impact of admitting patients to an intermediary level of care, the SDU. In contrast to the ICU case, it is possible that the impact of SDU care could be positive, neutral or even negative. For instance, high severity patients who should be admitted to the ICU, but are instead admitted to the SDU may experience worse outcomes as a result. On the other hand, SDU care may have no impact or even benefit low severity patients who would traditionally be cared

for in the ward. There are limited objective standards for who should be treated in the ICU (see Task Force of the American College of Critical Care Medicine, Society of Critical Care Medicine (1999) and Kim et al. (2015)), let alone for the SDU (Nasraway et al. 1998). Thus, such categorizations of patients are likely to be highly varied across different physicians. As such, we take a data-driven approach to stratifying patients by severity. Such an approach could be useful in other service settings where precisely defining a customer's type is a challenging, but necessary, step toward determining the costs and benefits associated with service by different server types.

5.1. Severity Categorization

In order to estimate the impact of SDU care for patients admitted from the ED, we categorize patients based on their severity and study each severity group separately. Specifically, we aim to identify a 'low severity' cohort, for which the decision is to admit patients to either the ward or SDU, and a 'high severity' cohort for which the decision is to admit to either the SDU or ICU. One can also consider a 'medium severity' cohort whose patients can be admitted to any one of the three levels of care. Certainly, it seems reasonable to expect the decision to admit a patient to the SDU will have a different impact on patients of varying severity.

We begin by considering how patient level characteristics influence whether a patient is admitted to the ICU, SDU or ward from the ED. Specifically, we use our data to estimate an Ordered Probit regression model using only patient characteristics.

$$Tx_i = \begin{cases} Ward, \text{ if } & Tx_i^* \leqslant t_1 \\ SDU, \text{ if } & t_1 < Tx_i^* \leqslant t_2 \\ ICU, \text{ if } & t_2 < Tx_i^* \end{cases} \quad \text{and} \quad Tx_i^* = X_i'\theta + \xi_i, \tag{6}$$

where X_i' is a vector of control variables for patient characteristics and ξ_i represents unobservable factors.

We use the observed latent variable $\widehat{Tx_i}^* = X_i'\theta$ to define each patient's severity. Intuitively, $\widehat{Tx_i}^*$ is a linear transformation of patient characteristics into a single continuous variable which can be interpreted as a measure for the desired amount of care for the patient. The larger the value of $\widehat{Tx}^*$, the more likely the patient will be routed to higher level units, e.g., the ICU; the lower the value, the more likely a patient will be routed to the ward.

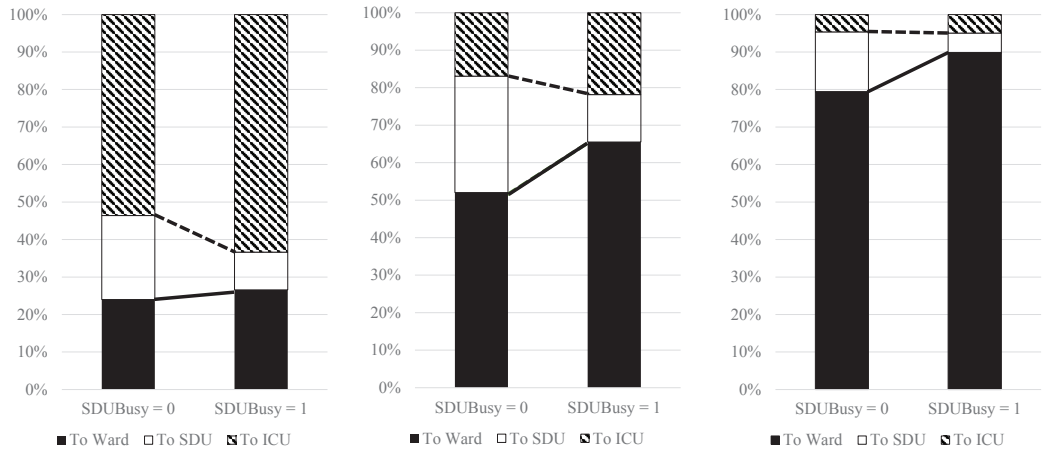
We differentiate patient severity groups by partitioning the $\widehat{Tx_i}^*$ space with thresholds. In theory, t_1 and t_2 from (6) partition the Tx_i^* space into patients who will be routed to the Ward, SDU, and ICU, so that patients with $Tx_i^* \leq t_1$ could be classified as low-severity patients and patients with $Tx_i^* > t_2$ could be classified as high-severity patients. However, because we do not observe ξ_i, we are only able to observe an estimate $\widehat{Tx_i}^*$, instead of Tx_i^*. Thus, some patients with $\widehat{Tx_i}^* \leq t_1$ will be routed to the SDU, or even the ICU. Similarly, patients with $\widehat{Tx_i}^* > t_2$ may be routed to the SDU or ward. Increasing t_2 will increase the proportion of patients with $\widehat{Tx_i}^* > t_2$ who are routed to the ICU and simultaneously decrease the proportion who are routed to the ward. Similarly, decreasing t_1 will increase the proportion of patients with $\widehat{Tx_i}^* \leq t_1$

being admitted to the ward and decrease the proportion being admitted to the ICU. Of course, this also comes at the cost of reducing the number of patients which satisfy these two criteria. Thus, we define the cutoffs to balance increasing the proportion of patients in the high (low) severity group who are routed to the ICU (ward) versus maintaining large enough patient cohorts to allow for meaningful statistical analysis. The tradeoff we are concerned with is close thresholds lead to increases in patient spill-over into the high or low severity groups, resulting in patients who do not comply with our instrument, versus far thresholds which reduce sample sizes, resulting in less statistical power. We use a data-driven approach and find that setting thresholds at the 95^{th} and 60^{th} percentiles of the distribution of $\widehat{Tx}_i^*$ seen in the data achieve this delicate balance. In Section 6, we discuss robustness checks using different thresholds.

We expect that when the SDU is congested, patients will be less likely to be admitted (e.g. see Section 3). Thus, we examine where patients are admitted when the SDU is busy, defined as done in Section 3. Figure 3 shows the proportion of high and low severity patients admitted to each unit, while Table 7 summarizes these results. Note that the ICU and SDU congestion have a correlation coefficient of 0.08, so the busyness of the ICU does not factor substantially into these results. Specifically, we ran t-tests comparing the proportion of patients admitted to each level of inpatient unit when the SDU is busy versus not busy. As we can see, when the SDU is busy, low severity patients will be rerouted to the ward ($p < 0.001$), rather than the ICU ($p = 0.327$). Conversely, when the SDU is busy, high severity patients tend to be rerouted to the ICU ($p = 0.002$), rather than the ward ($p = 0.212$). These results are suggestive that these severity categorizations are reasonable for our purposes.

Note that one can also define a 'medium severity' group as patients with $t_2 \leq \widehat{Tx}_i^* < t_1$. While a busy SDU does decrease the likelihood of SDU admission (Figure 3(b) and Table 7), the challenge with this cohort is that some patients who are discouraged from being admitted to the SDU will be admitted to the ICU, while others will be admitted to the general ward. Certainly, being bumped to a higher versus lower level of care will have a substantial impact on patient outcomes. As seen in Figure 3(b) and Table 7, there is a heterogenous effect of $SDUBusy$ on these patients, whereas high-severity patients are consistently bumped up to the ICU and low-severity patients are consistently bumped to the ward. This suggests that within the medium severity group a mix of high and low severity patients are being admitted to the SDU, so SDU admission can be beneficial or detrimental. By grouping these patients together into a medium severity classification, we cannot tease out the true impact of SDU admission. Still, for completeness, we will include results for this group of patients.

As summarized in Table 8, for the high severity group, 54.9% are admitted to the ICU, 20.76% to the SDU, and 24.35% to the ward. For low severity patients 4.65%, 14.73% and 80.62% are admitted to the ICU, SDU, and ward, respectively. We can see that even with our classifications, some high (low) severity patients will still be admitted to the ward (ICU). In order to focus on the impact of SDU admissions on patient outcomes, we exclude high (low) severity patients who are routed to the ward (ICU). For the medium



(a) High-severity Patients (b) Medium-severity Patients (c) Low-severity Patients

Figure 3 Proportions of ED patients who are routed to the ICU, SDU, and ward when $SDUBusy = 1$ vs $SDUBusy = 0$. Severity thresholds, t_1 and t_2, defined by 95^{th} and 60^{th} percentiles of $\widehat{Tx}_i^*$.

Table 7 Proportions of ED patients who are routed to the ICU, SDU, and ward when $SDUBusy = 1$ vs $SDUBusy = 0$ and results of t-tests which compare the difference in routing proportions.

	SDUBusy = 0			SDUBusy = 1			p-value of t-test		
	ICU	SDU	Ward	ICU	SDU	Ward	ICU	SDU	Ward
Low Severity	4.62%	15.97%	79.14%	4.94%	5.23%	89.93%	0.327	< 0.001	< 0.001
Medium Severity	16.90%	31.05%	52.04%	21.84%	12.64%	65.51%	0.013	< 0.001	0.001
High Severity	53.60%	22.39%	24.01%	63.38%	10.06%	26.56%	0.002	< 0.001	0.212

Note. Severity thresholds, t_1 and t_2, defined by 95^{th} and 60^{th} percentiles of $\widehat{Tx}_i^*$.

severity group, we consider patients admitted to all three levels of care as it is not clear whether the ICU or ward is the 'more desirable' unit if the SDU is not available. Tables 9 and 10 report summary statistics of patient demographics and outcomes for each severity group.

Table 8 Routing Statistics of Patients for Different Severity Groups

	Low-Severity Patients		Med-Severity Patients		High-Severity Patients	
Unit following the ED	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
ICU	2,067	4.65	4,529	17.47	2,034	54.90
SDU	6,549	14.73	7,514	28.98	769	20.76
Ward	35,836	80.62	13,885	53.55	902	24.35

Note. Severity thresholds, t_1 and t_2, defined by 95^{th} and 60^{th} percentiles of $\widehat{Tx}_i^*$.

5.2. IV Justification

We are again faced with the econometric challenge of endogeneity bias. Our econometric model is very similar to that of (2) and (3). The main difference is that for low (high) severity patients, $ADMITSDU_i$ is equal to 1 if the patient is admitted to the SDU and 0 if to the ward (ICU). For medium severity

Table 9 Summary Statistics of Patient Demographics for ED Cohort by severity classification

	Low Severity				Medium Severity				High Severity			
Variable	mean	std	min	max	mean	std	min	max	mean	std	min	max
Age	67.48	18.57	18	107	67.63	15.96	18	111	70.36	14.74	18	102
Male	0.43	0.49	0	1	0.54	0.50	0	1	0.57	0.50	0	1
LAPS2	59.48	26.89	0	158	89.22	32.53	0	200	155.57	31.85	16	272
COPS2	41.96	41.21	0	285	50.86	46.53	0	290	64.13	53.64	0	278
ED LOS (hrs)	1.38	1.99	0.02	62.73	1.56	2.41	0.02	118.68	1.66	2.92	0.02	113.50

Note: LAPS2 is a severity score based on lab results taken 72 hours preceding hospital admission.
COPS2 is a severity score based on comorbidities. SAPS3 is a severity score used for ICU patients.
Severity thresholds, t_1 and t_2, defined by 95^{th} and 60^{th} percentiles of $\widehat{Tx}_i^*$.

Table 10 Summary Statistics of Patient Outcomes for ED Cohort by severity classification: Mean (Number of observations or standard deviation for continuous variables)

	Low Severity		High Severity	
Outcome	SDU mean (N/std)	Ward mean (N/std)	SDU mean (N/std)	ICU mean (N/std)
Mortality	0.02 (6,549)	0.02 (35,836)	0.17 (769)	0.27 (2,034)
HospRemLOS (days)	3.97 (5.85)	3.95 (5.21)	6.68 (10.54)	9.35 (14.22)
HospReadm - 2 weeks	0.10 (6,431)	0.10 (35,258)	0.17 (636)	0.16 (1,483)

	Medium Severity		
Outcome	SDU mean (N/std)	Ward mean (N/std)	ICU mean (N/std)
Mortality	0.04 (7,514)	0.04 (13,885)	0.08 (4,529)
HospRemLOS (days)	4.20 (5.15)	4.22 (6.67)	6.11 (11.77)
HospReadm - 2 weeks	0.12 (7,202)	0.11 (13,310)	0.12 (4,156)

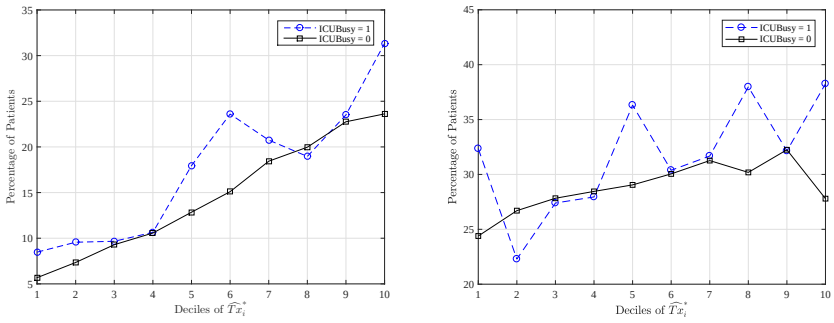
Severity thresholds, t_1 and t_2, defined by 95^{th} and 60^{th} percentiles of $\widehat{Tx}_i^*$.

patients, $ADMITSDU_i$ is equal to 1 if the patient is admitted to the SDU and 0 if to the ward or ICU. Detailed descriptions of the covariates are shown in Table 14 in the Appendix. Similarly, we also control for $AvgOccVisited_i$, i.e., the daily average occupancy level patient i experiences for all inpatient units s/he is admitted to *after leaving the ED and before leaving hospital*.

Similar to our models for the ICU Cohort, we consider using $SDUBusy_i$ as an instrumental variable. Additionally, we consider using $ICUBusy_i$ as an instrument as Kim et al. (2015) found that it is a good instrument when studying patients who are or are not admitted to the ICU, which is similar to our High severity group. Specifically, we define $SDUBusy_i$ ($ICUBusy_i$) as an indicator variable that equals one when the number of available beds in the SDU (ICU) one hour prior to patient i's transfer from the ED is less than or equal to two, and zero otherwise. On average, the proportions of patients who are transferred from the ED when the SDU is busy and the ICU is busy are approximately 12% and 6%, respectively.

As discussed previously, in order for a variable to be a valid instrument, it has to be 1) correlated with the endogenous variable, $ADMITSDU_i$, and 2) unrelated to the unobservable factors which affect patient outcomes. As seen in Table 7, when the SDU is busy, patients are less likely to be admitted to the SDU.

However, we find that ICU congestion does not appear to have a monotonic effect on SDU admission for low or medium severity patients. Specifically, we observe in Figure 4 that when we partition the (a) low or (b) medium severity patients into deciles of $\widehat{Tx}_i^*$, ICU congestion increases the percentage of SDU admissions for some patients, while it has no effect or even *decreases* the percentage of SDU admissions for other patients. Therefore, we conclude that $ICUBusy_i$ is not a valid instrument for low or medium severity patients. We see these effects more concretely when we analyze a Probit regression model, which controls for various patient characteristics and operational controls. We find with 0.1% significance level that SDU congestion reduces the likelihood of SDU admission for both low, medium and high severity patients, and that ICU congestion increases the chance of SDU admission for only high severity patients. The impact of ICU congestion for low and medium severity patients is not statistically significant.



(a) Low Severity (b) Medium Severity

Figure 4 **Percentage of (a) low or (b) medium severity patients admitted to the SDU from the ED when the ICU is busy ($ICUBusy = 1$) or not ($ICUBusy = 0$) for varying levels of severity as measured by deciles of $\widehat{Tx}_i^*$, given patients are classified as low severity: $\widehat{Tx}_i^* \leq t_1$. Low/Medium/High severity thresholds, t_1 and t_2, defined by 95^{th} and 60^{th} percentiles of $\widehat{Tx}_i^*$.**

We next examine whether our instruments are correlated with observable measures of severity. We again perform a two-sample Kolmogorov-Smirnov test to test the hypothesis that the distribution of LAPS2 is not statistically different when $SDUBusy = 1$ ($ICUBusy = 1$) from that when $SDUBusy = 0$ ($ICUBusy = 0$). For low severity patients, the p-value for the Kolmogorov-Smirnov test is 0.135, thus, we conclude that patients who leave the ED when $SDUBusy = 1$ are statistically similar to those who leave the ED when $SDUBusy = 0$. For medium severity patients, the p-value for the Kolmogorov-Smirnov test is 0.120. For high severity patients, the p-values are 0.141 and 0.358 for $SDUBusy$ and $ICUBusy$, respectively. Therefore, our models for low and medium severity patients use $SDUBusy_i$ as an instrument, while both $SDUBusy_i$ and $ICUBusy_i$ are used in the models for high severity patients.

5.2.1. Additional Instruments Apart from the congestion in the ICU and the SDU, we also consider other potential behavioral IVs discussed in Kim et al. (2015). The first factor is $RecentDischarge_i^{SDU}$, which accounts for the number of all SDU discharges in the 3-hr window before patient is admission to the first inpatient unit. The second behavioral factor, $RecentAdmission_i^{SDU}$, accounts for the number of SDU admissions in the 3-hr window before patient is admission to the first inpatient unit. To define $RecentDischarge_i^{SDU}$ and $RecentAdmission_i^{SDU}$, we normalize the number of discharges or admissions by the SDU capacity of each hospital. The third factor, $LastAdmitSeverity_i^{SDU}$, measures the severity of the last patient admitted to the SDU from the ED. We also consider $RecentDischarge_i^{ICU}$, $RecentAdmission_i^{ICU}$, $LastAdmitSeverity_i^{ICU}$, which are defined the same way but instead involve the ICU. Most of these variables demonstrate a heterogeneous impact on the SDU admission decision; for instance, amongst low severity patients, $RecentAdmission_i^{SDU}$ will increase the likelihood of SDU admission, while it will decrease likelihood for other patients. We find that only $RecentAdmission_i^{ICU}$ is a valid instrument and is valid only for high-severity patients. However, we do not include this as a third IV for high-severity patients in our main specifications because the results are similar.

6. ED Cohort: Results

We now present our main results for our ED cohort on the impact of SDU admission on patients being admitted to an inpatient unit from the ED. We start with the two patient cohorts for which the routing decision is more straight-forward (low and high severity patients). Then, for completeness, we include the results for the medium severity patients.

6.1. Low Severity

For low severity patients, a busy SDU is associated with a decrease in likelihood of SDU admission. For the mortality model, the coefficient on $SDUBusy_i$ is -0.5117 with standard error 0.0376 and p-value $< .1\%$. The results are similar for the other patient outcome models.

Table 11 summarizes our results. We find that SDU care may benefit low severity patients. Specifically, we find that SDU care is associated with lower mortality rate and shorter hospital remaining length-of-stay, as seen in the negative sign of SDU admission coefficient. We also present the predicted patient outcomes under two extreme scenarios: (i) the SDU is never busy ($\hat{P}_{SDUBusy=0}$) and (ii) the SDU is always busy ($\hat{P}_{SDUBusy=1}$). Our results indicate that, on average, availability of SDU care is associated with a reduction in mortality by 3.2% and 29 minutes of hospital remaining length-of-stay. We note that the estimated marginal effects are quite small as there is a substantial proportion (83%) of low severity patients who will not be admitted to the SDU even when the SDU is not busy. As such, the outcomes for these patients will be agnostic to whether the SDU is busy, since they will be admitted to the ward either way. Thus, our estimates are only for the remaining 17% of low severity patients whose routing from the ED is dictated by the state of the SDU. We do not find a statistically significant relationship between SDU care and the likelihood of

hospital readmission within 2 weeks. Based on the size of the low severity cohort, we estimate eliminating SDU busy-ness would translate into annual savings of 34 lives and 979 hospital days (weekly savings of 0.65 lives and 18.8 hospital days) aggregated across the 10 hospitals.

Table 11 Estimated Effect of SDU Admission Following the ED (γ) on Patient Outcomes for Low Severity Patients and Correlation between error terms (ρ) for the admission decision and patient outcomes: $N = 42,385$.

	With IV					Without IV
Outcome	γ (SE)	Predicted Outcome $\hat{P}_{SDUBusy=0}$	Predicted Outcome $\hat{P}_{SDUBusy=1}$	ρ (SE)	Test $\rho = 0$	γ (SE)
$Mortality$	-0.55** (0.28)	2.16%	2.23%	0.33+ (0.17)	0.07	0.06 (0.06)
$\log(HospRemLOS)$	-0.20*** (0.04)	2.92	2.94	0.18*** (0.03)	0.00	0.03* (0.01)
$HospReadm_{2w}$	-0.13 (0.12)	-	-	0.09 (0.07)	0.21	0.02 (0.03)

Note. Standard error in parentheses. $^{+}(p < 10\%)$,$^{*}(p < 5\%)$,$^{**}(p < 1\%)$,$^{***}(p < 0.1\%)$.
Low severity threshold, t_1, defined by 60^{th} percentile of $\widehat{Tx}_i^*$.
Predicted outcome: $\hat{P}_{SDUBusy=0}$ - Average predicted outcome if the SDU was never busy
$\hat{P}_{SDUBusy=1}$ - Average predicted outcome if the SDU was always busy.
Predicted $HospRemLOS$ (days) is shown instead of $\log(HospRemLOS)$

6.2. High Severity

We find that a busy SDU is associated with a decrease in likelihood of SDU admission for high severity patients, while a busy ICU is associated with an increase in likelihood of SDU admission. For the mortality model, the coefficient on $SDUBusy_i$ is -0.6325 with standard error 0.1043 and p-value $< .1\%$; for $ICUBusy_i$, the coefficient is 0.4072 with standard error 0.1352 and p-value $< .1\%$. The results are similar for the other patient outcome models. Table 12 summarizes the impact of SDU admission after ED transfer on the various patient outcomes for these patients.

Table 12 Estimated Effect of SDU Admission Following the ED (γ) on Patient Outcomes for High Severity Patients and Correlation between error terms (ρ) for the admission decision and patient outcomes: $N = 2,803$.

	With IV					Without IV
Outcome	γ (SE)	Predicted Outcome $\hat{P}_{SDUBusy=0}$	Predicted Outcome $\hat{P}_{SDUBusy=1}$	ρ (SE)	Test $\rho = 0$	γ (SE)
$Mortality$	0.75* (0.33)	23.64%	21.09%	-0.48* (0.18)	0.03	-0.05 (0.07)
$\log(HospRemLOS)$	0.45*** (0.12)	6.22	5.87	-0.57*** (0.07)	0.00	-0.32*** (0.04)
$HospReadm_{2w}$	1.27** (0.40)	18.28%	12.12%	-0.78* (0.20)	0.04	-0.08 (0.08)

Note. Standard error in parentheses. $^{+}(p < 10\%)$,$^{*}(p < 5\%)$,$^{**}(p < 1\%)$,$^{***}(p < 0.1\%)$.
High severity thresholds, t_2, defined by 95^{th} percentile of $\widehat{Tx}_i^*$.
Predicted outcome: $\hat{P}_{SDUBusy=0}$ - Average predicted outcome if the SDU was never busy
$\hat{P}_{SDUBusy=1}$ - Average predicted outcome if the SDU was always busy.
Predicted $HospRemLOS$ (days) is shown instead of $\log(HospRemLOS)$

For high severity patents, being admitted to the SDU appears to be associated with worse outcomes, as seen in the sign of SDU admission coefficient, which is positive and statistically significant in all outcome

measures. We again use our estimation results to predict the patients outcomes under two cases: (i) the SDU is never busy and (ii) the SDU is always busy. Our results suggest that being admitted to the hospital when the SDU is busy is associated with substantial degradation in patient outcomes. SDU bed availability is, on average, associated with an increase of 12.1% in in-hospital deaths, a 50.8% increase in hospital readmissions within 2 weeks, and .35 additional days in the hospital. As with the ICU cohort, the marginal effects estimates are based on the estimated treatment effect, which is averaged across all patients who comply with the instruments. Thus, one should interpret our results as demonstrating substantive and rigorous evidence to the statistical significance and direction of the treatment effect. Again, we see evidence of a correlation between the SDU admission decision and patient outcomes with the estimated correlation coefficient ρ being statistically different than 0. Based on the size of this cohort, we estimate eliminating SDU busy-ness would translate into annual increases of 71 deaths, 981 more hospital days, and 173 additional hospital readmissions within 2 weeks (weekly increases of 1.37 deaths, 18.9 hospital days, and 3.33 readmissions) aggregated across the 10 hospitals.

The results for LOS and hospital readmissions are consistent with Kim et al. (2015). Interestingly, we find that being admitted to the SDU is associated with an increase in mortality risk, while Kim et al. (2015) did not find an impact of non-ICU care on mortality. One potential explanation is that Kim et al. (2015) considered all patients admitted from the ED to a medical service, while we stratify our analysis to focus on only the high severity patients. As such, the results of Kim et al. (2015) may be distorted as SDU care may improve mortality risk for some patients within their cohort while also degrading mortality risk for other patients, thereby cancelling each other out. In contrast, since we focus on patients who are more likely to be admitted to the ICU (i.e. 54.90% compared to 11% in Kim et al. (2015)), we are able to provide a cleaner estimate.

6.3. Medium Severity

We now consider the impact of SDU admission on medium severity patients. For these patients, a busy SDU is associated with a decrease in likelihood of SDU admission. For the mortality model (when comparing admission to the SDU versus ICU or ward), the coefficient on $SDUBusy_i$ is -0.5503 with standard error 0.0377 and p-value $< 0.1\%$. The results are similar for the other patient outcome models.

Because of the aforementioned substantial heterogeneity within the medium severity cohort, when we run our models on this population, the results are not statistically significant and our instruments do not seem to be able to address potential endogeneity biases (see Table 13). The only result that is statistically significant is the remaining hospital LOS when considering whether a patient is admitted to the SDU versus ICU or ward. We're not sure what to make of this result due to the substantial heterogeneity of this group, as discussed before. We find that when the SDU is busy, patients are more likely to be rerouted – lower severity patients tend to go to the ward, while higher severity patients tend to go to the ICU. One possible

explanation for the statistically significant effect on hospital LOS is that the high severity patients who are admitted to the SDU are less sick that those who are not (and instead are sent to the ICU). Combining this with the lower severity patients who benefit from SDU care results in a statistically significant effect. Unfortunately, because of limitations in our data and the presence of unobservable factors, it is difficult to accurately assess the severity of the patients in this particular cohort, so we cannot be sure what is driving this result.

Table 13 Estimated Effect of SDU Admission Following the ED (γ) on Patient Outcomes for Medium Severity Patients and Correlation between error terms (ρ) for the admission decision and patient outcomes.

	With IV			Without IV
Outcome	γ (SE)	ρ (SE)	Test $\rho = 0$	γ (SE)
SDU vs. Ward: $N = 21,399$				
$Mortality$	-0.19 (0.25)	0.13 (0.15)	0.38	0.02 (0.04)
$\log(HospRemLOS)$	-0.02 (0.13)	0.10 (0.07)	0.14	0.08^{***} (0.01)
$HospReadm_{2w}$	0.32 (0.25)	-0.16 (0.15)	0.28	0.05^{+} (0.03)
SDU vs. ICU: $N = 12,043$				
$Mortality$	0.07 (0.20)	-0.18 (0.13)	0.15	-0.22^{**} (0.07)
$\log(HospRemLOS)$	-0.26 (0.08)	0.10 (0.06)	0.11	-0.18^{**} (0.03)
$HospReadm_{2w}$	-0.15 (0.19)	0.09 (0.12)	0.46	-0.01 (0.04)
SDU vs. ICU or Ward: $N = 25,928$				
$Mortality$	-0.23 (0.19)	0.10 (0.11)	0.37	-0.06^{+} (0.04)
$\log(HospRemLOS)$	-0.21^{**} (0.07)	0.17^{**} (0.05)	0.01	0.00 (0.01)
$HospReadm_{2w}$	-0.01 (0.25)	0.03 (0.15)	0.86	0.03 (0.03)

Note. Standard error in parentheses. $^{+}(p < 10\%)$,$^{*}(p < 5\%)$,$^{**}(p < 1\%)$,$^{***}(p < 0.1\%)$. Medium severity thresholds, t_1 and t_2, defined by 95^{th} and 60^{th} percentiles of $\widehat{Tx}_i^*$.

While it is challenging to consider this medium severity group and to interpret the results of our estimations for these patients, we highlight that this does not change the fundamental result of our work. In particular, we find that sending high severity patients from the ED to the SDU when the ICU is busy can have substantial adverse consequences. On the other hand, for some low severity patients care in the SDU may be beneficial with regard to some patient outcomes. These results indicate the need for further study of the use of the SDU for patients originating in the ED.

6.4. Robustness Checks

We now describe a number of robustness checks for our ED cohort. Due to the challenges associated with the medium severity patients, we focus our attention on the robustness of our results for the low and high severity patients. Our initial robustness checks are very similar to those we did with the ICU cohort. We tried different specifications of control variables, different time-lines for hospital readmission (within 1 week, 2 weeks, and 30 days after a patient is discharged from the hospital), and different ways to control for congestion during a patient's stay (max occupancy and no control). We also varied the definition of

our instrument(s) by varying the number of available beds from one to four, using indicator variables for whether the percentile occupancy level was above a threshold, a spline variable for the percentile occupancy level, as well as a time-varying capacity measure based on three 4-month periods. For our LOS models, we also considered specifications including patients with in-hospital death.

6.4.1. Low Severity Patients For the low severity patients, we found the results for LOS were very robust in sign, magnitude and statistical significance for all of these different specifications, including when we include patients with in-hospital death. The sign and magnitude for the in-hospital mortality results were also very robust. When the IV of SDU busy was based on very limited bed availability (i.e., ≤ 1 free beds or occupancy above the 95^{th} percentile), the coefficient estimates were not statistically significant. In these instances, we cannot reject the null hypothesis that the correlation between the two error terms in our model is 0, i.e. $\rho = 0$. This suggests that in these instances, the instrument is not able to adequately address the endogeneity biases in our data. We consistently found no statistically significant association between SDU admission and hospital readmissions within 1 week, 2 weeks, or 30 days.

6.4.2. High Severity Patients For the high severity patients, we found the results for $HospRemLOS$ were very robust in sign, magnitude and statistical significance for all of these different specifications. The mortality results were also quite robust. However, when defining $ICUBusy$ or $SDUBusy$ with a relatively low occupancy level (e.g. 80^{th} percentile), the statistical significance of the coefficients can drop to p-value $< 10\%$ or in some rare instances, is no longer statistically significant even at the 10% level. In these instances, we cannot reject the null hypothesis that the correlation between the two error terms in our model is 0, i.e. $\rho = 0$. This suggests that in these instances, the instrument is not able to adequately address the endogeneity biases in our data. Similarly, the hospital readmission results are always consistent in terms of sign and magnitude for these different specification. However, there are some instances when the results are not statistically significant. These instances correspond to when we cannot reject the null hypothesis that $\rho = 0$. This happens most frequently with the 1 week time to hospital readmission. We also found that all of our results were robust to including an additional instrumental variable based on the number of recent admissions to the ICU ($RecentAdmission_i^{ICU}$).

For the LOS models, we also considered the robustness of our results to including patients with in-hospital death. When including patients with in-hospital mortality in the high-severity ED Cohort, the sign of γ is negative (-1.83) and statistically significant at the p-value $< .01$ level. This raises questions as to the robustness of our LOS results for the high-severity group. However, we believe the main results as reported are more likely to be aligned with the true effect direction and size of SDU admission as it has been well established in the medical literature to exclude patients with in-hospital death for LOS models (e.g. Rapoport et al. (1996), Norton et al. (2007)).

6.4.3. Severity categorizations In our severity categorizations for the ED Cohort, we took a data-driven approach and used thresholds on $\widehat{Tx}_i^*$ to partition the patients into Low and High Severity groups. We varied the thresholds for these categorizations from the 45^{th} to 85^{th} percentile for low severity patients and from the 90^{th} to 97^{th} percentile for high severity patients. We then examined the robustness of our estimation results to these different thresholds.

6.4.4. Low Severity Patients As with our main specification, we do not find statistically significant results for the hospital readmission models. We find that the results for $HospRemLOS$ is very robust in magnitude and statistical significance to all of the different specifications of the low severity threshold. While the mortality results are robust to lowering the threshold, which reduces the sample size, we lose statistical significance when increasing the threshold above the 60^{th} percentile. This may be because as the sample size is increased, there are (moderately) high severity patients whose mortality risk may suffer with SDU admission are included in the cohort. When examining the LOS results more closely, we see that as the threshold is increased, the magnitude of the coefficient decreases, suggesting that the low severity cohort is including more patients for which SDU care is detrimental. Moreover, we cannot reject the null hypothesis that the correlation between our error terms is 0 ($\rho = 0$), which suggests that as the threshold increases, there are more non-compliers included in the cohort, making the instrument ineffective to address the endogeneity issues.

6.4.5. High Severity Patients In our readmission models, increasing the threshold for high severity patients above the 95^{th} percentile results in the regressions not converging. This is likely because the size of the cohort is being made smaller and smaller, and there are not enough samples to solve the FMLE optimization. These results suggest that the hospital readmission results are not very robust.

On the other hand, the $HospRemLOS$ results are quite robust to changes in the threshold. Similar to our observations for the low severity patients, we see that as the threshold decreases, the magnitude of the coefficient decreases. This may be because low severity patients who benefit from SDU care are entering into the high severity cohort as the threshold is decreased. A similar argument can be made for the mortality results. We find that when the threshold for high severity patients is less than the 93^{rd} percentile, the mortality and LOS results are no longer statistically significant. About 10% of the ED-medical patients are admitted to the ICU from the ED. Additionally, our admission model in Equation (6) incorporates an unobservable term ξ_i, such that if the observed latent variable $\widehat{Tx}_i^*$ plus ξ_i is above the threshold, the patient will be routed to the ICU. As the threshold gets closer to the 90^{th} percentile, there will be more spill over of patients for which SDU care is beneficial (instead of detrimental).

7. Conclusions and Managerial Insights

This paper studies the role of different levels of service for customers with uncertain needs. We examine this in a hospital setting where step-down units (SDUs) can be used to treat a variety of patients with very

different repercussions. We consider fundamental questions regarding the SDU: Does admitting a patient to the SDU improve or degrade patient outcomes? What is the magnitude of these effects? And, how does it vary across different types of patients? Our work represents an important first step towards answering these questions. We find that while the answer for patients discharged from the ICU (its original purpose) is fairly clear, for those admitted from the ED, it is quite nuanced – some patients will benefit, while others will not. Moreover, the impact of SDU care can be substantial, so it is essential to be able to carefully identify which patients are appropriate for SDU care. These findings suggest that while different levels of service may be used to serve multiple customer types, the costs and benefits associated with each level of service can be highly heterogeneous due to the different and sometimes uncertain needs of customers.

There are a number of opportunities for future work. Our empirical analysis relies on the variation in patient routings following ICU discharge or following admission from the ED due to SDU and/or ICU capacity constraints. Consequently, our estimates fundamentally apply to patients whose SDU admission comply with our instrumental variables. As such, it is not possible to make any statements about the impact of SDU care for patients whose care pathway is invariant to SDU (or ICU) bed availability. While it is difficult to extrapolate our results to make inferences on the precise magnitude of the effect of the SDU on individual patients, our results demonstrate strong evidence as to the directional impact of an SDU. Because SDUs go in and out of favor at individual hospitals, there may be opportunities for natural experiments to make such inferences without requiring an instrumental variable analysis. Alternatively, at a hospital system such as Kaiser Permanente, it might be possible to conduct a controlled randomized trial by randomizing which hospitals have SDUs. Of course, such a study would require substantial buy-in from hospital administrators and staff. Our empirical setting focuses on patients admitted to the hospital via the ED to a medical service. A number of studies in the medical literature consider the impact of SDUs on surgical patients (e.g. Eachempati et al. (2004)). The impact of SDU congestion is likely very different for surgical patients, where surgical procedures and schedules often dictate the precise care pathway for these patients. Hence, an alternative identification strategy is likely needed. In other service settings where experimentation is less costly (e.g. call-centers), randomized experiments may be a feasible approach to providing unbiased estimates of costs and benefits for different customers. That said, as routing of customers to servers in call-centers is often done by computers, the likelihood of biases due to unobservable factors may be lower, so an IV approach or randomized experiment may not be necessary in these settings.

Our approach to classifying patients could be used in other service settings where customers' needs are uncertain. For instance, in an increasing number of healthcare settings–including EDs, critical care, primary care and oncology, among others–Physician Assistants (PAs) and Nurse Practitioners (NPs) are used as lower cost alternatives to physicians (e.g. Hooker and McCaig (2001), Naylor and Kurtzman (2010), Hinkel et al. (2010), Doan et al. (2011), Gershengorn et al. (2011), Green et al. (2013), Gershengorn et al. (2016), etc.). PAs and NPs are trained in some, but not all, of the skills of physicians, raising important questions

as to which patient types and tasks can be safely and effectively handled by these healthcare professionals rather than by a physician. In such instances, a data-driven approach, such as the one taken in this paper, could be utilized to classify patients. It could also be useful in other service settings in which servers have different skill levels, such as call-centers or repair facilities, where customers are heterogeneous, but their needs are not known a priori. With the growing availability of customer information (e.g. demographics, spending habits, etc.), a data-driven approach to customer segmentation may be useful.

From a stochastic modeling point of view, there are a number of directions that could build upon this work. From the healthcare operations management standpoint, it would be interesting to study optimal control policies regarding where to transfer patients from the ED or following ICU discharge in the presence of an SDU. This would provide a system-level view that would capture the potential benefits of an SDU, including externalities on other patients, beyond the estimates of individual patients estimated in this work. This would complement the growing body of work which examines how to make patient transfer decisions from the ED as well as inpatient units (e.g. Mandelbaum et al. (2012), Barz and Rajaram (2015), Samiedaluie et al. (2017), Dai and Shi (2017), Kilinc et al. (2016), etc.). Additionally, one could consider how to determine the capacity of the SDU relative to the ICU and general ward given patient mix and arrival rates. One factor which would significantly impact this decision is whether to restrict use of the SDU to be a true step-down versus allowing admission of patients from non-ICU units, such as the ED. This work quantifies the impact of lack of access to care for various patient types and could be used to set performance benchmarks or to calibrate a cost minimization framework when determining bed capacity (e.g Yankovic and Green (2011), Yom-Tov and Mandelbaum (2014), Best et al. (2015), Armony et al. (2017), etc.). Analysis of these questions can also provide insights into how to utilize nested levels of service and routing policies in other types of service settings with heterogeneous customers (e.g. call-centers (Gans et al. 2003), retail stores, restaurants, etc.). For instance, it would be interesting to understand what factors such as number of customer types, differences in customer demand, service times, and costs impact the optimal number of levels of service and the optimal capacity to allocate to each level.

In understanding the benefits of the nested structure, an interesting tradeoff arises where increasing the number of levels reduces pooling benefits and may increase delays or reroutings. On the other hand, increasing the number of levels of care allows for more specialization that may result in efficiencies that reduce service times and improve outcomes. In many nested service systems, including the hospital situation studied here, it would be interesting to examine the potential tradeoffs between pooling and efficiency, similar in spirit to the work in Song et al. (2015) and how this would impact the allocation of servers (e.g. beds) to different levels of care (e.g. Best et al. (2015)).

Acknowledgments

We gratefully acknowledge the editors and reviewers for their many helpful suggestions and comments, which have greatly improved this paper. We thank Marla Gardner, John Greene, and Benjamin Turk for their help in preparing the

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. The work by Carri W. Chan and Gabriel Escobar was supported in part by NSF/AHRQ grant number CMMI-1233547. Dr. Escobar was also supported by The Permanente Medical Group, Inc., and Kaiser Foundation Hospitals, Inc.

References

Armony, M., C. W. Chan, B. Zhu. 2017. Critical care capacity management: Understanding the role of a step down unit. *Working Paper, Columbia Business School* .

Baker, D., D. Einstadter, S. Husak, R. Cebul. 2003. Changes in the use of do-not-resuscitate orders after implementation of the patient self-determination act. *Journal of General Internal Medicine* **18** 343 – 349.

Barz, C., K. Rajaram. 2015. Elective patient admission and scheduling under multiple resource constraints. *Production and Operations Management* **24**(12) 1907–1930.

Batt, Robert J., Christian Terwiesch. 2017. Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science, to appear* .

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

Berry Jaeker, J. A., A. L. Tucker. 2016. Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science* .

Best, T., B. Sandikci, D. Eisenstein, D. Meltzer. 2015. Managing hospital inpatient bed capacity through partitioning care into focused wings. *Manufacturing & Service Operations Management* **17**(2) 157–176.

Brown, S., S. Ratcliffe, S. Halpern. 2013. An empirical derivation of the optimal time interval for defining ICU readmissions. *Medical Care* **51** 706 – 714.

Byrick, R. J., J. D. Power, J. O. Ycas, K. A. Brown. 1986. Impact of an intermediate care area on ICU utilization after cardiac surgery. *Critical care medicine* **14** 869–872.

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

Centers for Medicare & Medicaid Services. 2016. National health expenditure tables. URL `www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/Tables.zip`. Accessed, 06/07/2016.

Chan, C. W., L. V. Green, Y. Lu, N. Leahy, R. Yurt. 2013. Prioritizing burn-injured patients during a disaster. *Manufacturing & Service Operations Management* **15**(2) 170–190.

Christensen, S., M. Johansen, C. Christiansen, R. Jensen, S. Lemeshow. 2011. Comparison of charlson comorbidity index with saps and apache scores for prediction of mortality following intensive care. *Journal of Clinical Epidemiology* **3** 203–211.

Coopersmith, C., H. Wunsch, M. Fink, W. Linde-Zwirble, K. Olsen, M. Sommers, K. Anand, K. Tchorz, D. Angus, C. Deutschman. 2012. A comparison of critical care research funding and the financial burden of critical illness in the United States. *Critical Care Medicine* **40** 1072 – 1079.

Dai, J. G., P. Shi. 2017. Inpatient bed overflow: An approximate dynamic programming approach. *Working Paper* .

Doan, Q., V. Sabhaney, N. Kissoon, S. Sheps, J. Singer. 2011. A systematic review: the role and impact of the physician assistant in the emergency department. *Emergency Medicine Australasia* **23**(1) 7–15.

Doran, K., K. Ragins, A. Iacomacci, A. Cunningham, K. Jubanyik, G. Jenq. 2013. The revolving hospital door: hospital readmissions among patients who are homeless. *Medical Care* **51** 767 – 773.

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

Edie, Leslie C. 1954. Traffic delays at toll booths. *Journal of the operations research society of America* **2**(2) 107–138.

Escobar, G. J., M. N. Gardner, J. D. Greene, D. Draper, P. Kipnis. 2013. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. *Medical Care* **51**(5) 446–453.

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

Freeman, M., N. Savva, S. Scholtes. 2016. Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science* .

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141.

Gelman, A., J. Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gershengorn, H. B., H. Wunsch, R. Wahab, D. Leaf, D. Brodie, G. Li, P. Factor. 2011. Impact of nonphysician staffing on outcomes in a medical icu. *CHEST Journal* **139**(6) 1347–1353.

Gershengorn, H. B., Y. Xu, C. W. Chan, M. Armony, M. N. Gong. 2016. The impact of adding a physician assistant to a critical care outreach team. *PloS one* **11**(12) e0167959.

Gibbons, J., S. Chakraborti. 2011. *Nonparametric Statistical Inference*. 5th ed. Boca Raton, FL: Chapman & Hall/CRC.

Green, Linda. 1985. A queueing system with general-use and limited-use servers. *Operations Research* **33**(1) 168–182.

Green, Linda V, Sergei Savin, Yina Lu. 2013. Primary care physician shortages could be eliminated through use of teams, nonphysicians, and electronic communication. *Health Affairs* **32**(1) 11–19.

Greene, W. H. 2012. *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.

Gurvich, I., M. Armony, A. Mandelbaum. 2008. Service-level differentiation in call centers with fully flexible servers. *Management Science* **54**(2) 279–294.

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

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

Gustafsson, U. O., J. Hausel, A. Thorell, O. Ljungqvist, M. Soop, J. Nygren, Enhanced Recovery After Surgery Study Group. 2011. Adherence to the enhanced recovery after surgery protocol and outcomes after colorectal cancer surgery. *Archives of Surgery* **146**(5) 571–577.

Harding, A. D. 2009. What can an intermediate care unit do for you? *Journal of Nursing Administration* **39** 4 – 7.

Hasija, Sameer, Edieal J Pinker, Robert A Shumsky. 2005. Staffing and routing in a two-tier call centre. *International Journal of Operational Research* **1**(1-2) 8–29.

Hinkel, J. M., J. L. Vandergrift, S. J. Perkel, M. B. Waldinger, W. Levy, F. M. Stewart. 2010. Practice and productivity of physician assistants and nurse practitioners in outpatient oncology clinics at national comprehensive cancer network institutions. *Journal of oncology practice* **6**(4) 182–187.

Hooker, R. S., L. F. McCaig. 2001. Use of physician assistants and nurse practitioners in primary care, 1995–1999. *Health Affairs* **20**(4) 231–238.

Huang, JunfeiJ., B. Carmeli, A. Mandelbaum. 2015. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research* **63**(4) 892–908.

Huh, W. T., N. Liu, V.-A. Truong. 2013. Multiresource allocation scheduling in dynamic environments. *Manufacturing & Service Operations Management* **15**(2) 280–291.

Jayes, R., J. Zimmerman, D. Wagner, E. Draper, W. Knaus. 1993. Do-not-resuscitate orders in intensive care units. current practices and recent changes. *JAMA* **270** 2213 – 2217.

Jerath, K., A. Kumar, S. Netessine. 2015. An information stock model of customer behavior in multichannel customer support services. *Manufacturing & Service Operations Management* **17**(3) 368–383.

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

Keenan, S. P., W. J. Sibbald, K. J. Inman, D. Massel. 1998. A systematic review of the cost-effectiveness of noncardiac transitional care units. *Chest* **113** 172 – 177.

Kilinc, D., S. Saghafian, S. Traub. 2016. Dynamic assignment of patients to primary and secondary inpatient units: Is patience a virtue. *Working paper, Harvard University* .

Kim, S. H., C. W. Chan, M. Olivares, G. Escobar. 2015. ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science* **61**(1) 19–38.

Kuntz, L., R. Mennicken, S. Scholtes. 2015. Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science* **61**(4) 754–771.

Kuntz, L., S. Scholtes, S. Sülz. 2016. Separate & concentrate: Accounting for process uncertainty in the design of regional hospital systems. *Working Paper, Cambridge University* .

Lassen, K., M. M. E. Coolsen, K. Slim, F. Carli, J. E. de Aguilar-Nascimento, M. Schäfer, R. W. Parks, K. C. H. Fearon, D. N. Lobo, N. Demartines, M. Braga, O. Ljungqvist, C. H. C. Dejong. 2013. Guidelines for perioperative care for pancreaticoduodenectomy: Enhanced recovery after surgery (eras) society recommendations. *World Journal of Surgery* **37**(2) 240–258.

Lee, Hsiao-Hui, Edieal J Pinker, Robert A Shumsky. 2012. Outsourcing a two-level service process. *Management Science* **58**(8) 1569–1584.

Luo, J., J. Zhang. 2013. Staffing and control of instant messaging contact centers. *Operations Research* **61**(2) 328–343.

Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Operations Research* **53**(2) 242–262.

Maguluri, S. T., R. Srikant, L. Ying. 2012. Stochastic models of load balancing and scheduling in cloud computing clusters. *INFOCOM, 2012 Proceedings IEEE*. IEEE, 702–710.

Mandelbaum, A., P. Momcilovic, Y. Tseytlin. 2012. On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science* **58**(7) 1273–1291.

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

Mbongo, C., P. Monedero, F. Guillen-Grima, M. Yepes, M. Vives, G. Echarri. 2009. Performance of saps3, compared with apache ii and sofa, to predict hospital mortality in a general ICU in southern europe. *European Journal of Anaesthesiology* **26** 940–945.

Miller, T. E., J. K. Thacker, W. D. White, C. Mantyh, J. Migaly, J. Jin, A. M. Roche, E. L. Eisenstein, R. Edwards, K. J. Anstrom, R. E. Moon, T. J. Gan, Enhanced Recovery Study Group. 2014. Reduced length of hospital stay in colorectal surgery after implementation of an enhanced recovery protocol. *Anesthesia & Analgesia* **118**(5) 1052–1061.

Nasraway, S. A., I. L. Cohen, R. C. Dennis, M. A. Howenstein, Diana K. Nikas, J. Warren, S. K. Wedel. 1998. Guidelines on admission and discharge for adult intermediate care units. *Critical care medicine* **26**(3) 607–610.

Naylor, M. D., E. T. Kurtzman. 2010. The role of nurse practitioners in reinventing primary care. *Health Affairs* **29**(5) 893–899.

Norton, S.A., L.A. Hogan, R.G. Holloway, H. Temkin-Greener, M.J. Buckley, T.E. Quill. 2007. Proactive palliative care in the medical intensive care unit: Effects on length of stay for selected high-risk patients. *Crit Care Med* **35** 1530–1535.

Ouanes, I., C. Schwebel, A. Francais, C. Bruel, F. Philippart, A. Vesin, L. Soufir L, C. Adrie, M. Garrouste-Orgeas, J. Timsit, B. Misset. 2012. A model to predict short-term death or readmission after intensive care unit discharge. *Journal of Critical Care* **27** 422 e1 – e9.

Rapoport, J., D. Teres, S. Lemeshow. 1996. Resource use implications of do not resuscitate orders for intensive care unit patients. *Am J Respir Crit Care Med* **153** 185–190.

Samiedaluie, S., B. Kucukyazici, V. Verter, D. Zhang. 2017. Managing patient admissions in a neurology ward. *Operations Research, to appear* .

Shi, P., M. Chou, J. Dai, D. Ding, J. Sim. 2014. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science, to appear* .

Shumsky, Robert A, Edieal J Pinker. 2003. Gatekeepers and referrals in services. *Management Science* **49**(7) 839–856.

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

Song, Hummy, Anita L Tucker, Karen L Murrell. 2015. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* **61**(12) 3032–3053.

Stacy, K. M. 2011. Progressive care units: Different but the same. *Critical Care Nurse* **31** 77 – 83.

Stowell, A., P.-G. Claret, M. Sebbane, X. Bobbia, C. Boyard, R. G. Grandpierre, A. Moreau, J.-E. de La Coussaye. 2013. Hospital out-lying through lack of beds and its impact on care and patient outcome. *Scand J Trauma Resusc Emerg Med* **21** 17.

Strand, K., H. Flaatte. 2008. Severity scoring in the ICU: a review. *Acta Anaesthesiologica Scandinavica* **52** 467 – 478.

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

Tezcan, T., B. Behzad. 2012. Robust design and control of call centers with flexible interactive voice response systems. *Manufacturing & Service Operations Management* **14**(3) 386–401.

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

Tezcan, T., J. Zhang. 2014. Routing and staffing in customer service chat systems with impatient customers. *Operations research* **62**(4) 943–956.

Thiele, R. H., K. M. Rea, F. E. Turrentine, C. M. Friel, T. E. Hassinger, B. J. Goudreau, B. A. Umapathi, I. L. Kron, R. G. Sawyer, T. L. Hedrick, T. L. McMurry. 2015. Standardization of care: Impact of an enhanced recovery protocol on length of stay, complications, and direct costs after colorectal surgery. *Journal of the American College of Surgeons* **220**(4) 430 – 443.

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

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

Wang, G., J. Li, W. J. Hopp, F. L. Fazzalari, S. Bolling. 2016. Using patient-centric quality information to unlock hidden health care capabilities. *Working Paper, Ross Business School* .

Yankovic, N., L. V. Green. 2011. Identifying good nursing levels: A queuing approach. *Operations research* **59**(4) 942–955.

Yom-Tov, G., A. Mandelbaum. 2014. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* **16** 283 – 299.

Zimmerman, J. E., D. P. Wagner, W. A. Knaus, J. F. Williams, D. Kolakowski, E. A. Draper. 1995. The use of risk predictions to identify candidates for intermediate care units. *Chest* **108** 490 – 499.

Appendix A: Supplementary Tables

Table 14 Control variables for patient characteristics and hospital care

Variable	Description	ICU Cohort	ED Cohort
Gender	Dummy variable: Males were coded 1 and females 0	✓	✓
Age	Continuous variable: Coded as piecewise linear spline variables with knots at its 50^{th} and 80^{th} percentiles (65 and 81)	✓	✓
LAPS2	Laboratory-based Acute Physiology Score; measures physiologic derangement at admission and is mapped from 14 laboratory test results, such as arterial pH and white blood cell count, obtained 72 hours preceding hospitalization to an integer value that ranges from 0 to 262 in our data set(higher scores indicate poorer condition); coded as piecewise linear spline variables with knot at its 50^{th} and 80^{th} percentiles (94 and 134)	✓	✓
COPS2	Comorbidity Point Score; measures the chronic illness burden and is based on 41 comorbidities, such as diabetes, to which patients are categorized using outpatient and inpatient data from the 12 months preceding hospitalization. It ranges from 0 to 267 in our data set, a higher score indicates a higher comorbid illness burden, it was coded as piecewise linear spline variables with knot at its 50^{th} and 80^{th} percentiles (33 and 87)	✓	✓
SAPS3	Simplified Acute Physiology Score; measures the severity of illness and predict vital status at hospital discharge based on ICU admission data. SAPS3 score is associated with each ICU admission and is calculated based on data obtained within on hour of ICU admission. SAPS3 ranges from 14 to 100 in our data set; coded as piecewise linear spline variables with knot at its 50^{th} and 80^{th} percentiles (52 and 61)	✓	
Admitting diagnosis	A way of classifying ICD9 codes. This clinical classification system was developed by HCUP and buckets ICD9's into about 200 groups. A further grouping of the variable HCUP developed by Gabriel Escobar to condense the HCUP grouping into 38 groups so it could be used in a similar fashion as PRIMCOND3.	✓	✓
Seasonality	Month/day-of-week/time-of-day; Category variable for each month and day-of-week. For time-of-day, we use category variables for nurse shifts happening three times a day at 7am, 15pm, and 23pm.	✓	✓
Previous unit	Category variable to track inpatient unit a patient is admitted to immediately before ICU admission.	✓	
LOS before ICU	Continuous variable that is the total length-of-stay (hrs) prior to the ICU admission. It measures how long a patient has been in hospital before being admitted to the ICU, coded as piecewise linear spline variables with knot at its 50^{th} and 80^{th} percentiles (2 and 31).	✓	
ICU LOS	Continuous variable that is the length-of-stay (hrs) at the first ICU. It measures how long a patient has been taking care of at ICU, coded as piecewise linear spline variables with knot at its 50^{th} and 80^{th} percentiles (38 and 83).	✓	
ED LOS	Continuous variable that is the length-of-stay (hrs) at the first ED. It measures how long a patient has been taking care of at ED.		✓

NOTE: To account for potential non-linear effects of some of the variables used to control for patient severity, we code them as piecewise linear spline variables.

Table 15 ICU Cohort: Percentage of patients who are discharged from ICU when SDU is busy

		% when number of available SDU beds			
Hosp	SDU Size	≤ 1	≤ 2	≤ 3	≤ 4
1	24	0.93	3.57	7.80	12.17
2	25	0.66	2.95	7.54	12.46
3	14	0.56	7.94	24.29	45.63
4	19	3.17	12.68	27.07	41.59
5	24	0.28	1.54	3.93	7.87
6	19	0.82	3.34	6.76	15.37
7	20	0.00	2.84	16.74	36.77
8	27	2.81	9.34	18.80	31.74
9	11	9.76	37.72	63.94	80.34
10	32	0.34	2.66	6.19	12.71
All hosp		2.52	10.64	21.70	34.00

Table 16 Post-ICU location

Unit	Frequency	Percent
Ward	7,226	54.77
SDU	3,832	29.05
Death in ICU	985	7.47
Out of hospital (alive)	1,150	8.72
Total	13,193	100

Table 17 Effect of SDU Admission Following the ED (γ) on $HospRemLOS$ When Including Patients with In-Hospital Death

Cohort	γ (SE)	Predicted Outcome		ρ (SE)	Test
		$\hat{P}_{SDU}$	$\hat{P}_{ICU}/\hat{P}_{Ward}$		$\rho = 0$
ED Cohort - High Severity	-1.83** (0.07)	1.33%	8.31% ($\hat{P}_{ICU}$)	0.81*** (0.02)	0.00
ED Cohort - Low Severity	-0.20*** (0.04)	2.38	2.92 ($\hat{P}_{Ward}$)	0.18*** (0.03)	0.00

Note. Standard error in parentheses. $^{+}(p < 10\%)$,$^{*}(p < 5\%)$,$^{**}(p < 1\%)$,$^{***}(p < 0.1\%)$.
Predicted outcome: $\hat{P}_{SDU}$ - Average predicted outcome if all patients could be routed to the SDU and $\hat{P}_{ICU}$ ($\hat{P}_{Ward}$) if no SDU and everyone is routed to the ICU (Ward).
Predicted $HospRemLOS$ (days) is shown instead of $\log(HospRemLOS)$