Structural Estimation of Intertemporal Externalities on ICU Admission Decisions

Yiwen Shen
Decision, Risk, and Operations, Columbia Business School, yshen21@gsb.columbia.edu

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

Fanyin Zheng
Decision, Risk, and Operations, Columbia Business School, fanyin.zheng@columbia.edu

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

Service systems' behavior can be affected by multiple factors. In the case of intensive care units (ICUs), which admit patients from four primary loci (the emergency department (ED), scheduled patients, planned transfers from other ICUs, and unplanned transfers), it is known that admission rates of some patients decrease as occupancy increases. It is also known that, for at least some conditions, ICU admission is not just a function of patients' illness, and that a significant proportion of the variation in ICU admission rates is due to hospital, not patient, factors. In this paper, we employ two years of data from patients admitted to 21 Kaiser Permanente Northern California ICUs from the ED. We quantify the variation in ICU admission from the ED under varying degrees of ICU and ED occupancy. We find that substantial heterogeneity in admission rates is present, and that it cannot be explained either by patient factors or occupancy levels alone. We use a structural model to understand the extent that intertemporal externalities could account for some of this variation. Using counterfactual simulations, we find that, if hospitals had more information regarding their behaviors, and if it were possible to alter hospital admission processes to incorporate such information, hospitals could achieve greater efficiency safely.

Key words: healthcare, structural estimation, empirical operations management, intensive care unit, behavioral operations

1. Introduction

In managing service systems, one often has to consider the short term and long term implications of each decision. This is particularly relevant when the system is operated in a resource constrained environment. Indeed, managers often need to balance carefully between providing immediate access to service in order to address immediate considerations such as minimizing waiting time and the impact such actions may have on the ability to provide service to future customers. In this work, we study how hospitals balance this trade-off between short-term versus long-term considerations when making Intensive Care Unit (ICU) admission decisions and how this behavior impacts various system performance metrics.

The tension between the short-term and long-term considerations is particularly relevant in health care settings, which are often resource constrained and where timely access can have substantial clinical impli-

cations. We study this balancing act in the case of ICUs where one could think of the short-term horizon as the next two to six hours versus the longer-term horizon of the next 24 hours. ICUs are specialized inpatient units which provide the highest level of care for the most critically ill patients. They are expensive medical resources with the cost of patient care being several times higher than regular wards (see, e.g., Halpern and Pastores (2015)). In the United States, ICU costs comprise 17.4%-39% of total hospital costs (Coopersmith et al. (2012)) but less than 10% of inpatient beds. Additionally, ICUs often operate at high occupancy levels (see e.g., Green (2002), Halpern and Pastores (2010, 2015)). Consequently, the ICU is often identified as a critical process bottleneck; ICU congestion can have serious repercussions on patient flow and patient outcomes (see, e.g., Kc and Terwiesch (2012), Allon et al. (2013), Kim et al. (2015), Chan et al. (2016)). All these factors make ICU beds a key resource in a hospital that must be managed effectively.

In this work, we study ICU admissions for patients admitted to the hospital from the emergency department (ED) in a large hospital network. When a patient arrives in the ED, an ED physician will stabilize the patient and assess the patient's needs. If the patient may require ICU admission, an intensivist is called for a consultation. The ultimate outcome in terms of the patient's disposition is the result of the coordination of many people including the ED physician, the intensivist, and hospital administrators, as well as various aspects of the hospital system response which could include temporizing actions or adding additional resources (e.g. floating nurses). For conciseness, throughout this manuscript, we will often refer to this composite decision maker which is comprised of complex interactions of clinicians as well as the system response as the 'system' or the 'hospital'. When considering the admission of patients to the ICU, the system must balance the benefit of providing ICU care to the current ED patient versus the potentially negative impact of increased ICU occupancy level as a result of the admission.

As the ICU provides the highest level of care among all hospital units, swift admission generally benefits the patient. However, the admitted patient will occupy an ICU bed, which may lead to ICU congestion and restrict access to ICU care for future, perhaps more severe, patients. Moreover, the system might also take other actions in response to the increasing occupancy in the ICU which might affect the admission of the current patient. For example, the system might increase the efforts to stabilize the less severe patients to prepare for the increasing needs for ICU beds in advance. In addition, the system might need to obtain additional resources such as nursing staff and respiratory care technicians in order to prepare for the increasing occupancy in the ICU, which might also slow down the admission of current patients. In other words, the system's admission response not only impacts current patients, but also changes the system state which affects future patients. This trade-off introduces *intertemporal externalities* on the ICU admission decisions, i.e., both current and future 'payoffs'[1] matter. While there are undoubtedly many factors which influence the

[1] Note that throughout the manuscript, we will use the economic term 'payoff' to capture clinical, operational, and (possibly) financial impacts of each decision. This term is not intended to imply there are explicit financial payoffs associated with each decision.

ICU admission decision, our goal in this manuscript is to develop an understanding of how much variation in which hospitals react to ICU congestion could, in part, be explained by intertemporal externalities.

A priori, it is not clear how much the system takes actions in advance or thinks forward in response to the increasing ICU occupancy as there are supporting arguments for responding early as well as responding late. On one hand, the system's primary goal is to provide the best care for the patient. Thus, if ICU care will benefit the patient, the system should be more likely to admit the patient as long as there is an available bed in the ICU. This would suggest that the system behaves in a manner which does not respond much in advance to the increasing ICU occupancy. On the other hand, there is empirical evidence indicating that the system may reduce the likelihood of ICU admissions in order to 'save' ICU capacity for future, potentially sicker, patients, or take other actions in advance to prepare for the higher needs of ICU beds in the future. Such behavior is particularly noticeable when the ICU occupancy is high – i.e., when the remaining capacity and resources for future patients is limited. When the ICU is congested, the system is less likely to admit ED patients across all severity levels (Kim et al. 2015), and patients tend to board longer in the ED before being admitted to the ICU (Chan et al. 2016). These empirical findings suggest that the hospital might indeed take into consideration the ability to service future patients and the efforts needed to obtain additional resources in preparation for future admissions. Thus, at least to some extent, the system behaves in a manner that is consistent with forward-thinking and accounts for the long-term when making ICU admission decisions. Moreover, given the complexity of the admission process and the various constraints in the system, the system might not be fully aware of how much forward-thinking there is in the admission process. This paper will shed light on this debate by utilizing outcome data on which hospital unit patients are admitted to in order to empirically measure the degree of forward-thinking behavior embedded in the system's ICU admission decisions. While there are undoubtedly many factors which influence the flow of patients, we focus primarily on isolating the impact of one of these factors – the degree of forward-thinking behavior.

The degree of the system's forward-thinking behavior is hard to empirically quantify directly. First, to the best of our knowledge, there is limited data on the decision-making process itself. Rather, we utilize observational data on the final decision and some system information which could (or could not) influence the decision without explicit information into the decision-making process. Some crude measures can be constructed from system statistics regarding how the system reacts to ICU occupancy in their admission decisions. For example, if we see a larger drop in admission probability or a greater increase in ED boarding time (i.e. the time spent waiting in the ED once an admission decision has been made) as the ICU becomes congested, this would suggest the system is more active in saving ICU beds or obtaining additional resources for future patients and thus consider the long-term when making decisions. However, such crude measures have several drawbacks. First, the hospital's reaction to ICU occupancy is also affected by various other factors, such as the ED and ICU capacities, the composition of the entire patient cohort treated at the hospital, as well as the arrival rate and average length-of-stay (LOS) of patients. Thus, while such measures

can provide some evidence of the hospital's internalization of the short versus longer term, they cannot do so in an accurate and straightforward way. Consequently, it is also difficult to use these measures for comparing behavior across different hospitals, as hospitals may differ dramatically in their sizes, workloads, resources, and patient cohorts. More importantly, these indirect measures do not allow us to conduct counterfactual studies to analyze the impact of the forward looking behavior on various hospital performance metrics and patient outcomes.

To handle these difficulties, we take a structural estimation approach to measure the degree to which the system is forward-thinking from observed data and quantify the impact of such behavior on key system performance metrics. In particular, we leverage a new econometric approach on a large retrospective data set (more than 300,000 hospitalizations from 21 hospitals) to estimate the *inter-temporal discount factor* in a dynamic discrete choice model for ICU admission decisions. In this model, the hospital may consider the longer-term system dynamics and performance and take relevant actions in advance, to maximize their expected accumulative utility by choosing the optimal action for the current patient. Their actions impact the current period utility, but also affect future utility through the transitions of the system states. The discount factor denotes the relative weight of the next period's utility in the hospital's objective function, and, thus, determines how they balance current and future payoffs when making admission decisions. We want to emphasize that our interpretation of the discount factor in the model includes the following forward-thinking behaviors. First, the system might actively save beds for future patients when the ICU occupancy increases. Second, the system might also spend efforts to stabilize less severe patients to prepare for the increasing needs for ICU beds in the future. Finally, the system sometimes needs to spend efforts to obtain additional resources in order to admit more patients into the ICU, which might also slow down admissions. We consider all of these possible behaviors as forward-thinking or longer-term which are captured by the discount factor in the model. We acknowledge that the system might not be fully aware of their forward-thinking behavior or be able to choose the level of forward-thinking given the complexity of the admission process and various constraints it faces. In other words, the discount factor in our setting should be interpreted as the 'perceived discount factor'. For conciseness and consistency with the rest of the literature, we use the term 'discount factor' and 'perceived discount factor' interchangeably throughout the rest of the paper. A larger perceived discount factor implies that the system is more forward-thinking, while a smaller perceived discount factor implies they are less forward-thinking, and care more about the short-term.

We estimate the perceived discount factor and the costs associated with the system's actions *jointly* from the data. Most empirical studies with dynamic models in the literature assume the discount factor to be known and estimate the cost parameters (see, e.g., Rust (1987), Bajari et al. (2007), and Mehta et al. (2017)). Moreover, the discount factor is generally set at a large level that is close to 1 – e.g., 0.90 or even 0.99. In other words, most of these works make an implicit assumption that the agent is relatively forward-looking. However, this assumption lacks empirical support and formal justifications. More importantly, the

discount factors in dynamic models can vary dramatically depending on the context of the problem, or as a consequence of the behavioral variation of the decision makers. In the context of ICU admissions, using a prespecified perceived discount factor can be inappropriate, as there are conflicting arguments for both the immediate and longer-term aspects of the hospital's behaviors. The near versus long term behavior captured by the perceived discount factor may vary across hospitals as well. To address these issues, we identify the perceived discount factor from observed data instead of anchoring our analysis with an assumed, fixed perceived discount factor.

Various literature show that the dynamic model cannot be identified from choice data without further restriction (see e.g., Manski (1993), Rust (1994), Magnac and Thesmar (2002)), which is the primary reason that most empirical studies assume a known discount factor. As shown in Rust (1994), the non-identification of dynamic models stems from the existence of observationally equivalent structures: different combinations of discount factors and costs can lead to the same choice probabilities for all states, making the discount factor and costs not jointly identifiable. In this paper, we circumvent this difficulty by leveraging the new econometric approach developed by Komarova et al. (2018). They establish a one-dimensional criterion function for the identification of an empirical model with linear payoff specification. We apply their result to our setting and then use the observed data to show our dynamic discrete choice model can be identified, i.e., the discount factor and costs can be jointly estimated.

With the estimated structural model, we quantify the impact of the degree to which the hospital balances the short-term versus long-term considerations on key performance metrics of the ICU. This behavior can have a multifaceted impact on patient and system outcomes. By having a larger discount factor (longer-term focused behavior), the hospital can reduce ICU congestion by saving beds in advance, but this may also lead to longer boarding times for ED patients. It is well known that being myopic is not always the best thing to do when optimizing over a longer horizon. While it is unlikely that the hospital focuses only on the short-term when making ICU admissions (e.g., with discount factor being zero in the dynamic model), we explore the potential impact on hospital flow and patient outcomes if it were possible to increase the decision-making team's discount factor *only*. To provide insights for this question, we conduct counterfactual analyses to quantify the impact of increasing the degree of longer-term considerations on ICU congestion and ED boarding.

From a broader perspective, our study builds an understanding of how human servers make decisions in a resource limited environment. We focus on the particular aspect of how servers internalize the trade-off between near versus longer term considerations. Our model can be viewed as a finite buffer queueing system. ED patients are admitted to the ICU or medical-surgical ward according to the system's decisions. Consequently, the transition of the system state is determined by random arrivals and departures as well as the behaviors of the decision-making team. In most queueing systems, customers enter service immediately as long as there are available servers. However, in our setting, there are limited clear and/or objective criteria

for which patients to admit to the ICU, as well as when to admit them. These decisions inevitably depend on a patient's severity level and the system state (e.g. the availability of a bed), but also hinge on the behaviors of the hospital. In this work, we focus on how the hospital internalizes the intertemporal externalities on ICU admissions. This introduces a behavioral perspective to the queueing system which we will see has a substantial impact on system dynamics. Our main contributions can be summarized as follows:

- We build a structural model of the system's ICU admission decisions that incorporates both consideration of the current patients as well as the decision's impact on the system's capacity to serve future patients. The structural model accounts for the observed patient severity level, unobserved patient characteristics that are only available to the hospital, random arrivals and departures, as well as the system capacity constraints. We measure the hospital's degree of near versus longer-term looking behavior by the discount factor in the model. Leveraging recent identification results in the econometrics literature, we demonstrate that we can jointly estimate the discount factor and cost parameters from our data.
- We estimate the structural model with an extensive data set consisting of more than 300,000 hospitalizations from 21 Kaiser Permanente hospitals. We find that while, on average, the hospitals by and large tend to focus more on the near-term, there is large heterogeneity in the estimated discount factors across hospitals – i.e., some account more for the near-term, while others consider more longer-term impacts. In contrast to the standard approach in the literature, in the context of ICU admission decisions, it is inappropriate to assume a prespecified level for the discount factor without empirical support. We show that the heterogeneity estimated from the structural model is consistent with evidence from a reduced-form multinomial logistic regression. Moreover, we find that hospitals with busier ICUs tend to have larger discount factors.
- With our estimated structural model, we perform counterfactual studies to evaluate the impact of the hospital's behavior on ICU performance metrics. We conduct the analyses in the following two steps. First, we measure the benefit of adding one ICU bed. We show that adding one ICU bed leads to significant reduction in ICU congestion. For several hospitals, the annual reduction translates to 20 fewer days under a high congestion state and ten fewer patients who cannot be accepted upon arrival because the ICU is full. In the second step, we quantify the impact of increasing the discount factor from its current estimated level to 0.9 – i.e., changing the hospital's behavior so they use longer-term discounting – and compare its impact to the benefit of adding one ICU bed computed in the first step. We show that for some hospitals, this purely behavioral change can lead to reductions in ICU congestion that are comparable to the costly act of adding one ICU bed. This highlights the importance of understanding the behavior of the decision-making team in ICU capacity management.

The rest of the paper is organized as follows. We conclude this section with a brief literature review. Section 2 describes the setting and data. Section 3 develops a reduced-form model to estimate factors which

impact the hospital's decisions and provides suggestive evidence that they indeed internalize the intertemporal externalities on ICU admission decisions. Section 4 develops the main structural model to measure the degree of near versus longer term looking behavior of the system, and establishes the identification results and algorithmic approach for estimation. Section 5 provides the estimation results and conducts counterfactual studies. Section 6 concludes the paper and discusses future research directions.

1.1. Literature Review

Our work is related to four main streams of literature: (1) empirical healthcare operations management, particularly those related to ICU decisions; (2) structural estimation in operations management; (3) behavioral operations; and (4) econometrics tools for identifying dynamic models.

There has been a growing literature in the field of empirical healthcare management that examines patient flow in hospitals. A number of works study the effect related to admission and/or rerouting decisions as we do in this work. For instance, Shmueli et al. (2003), Edbrooke et al. (2011) and Kim et al. (2015) study the impact of ICU admissions on patient outcomes including mortality, hospital length of stay (LOS), readmission rate, and patient transfers to higher levels of care. Patients who are not admitted to the unit of choice are typically rerouted to alternative units or even different levels of care. Song et al. (2019) and Dong et al. (2018) study off-placement of patients when bed availability in the primary unit is limited. Such off-placement has important clinical and operational implications as it can result in longer LOS. Dong et al. (2018) find that by carefully coordinating admissions within the internal hospital network, ED boarding can be reduced. Indeed, patients waiting for access to care is highly undesirable. Chan et al. (2016) finds that delays in ICU admission can increase ICU LOS, which, in turn, can create more congestion in an already busy unit. In the ED, another consequence of long waits is an increase in the likelihood of patients leaving without being seen (Batt et al. 2019). In contrast to this body of work which primarily focuses on the impact of admission decisions and waiting on outcomes, we focus on measuring the degree of forward-looking behavior in admission decisions and quantify its impact on hospital and patients.

Clearly, bed availability, or conversely congestion, has substantial impacts on access to care and whether a patient is admitted, is rerouted, or waits. Indeed, there is substantial evidence the congestion influences patient flow. For instance, it can impact who is admitted to the ICU (Kim et al. 2015) and when patients are discharged (Kc and Terwiesch 2012). These works demonstrate that clinicians may alter their actions based on congestion, potentially to the detriment of patient outcomes. There is evidence this impact of congestion on clinical behaviors also arises in the ED. Batt and Terwiesch (2016) find that when the ED becomes congested, certain test are initiated at triage in hopes of improving flow in the congested department.

Our work is also closely related to the literature on structural estimation in operations management. Structural models have been widely used in different fields of operations management including supply chains (Bray et al. 2019) and, more closely related to our work, service operations (Li et al. 2014). In

particular, Akşin et al. (2013) take the structural estimation approach to study caller abandonment behaviors in a call center. Subsequent work including Akşin et al. (2016) and Yu et al. (2016) study the impact of delay announcement in call centers. The element of human customers and human servers often introduces interesting dynamics. For instance, Lu et al. (2013) find that observed queue lengths impact the purchasing behavior of customers in a super market setting. Emadi and Staats (2019) find that the attrition of agents at a management firm appears to be insensitive to salary.

Structural models have also been used specifically in the healthcare operations management literature. Olivares et al. (2008) use a newsvendor model to study how a hospital balances the costs of reserving too much versus too little operating room capacity for cardiac surgery cases. In a different operating room setting, Rath and Rajaram (2018) use a choice model to estimate costs associated with operating room scheduling of anesthesiologists. Our work contributes to the application of structural estimation in healthcare operations management, but in the ICU setting. To the best of our knowledge, we are the first to estimate a dynamic structural model in healthcare operations. In particular, we develop a structural model to estimate the ICU admission decision and use the estimated discount factor as a measure of the hospital's discounting behavior.

The third stream of relevant literature is behavioral operations management and, particularly, its application in healthcare. In healthcare settings, there is evidence that the behaviors of physicians, staffs, or patients can have substantial impact. Green et al. (2013) studies nurse absenteeism and finds nurses exhibit aversion to higher levels of anticipated workload, leading to endogenous absenteeism rates that must be considered in nurse staffing. Song et al. (2015) compare the ED wait time between a system with dedicated queues versus a pooled queue. They find the wait time decreases when de-pooling and suggest that such a phenomenon has a behavioral explanation where physicians feel an increased ownership of patient wait time when faced with a dedicated queue. Ibanez et al. (2017) examines how radiologists view scans given their complete discretion to determine the order to complete tasks. This discretion can lead to inefficiencies in completing tasks. In a setting very similar to ours, Kim et al. (2019) studies the ICU admission decision from a behavioral perspective. They propose a behavioral model and use controlled experiments to understand whether and how physicians are impacted by occupancy when making admission decisions. They identify a number of factors, such as the availability of information, which can bias physician decisions. While we also look at the ICU admission decision, we focus on the behavior of the hospital and take a structural estimation approach to estimate the discount factor from data.

Finally, from the methodological aspect, our work is related to the literature on identification and estimation of dynamic discrete choice models. The dynamic discrete choice model we use resembles the work in the econometrics community pioneered by Rust (1987). Our study extends this line of literature by applying the dynamic discrete choice model in the ICU context to study the hospital's admission decisions. Furthermore, we estimate the discount factor and cost parameters jointly from empirical data. The identification of

the discount factor is generally a very hard problem for dynamic models (Magnac and Thesmar 2002). We rely on recent developments in Komarova et al. (2018) that establishes joint identification of discount factor and payoff parameters for dynamic choice models with linear structure.

2. Setting and Data

We utilize a large data set collected from 21 Kaiser Permanente Northern California (KPNC) hospitals. The data contains over 300,000 hospitalizations over the period of two years. All patients are covered by KPNC insurance and received care at one of the KPNC hospitals.

Each observation in our data corresponds to a single hospitalization. For each hospitalization, we have patient level information such as age, gender, admitting hospital, admitting diagnosis, and three severity scores. The severity scores include a measure of the patient's chronic disease burden (COPS2), an acuity score (LAPS2), and a predicted in-hospital mortality risk score (CHMR). The LAPS2 score is the main severity measure we use in the analysis. It is assigned at hospital admission and measures the clinical severity of a patient based on labs and vital signs taken in the last 72 hours prior to admission – including any that may have been taken in the ED; a score of 110 is generally considered to capture a critically ill patient. More details about these scores can be found in Escobar et al. (2012) and Escobar et al. (2013). In addition to the patient level information, we also observe the admission and discharge time for each unit each patient stayed in during the hospitalization, as well as the type of care the unit provides – i.e., ICU, transitional care unit (TCU), general medical-surgical ward, operating room (OR), or the postanesthesia care unit (PAR). It is important to note that while we are able to see the full trajectory of each patient, including the ultimate decision outcome, the data does not include direct information on the decision making process. Additionally, while we have a rich dataset that includes detailed patient data, we do not have data on other factors that may influence the decision such as (i) nurse staffing availability, (ii) ED patient census, (iii) diversion policies, (iv) specific physician coverage policies, and (v) the possibility to flex capacity.

We utilize the data from all 312,306 hospitalizations to compute the maximum capacity and real-time occupancy level of the ICU in each hospital. The maximum ICU capacity varies from 7 to 36 beds across hospitals, and the average occupancy level varies from 34% to 76%. Among all ICU admissions, 63% are admitted via the ED to a medical service; 10% are admitted through a non-ED unit to a medical service;10% are admitted via the ED to a surgical service, and 17% are admitted through a non-ED unit to a surgical service (i.e., these patients are scheduled surgeries).

Our study focuses on how the hospital internalizes the intertemporal externalities in ICU admissions. Since non-ED patients are more likely to be scheduled arrivals to the hospital and because there are often fixed care protocols for surgical patients, our study is most relevant for patients admitted to a medical service via the ED. The ICU admission decision for these patients is made as follows. After a patient is stabilized in the ED, the ED physician provides an initial assessment about whether the patient needs to be admitted

to the hospital. If the ED physician believes the patient needs to go to the ICU, an intensivist will be called to the ED for a consultation. While the intensivist makes the ultimate decision about whether and when the patient is admitted to the ICU, it is important to emphasize that the decision is determined by a *system* including various physicians, administrators, and possibly, patient family members, as well as the types of alternative interventions available (e.g. flexing capacity).

Next, we describe the data selection process for our study cohort. We start from a total of 312,306 hospitalizations. We restrict our study to the hospitalizations admitted to a medical service via the ED, which comprises the largest proportion of admitted patients ($> 60\%$). Note that for patients who are admitted via the ED, they appear in our data set as soon as the admission decision has been made; as such, we do not have information about patients discharged home from the ED nor patients for whom a disposition decision has not yet been made. We then drop 12 hospitalizations with unknown gender. We also eliminate 9,128 (4.8%) hospitalizations for patients who experience hospital transfers or transports outside of KPNC. As we explain in more detail in Section 3, our study focuses on three possible decisions for each patient in each decision epoch: keep the patient waiting in the ED, admit the patient to the ICU, or reroute the patient to the ward (or the TCU if the hospital has one). We drop 3,066 (1.7%) hospitalizations where the patient was admitted to other units – e.g., OR or PAR, from the ED. Finally, we drop 1,675 (1%) hospitalizations with ED waiting time longer than 12 hours as these episodes can be considered outliers (the average waiting time is shorter than two hours).

Because our data spans over two years, some hospitals might adjusted their ICU capacities during the sample period. As a result, we restrict our study cohort to the periods of each hospital with stable ICU capacity and occupancy. We follow three steps to select the sample. First, we discard the first and last month of data for all hospitals. Second, for several hospitals, we drop the period at either end of the sample where the ICU occupancy dramatically fluctuates or significantly differs from the more stable period in the middle. Finally, for hospital 21, we find that its ICU capacity experienced a substantial increase during the sample period (from 13 to 16). As a result, we split it into two parts, i.e., before and after the capacity change, and treat them as two hospitals in the estimation. We refer to 22 hospitals in our study cohort from here on. The number of days and hospitalizations for each hospital in the final study cohort are summarized in Table 14 in Appendix A.4. In total, we drop 11,268 (6.4%) hospitalizations that are outside the stable periods.

The final study cohort consists of 164,167 hospitalizations. Out of them, 19,683 (12.0%) are admitted to the ICU, and the rest are rerouted to the ward. In Table 1 below, we summarize the patient characteristics of the complete final study cohort and the subset which are admitted to the ICU in the left and right panel, respectively. As expected, the admitted cohort has higher average severity scores than the complete cohort.

3. Descriptive Evidence of Discounting Behavior

In this section, we conduct reduced form regressions to analyze the main determinants of the system's ICU admission decisions. The goal of these analyses is two fold. First, they provide descriptive evidence directly

Table 1 Summary Statistics of Patient characteristics of final study cohort and the subset of patients who are admitted to the ICU.

Final study cohort: N=164,167

	Min	Max	Mean	Median	SD
LAPS2	0.00	294.00	74.11	70.00	37.47
COPS2	0.00	306.00	44.97	28.00	43.09
CHMR	0.00	0.99	0.04	0.02	0.08
Male			0.53		
Age (years)	18.00	113.00	67.27	70.00	17.59
EDWait (hours)	0.02	12.00	1.30	0.88	1.41

Admitted cohort: N=19,683

	Min	Max	Mean	Median	SD
LAPS2	0.00	294.00	105.03	102.00	45.98
COPS2	0.00	276.00	48.67	30.00	47.27
CHMR	0.00	0.99	0.10	0.04	0.15
Male			0.48		
Age	18.00	111.00	64.52	67.00	17.48
EDWait	0.02	11.98	1.36	0.90	1.45

Note. LAPS2, COPS2, and CHMR are severity of illness scores. EDWait corresponds to the ED boarding time.

from the data on the behavior of the hospitals with respect to the short versus longer term considerations. Second, they provide motivations for the decision model which describes this behavior in the next section.

At a high level, one can think of the discounting behavior as how far into the future the hospital considers when making decisions. A hospital with a low discount factor likely focuses primarily on the implications of the admission decision on the current patient and may consider system dynamics over a shorter horizon of two to six hours. On the other hand, a hospital with more a longer-term focus is concerned with the current patient as well as the impact any decision will have on the ability to treat patients who may arrive later, for example, within the next 24 hours. This discounting behavior can manifest itself by how much the hospital alters its admission decisions based on congestion. While the hospital may not be fully aware of its potential change in behavior due to congestion, we aim to understand the *perceived discount factor* that can be elicited from the observed flow of patients.

3.1. Model

We apply a multinomial logit model to estimate the ICU admission decisions. In each period, the hospital chooses one of the three options for each patient: admit the patient into the ICU; reroute the patient the medical/surgical ward; or make the patient wait in the ED. Note that the decision to keep the patient waiting in the ED is often necessitated by system-level considerations; in the absence of capacity constraints, a patient would not be kept waiting for admission to the hospital. We include patient characteristics, system state variables, and seasonality effects as the potential determinants of these decisions.

A period is two hours. This granularity provides a reasonable amount of time for transferring the patient from one unit to the next after the admission request is issued. In addition, the time stamp data exhibits evidence of 'rounding' to the nearest hour. As a result, using two-hour periods avoids potential measurement errors due to the rounding behavior (see Section 4.3 of Hu (2018)).

At the start of each period, we construct system "snapshots" which includes detailed information (e.g., sex, age, and severity scores) for each patient in the ED, as well as the total number of patients boarding in the ED and the number of patients in the ICU.

The system's decision on patient i in period t is determined by two types of variables in the model: patient i's characteristics $\mathbf{X}_i$ and system state variables in period t, $\mathbf{S}_t$. $\mathbf{X}_i$ includes patient i's sex, age, as well as three severity scores–i.e., LAPS2, COPS2, and CHMR. To account for potential differences between hospitals, we also include a categorical variable to represent the hospital in which patient i is treated. In summary, we have

$$\mathbf{X}_i = \{\text{Gender}_i, \text{Age}_i, \text{LAPS2}_i, \text{COPS2}_i, \text{CHMR}_i, \text{Hosp}_i\}.$$

The system state vector $\mathbf{S}_t$ includes the following variables

$$\mathbf{S}_t = \{\text{ICUOccu}_t, \text{EDNum}_t, \text{DepPre}_t, \text{AvgLAPS2}_t, \text{WeekDay}_t, \text{DayTime}_t\},$$

where ICUOccu_t denotes the current ICU occupancy level. As the ICU sizes vary dramatically across the hospitals, we use the ICU percentile rank to measure occupancy. EDNum_t denotes the number of current ED patients for whom a decision to admit them to the hospital has been made but for which a decision about when and where to admit them needs to be made, and DepPre_t denotes the number of patients who left the ICU (i.e. discharged from the ICU or died) in the previous period. AvgLAPS2_t denotes the average severity level measured by the LAPS2 score of the current ICU patients in period t. Finally, the indicator variables WeekDay_t and DayTime_t capture potential seasonality in the admission decisions: WeekDay_t equals to one if the current period is a weekday, and zero otherwise; DayTime_t equals to one if the current period is between 7AM to 7PM, and zero otherwise.

For patient i who is in the ED at the start of period t, we estimate the system's decision d_{it} using a multinomial logit model:

$$\begin{aligned}\ln\left[\frac{\Pr(d_{it}|\mathbf{X}_i,\mathbf{S}_t)}{\Pr(\text{Rer}_{it}|\mathbf{X}_i,\mathbf{S}_t)}\right] &= \gamma_{0,d} + \gamma_{G,d}\text{Gender}_i + \gamma_{A,d}\text{Age}_i + \gamma_{L,d}\text{LAPS2}_i + \gamma_{CP,d}\text{COPS2}_i + \gamma_{CH,d}\text{CHMR}_i \\ &\quad + \gamma_{H,d}\text{Hosp}_i + \gamma_{ICU,d}\text{ICUOccu}_t + \gamma_{ED,d}\text{EDNum}_t + \gamma_{Dep,d}\text{DepPre}_t \\ &\quad + \gamma_{AvgL,d}\text{AvgLAPS2}_t + \gamma_{WD,d}\text{WeekDay}_t + \gamma_{DT,d}\text{DayTime}_t + \epsilon_{it}, \end{aligned} \tag{1}$$

where $d_{it} \in \{Wait_{it}, Adm_{it}\}$. $\Pr(d_{it}|\mathbf{X}_i\mathbf{S}_t)$ is the probability of d_{it} conditional on $(\mathbf{X}_i, \mathbf{S}_t)$. $\Pr(\text{Rer}_{it}|\mathbf{X}_i\mathbf{S}_t)$ is the probability of rerouting patient i in period t conditional on $(\mathbf{X}_i, \mathbf{S}_t)$. We use the rerouting decision as the base case, and estimate the probabilities of the ICU admission ($d = Adm$) and waiting ($d = Wait$) decisions relative to the rerouting decision respectively.

We first estimate the model by combining the patient data from all hospitals. Then, considering the heterogeneity across hospitals, we also estimate the model for individual hospitals separately after dropping the categorical variable term $\gamma_{H,d}\text{Hosp}_i$ in (1). For robustness, we also consider the binary version of model (1) with only admission and rerouting decisions (see Appendix A.3). That is, we estimate the patients' destination unit decisions (admission relative to rerouting) regardless of their waiting time. Only select coefficients from the multinomial logistic regression will be included in the main body; the full regression results and the results from the logistic regression can be found in the Appendix A.4.

3.2. Results

In this section, we report the estimation results for the multinomial logit model (1). We report the estimated coefficients for three main variables: LAPS2_i, ICUOccu_t, and EDNum_t. Table 2 below shows the estimation results for model (1) with all hospitals combined. Note that in the estimation, each hospitalization may be counted multiple times if the patient waits in the ED for more than one period. Thus, the sample size (183,691) is larger than the number of total hospitalizations (164,167).

Table 2 Estimation results for Multinomial-Logistic Regression (1), N = 183,691, R-squared = 0.11

	LAPS2_i	ICUOccu_t	EDNum_t
	γ_L	γ_{ICU}	γ_{ED}
Waiting	0.009^{***}	1.590^{***}	0.129^{***}
	(0.000)	(0.028)	(0.005)
Admission	0.027^{***}	-0.292^{***}	-0.051^{***}
	(0.000)	(0.028)	(0.006)

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Select coefficients (LAPS2 score, ICU occupancy level, and number of ED boarding patients) from (1) for waiting and admission decisions respectively.

The results in Table 2 show that all the coefficients are statistically significant at the 0.01 level and have the expected signs. In particular, higher LAPS2 score increases the probability of admission relative to rerouting, as these patients are more critically ill. For these severe patients, the system may also need to keep them waiting in the ED when the ICU is congested. Thus, higher LAPS2 score also increases the probability of waiting, although the magnitude of the impact is smaller. More importantly, the estimates of γ_{ICU} and γ_{ED} suggest that, even after controlling for patient characteristics and fixed effects, a busier system state (more congested ICU or more congested ED) decreases the probability of admission and increases the probability of waiting, respectively. Such evidence suggests that the system indeed internalizes the intertemporal externalities on the ICU admission decisions by adjusting their behaviors according to the current system state: When ICU or ED is more congested, they are more likely to delay the admission of current patients to save ICU beds for future patients.

We also report the McFadden's pseudo R^2 (see Appendix A.3 for details of this measure) in Table 2. While 0.11 is relatively low, we emphasize that the pseudo R^2 is computed using the null model where hospital fixed effects are included. Moreover, it is consistent with the magnitude seen for models of operational decisions in healthcare systems (see, e.g., Kim et al. (2015), Chan et al. (2016), and Song et al. (2019) among others).

In Table 3, we report the coefficients for the three main covariates when we estimate model (1) for each hospital separately. In the last column, we also provide the McFadden's R^2 of the model for each hospital. The results are qualitatively similar to that for all hospitals combined in Table 2. All coefficients

Table 3 **Estimation results for logistic regression (1): Individual hospitals**

		for *Waiting* decision			for *Admission* decision			
Hosp	Size	LAPS2	ICUOccu	EDNum	LAPS2	ICUOccu	EDNum	R^2
		γ_L	γ_{ICU}	γ_{ED}	γ_L	γ_{ICU}	γ_{ED}	
1	13,964	0.006^{***}	1.232^{***}	0.171^{***}	0.022^{***}	-0.178	-0.033	0.11
		(0.001)	(0.086)	(0.015)	(0.001)	(0.093)	(0.019)	
2	12,871	0.008^{***}	1.191^{***}	0.053^{***}	0.026^{***}	-0.485^{***}	-0.082^{***}	0.09
		(0.001)	(0.079)	(0.013)	(0.001)	(0.103)	(0.018)	
3	8,391	0.007^{**}	0.834^{***}	0.256^{***}	0.031^{***}	-0.245	-0.017	0.11
		(0.002)	(0.195)	(0.045)	(0.002)	(0.161)	(0.044)	
4	16,162	0.013^{***}	2.852^{***}	0.124^{***}	0.031^{***}	-0.104	-0.022	0.18
		(0.001)	(0.121)	(0.016)	(0.001)	(0.115)	(0.021)	
5	5,499	0.012^{***}	1.279^{***}	0.312^{***}	0.031^{***}	-0.527^{***}	-0.083	0.11
		(0.002)	(0.179)	(0.046)	(0.002)	(0.149)	(0.048)	
6	11,698	0.012^{***}	2.426^{***}	0.168^{***}	0.029^{***}	0.151	-0.082^{**}	0.13
		(0.001)	(0.127)	(0.024)	(0.001)	(0.111)	(0.026)	
7	5,200	0.010^{***}	0.978^{***}	0.130^{*}	0.019^{***}	-1.108^{***}	0.076	0.08
		(0.002)	(0.182)	(0.051)	(0.002)	(0.204)	(0.064)	
8	9,382	0.011^{***}	2.623^{***}	0.161^{***}	0.024^{***}	-0.432^{***}	-0.053^{*}	0.11
		(0.001)	(0.137)	(0.026)	(0.001)	(0.104)	(0.026)	
9	14,774	0.007^{***}	2.045^{***}	0.151^{***}	0.029^{***}	-0.250^{*}	-0.041^{*}	0.13
		(0.001)	(0.090)	(0.012)	(0.001)	(0.111)	(0.018)	
10	6,032	0.008^{**}	1.728^{***}	0.103	0.031^{***}	-0.277	-0.119^{*}	0.12
		(0.003)	(0.290)	(0.067)	(0.002)	(0.173)	(0.051)	
11	3,334	0.008^{***}	0.924^{***}	0.150^{*}	0.028^{***}	-0.926^{***}	-0.088	0.09
		(0.002)	(0.162)	(0.065)	(0.002)	(0.198)	(0.089)	
12	8,413	0.013^{***}	1.232^{***}	0.078^{***}	0.034^{***}	-0.696^{***}	-0.076^{*}	0.10
		(0.001)	(0.106)	(0.017)	(0.002)	(0.164)	(0.030)	
13	8,620	0.008^{***}	1.777^{***}	0.135^{***}	0.032^{***}	-0.154	-0.061	0.13
		(0.002)	(0.165)	(0.037)	(0.002)	(0.139)	(0.038)	
14	14,209	0.011^{***}	1.607^{***}	0.056^{**}	0.022^{***}	-0.135	-0.017	0.10
		(0.001)	(0.104)	(0.018)	(0.001)	(0.105)	(0.021)	
15	7,073	0.006	0.634^{*}	0.059	0.030^{***}	-0.046	-0.015	0.12
		(0.003)	(0.302)	(0.076)	(0.002)	(0.152)	(0.042)	
16	7,517	0.009^{***}	1.918^{***}	-0.011	0.031^{***}	-0.289	-0.089^{*}	0.13
		(0.002)	(0.203)	(0.049)	(0.002)	(0.153)	(0.045)	
17	3,896	0.014^{***}	1.593^{***}	0.281^{***}	0.027^{***}	-0.974^{***}	-0.060	0.13
		(0.002)	(0.191)	(0.063)	(0.002)	(0.187)	(0.078)	
18	7,550	0.005^{**}	0.357^{*}	0.191^{***}	0.029^{***}	0.114	-0.031	0.10
		(0.002)	(0.162)	(0.038)	(0.002)	(0.163)	(0.042)	
19	7,732	0.008^{***}	1.594^{***}	0.268^{***}	0.028^{***}	-0.022	-0.081^{*}	0.13
		(0.002)	(0.231)	(0.050)	(0.001)	(0.130)	(0.034)	
20	5,678	0.002	1.219^{***}	0.334^{***}	0.029^{***}	-0.351^{*}	-0.046	0.12
		(0.002)	(0.168)	(0.043)	(0.002)	(0.177)	(0.058)	
21	3,154	0.007^{***}	0.635^{***}	0.124^{**}	0.032^{***}	-0.943^{***}	0.041	0.10
		(0.002)	(0.145)	(0.040)	(0.003)	(0.228)	(0.071)	
22	2,542	0.012^{***}	1.397^{***}	0.178^{**}	0.027^{***}	-0.314	-0.142	0.10
		(0.002)	(0.190)	(0.056)	(0.003)	(0.205)	(0.075)	

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Select coefficients (LAPS2 score, ICU occupancy level, and number of ED boarding patients) from (1) for waiting and admission decisions respectively.

have the expected signs for most hospitals, although some are not statistically significant as the sample size of each individual hospital is much smaller than all hospitals combined. In particular, γ_L is positive for both admission and waiting decisions in all 22 hospitals; γ_{ICU} is positive for the waiting decision in all 22 hospitals, and negative for admission decision in 20 hospitals; γ_{ED} is positive for the waiting decision in 21 hospitals, and negative for admission decision in 20 hospitals . As shown in Table 18 in Appendix A.4, the results from the binary logistic regression with only admission and rerouting decisions are also similar.

The consistent pattern of the coefficients in Table 3 suggests that the hospitals do indeed internalize the intertemporal externalities on admission decisions. However, the large variation in the coefficients suggests that they may have very different *degrees* of how much they internalize such intertemporal externalities. If the hospital truly only considered the short-term implications of their decisions, the admission probabilities would be insensitive to the system state as they mostly care about the current patients only. On the other hand, the hospital may alter their admission probabilities when the system is congested in order to save ICU beds for future patients. The heterogeneity observed in the reduced-form regression highlights the importance to measure and understand how far into the future the hospital accounts for when making admission decisions. More importantly, the reduced form regression results do not allow us to conduct counterfactual analyses to quantify the impact of the forward looking behavior on key system performance metrics. We next propose a structural model to address these questions.

4. Structural Estimation

In this section, we first introduce the structural model which describes the ICU admission process within a hospital. After that, we explain the identification of the discount factor and describe the estimation procedure.

4.1. Dynamic discrete choice model

In the development of our structural model, we consider the admission decision at the level of the hospital. We model the ICU admission decision using a dynamic discrete choice model. There are three key features of the model: 1) in each period, the hospital considers three options for each patient: admitting the patient, rerouting the patient, or keeping the patient waiting in the ED; 2) the decision depends on both patient conditions and the current system status; 3) finally, the model allows for the hospital to consider the future and to make dynamic decisions. We provide detailed descriptions of the model below.

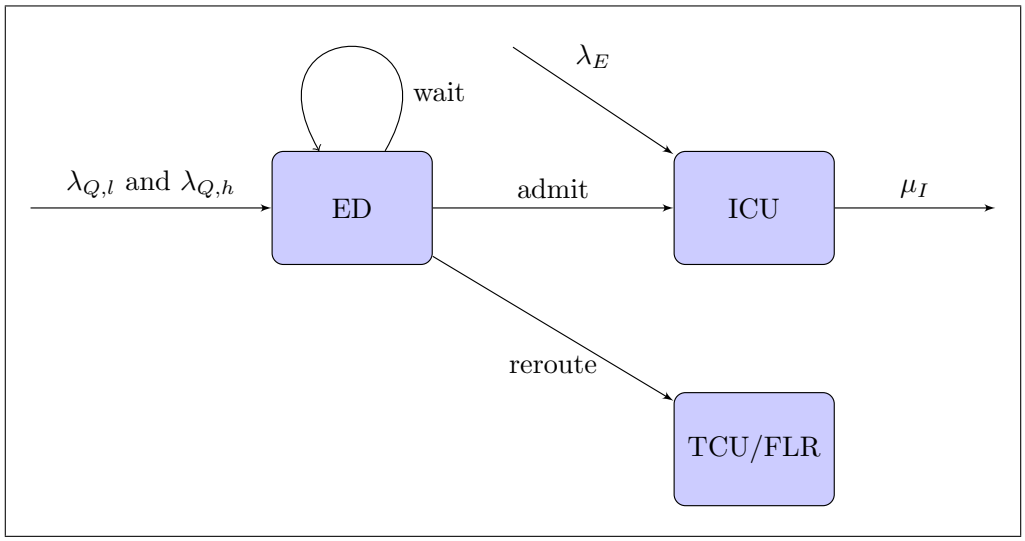
We divide the focal patient cohort into two classes: the low severity class and the high severity class, represented by subscripts l and h, respectively. We define the two classes based on the patient's LAPS2 score, which has been shown to be an effective severity measure of illness (Escobar et al. 2013). We assume that the ED has capacities Q_l and Q_h for the low and high classes, respectively. Note that in our model a patient arrival in the ED corresponds to when the decision to admit the patient to the hospital has been made (i.e. we only capture ED boarding patients), so the ED capacities can be regarded as capturing the limited

boarding space in the ED. We defer the discussion about how various parameters in the model, including Q_l and Q_h, are estimated from the data to Section 4.3. The two classes of patients arrive in the ED every period according to the following distribution. For $i \in \{l, h\}$, let $A_{i,t}$ be the number of class i patients arriving to the ED in time period t. $A_{i,t}$ follows a truncated Poisson distribution with rate $\lambda_{Q,i}$. While in theory $A_{i,t}$ can be unbounded, we truncate it at the maximum number of patients arriving at ED in each period (denoted by M_{A_i}) observed in our data. This limits the state space of the model and helps to keep the estimation computationally feasible.

In addition to our focal patient cohort, there are other patients who can occupy ICU beds. E_t denotes the number of surgical and non-ED medical patients arriving in the ICU in period t, referred to as the external arrivals. E_t is distributed according to a Poisson distribution with arrival rate λ_E.

For each individual patient in the ICU, including both ED patients and external arrivals, their service completion probability per period is given by μ_I. Thus, D_t, the number of patients departing from ICU in each period, follows a binomial distribution. We denote the ICU capacity by B, which is the total number of beds in the ICU. We assume the ward units in the hospital have ample capacity. An overview of the system flow is summarized in Figure 1 below.

Figure 1 Overview of patient flow and potential paths in the ED-ICU/ward system



At the beginning of period t, the system state is given by a three dimensional vector,

$$s_t = (n^E_{l,t}, n^E_{h,t}, n^I_t),$$

where $n^E_{i,t}$ is the number of class i patients in the ED, $i \in \{l, h\}$, and n^I_t is the number of patients in the ICU in period t. By the capacity constraints, $n^E_{i,t} \leq Q_i$ for $i \in \{l, h\}$ and $n^I_t \leq B$. For each patient in the ED, the hospital determines one of the following three decisions: admit to ICU, reroute to wardir, or keep them

waiting in the ED. Since the patients are treated as identical in terms of their observables within each class, the hospital's action can be described by the following four dimensional vector

$$d_t = (a_{l,t}, r_{l,t}, a_{h,t}, r_{h,t}),$$

where $a_{i,t}$ and $r_{i,t}$ denote the numbers of admitted and rerouted patients of class i, $i \in \{l, h\}$, respectively. Due to the capacity constraint in the ICU, the admissible action set for system state s_t is

$$\Pi(s_t) = \left\{(a_{l,t}, r_{l,t}, a_{h,t}, r_{h,t}) : a_{l,t} + r_{l,t} \leq n^E_{l,t}, a_{h,t} + r_{h,t} \leq n^E_{h,t}, a_{l,t} + a_{h,t} \leq B - n^I_t\right\}. \tag{2}$$

$\Pi(s_t)$ specifies the following set of constraints: The first two constraints state that the sum of admitted and rerouted patients must be smaller than or equal to the total number of patients currently in the ED. The last constraint requires that the total number of admitted patients must be smaller than or equal to the current number of available beds in the ICU.

In each period, a cost $c(s_t, d_t)$ associated with state s_t and action d_t is incurred. We note that these costs represent the hospital's assessment of the 'costs', or disutility, incurred for each state-action pair. They may capture clinical costs, operational costs, financial costs, etc. These costs are not necessarily consistent with the hospital's or patients' assessment of the costs associated with each state-action pair; they can also capture various system-level constraints. We assume admitting a patient to the ICU has *zero* costs for both classes of patients. This assumption is a reasonable approximation for the following reasons. First, ICUs have very high fixed operating costs, therefore, the marginal cost of admitting one patient is very small for either class of patients (Roberts et al. 1999, Kahn et al. 2008). For example, the extremely high set-up cost for specialized equipment such as ventilators and monitors in the ICU can be regarded as sunk costs, which do not affect the hospital's ICU admission decision. Additionally, hospitals tend to staff their ICU beds based on fixed nurse-to-bed ratios and rarely adjust staffing levels based on occupancy and severity of patients. Finally, ICU patients are the most severe type in a hospital. Therefore, it is reasonable to assume that the hospital's primary concern is to prioritize patient outcomes, and the immediate cost of ICU admission is negligible in comparison. In other words, the hospital's primary goal is to minimize undesirable patient outcomes due to lack of ICU care. As such, we assume zero admission costs for both classes of patients. As we discuss in Section 4.2, the assumption of zero admission cost is crucial for the identification of the discount factor in the dynamic discrete choice model.

We assume that rerouting each low (high) severity patient to the ward incurs a rerouting cost $c_{r,l}$ ($c_{r,h}$), while keeping each low (high) severity patient waiting in the ED incurs a waiting cost $c_{w,l}$ ($c_{w,h}$). These cost parameters represent the *average* rerouting or waiting cost across patients within each class. We restrict the waiting costs for both classes to be positive, i.e., $c_{w,l}, c_{w,h} > 0$. This is a reasonable assumption since longer ED boarding has been shown to be associated with increased mortality risk and hospital LOS (Singer

et al. 2011). Thus, on average, keeping patients in the ED is more likely to lead to negative outcomes compared with admitting them to ICU. As we assume admission incurs zero costs for both classes, the average waiting costs $c_{w,l}$ and $c_{w,h}$ should both be positive as longer waiting time is less desirable than immediate ICU admission. On the other hand, we do not restrict the sign of the rerouting costs $c_{r,l}$ and $c_{r,h}$. Patients who are not critically ill can often receive sufficient care in the ward, so rerouting them to the ward is not necessarily worse than admitting them to the ICU. Therefore, the average cost of rerouting, compared with ICU admission, can be positive or negative. Finally, there are no costs associated with external arrival patients in our model. The majority ($> 70\%$) of the external arrivals to the ICU are surgical patients (63% of which are scheduled surgeries) for whom ICU beds are often reserved in advance; thus, our model focuses on the costs associated with the patients for whom there are less clear protocols – those admitted to a medical service via the ED.

Given the system state, s_t, and the action, d_t, the total per period cost is given by,

$$c(s_t, d_t) = c_{r,l} r_{l,t} + c_{w,l} \left(n^E_{l,t} - a_{l,t} - r_{l,t} \right) + c_{r,h} r_{h,t} + c_{w,h} \left(n^E_{h,t} - a_{h,t} - r_{h,t} \right), \tag{3}$$

which is the sum of rerouting and waiting costs for the two classes of patients. Then, the hospital's per period utility can be written as,

$$U(s_t, d_t, \varepsilon_t) = -c(s_t, d_t) + \varepsilon_t(d_t) \tag{4}$$

where $\varepsilon_t(d_t)$ is the idiosyncratic utility component associated with action d_t, which is observed by the hospital when making the decision, but not to the researcher. The additively separable form (4) is similar to the assumption in Rust (1987) and numerous works in the structural estimation of dynamic discrete choice literature.

At the beginning of period t, the hospital observes the system state, s_t and the idiosyncratic utility component, ε_t, then they chooses the optimal action d_t that solves following infinite horizon utility maximization problem:

$$\sup_{d_t \in \Pi(s_t)} \mathsf{E} \left\{ \sum_{j=t}^{\infty} \beta^{j-t} U(s_j, d_j, \varepsilon_j) | s_t, \varepsilon_t \right\}. \tag{5}$$

The discount factor $\beta \in (0, 1)$ captures the trade-off between current and future utility, which is the focus of our study. The expectation is taken over both the random component ε_t and the transitions of the system– i.e., the arrivals and departures of patients in each period. In addition, note that the expectation in (5) is conditional on both s_t and ε_t, as the random component is observable to the hospital before making the decision in period t. As noted earlier, because the costs capture the hospital's assessment, this utility maximization problem is employed to model the hospital's decision making and it's relationship to the observed flows of patients. In other words, as the system of different players and system-level effects which determines the outcomes of the patients, these costs does not necessarily coincide identically with the goals

of the patients or hospital administrators. We define the value function as the objective in (5) given the optimal action sequence, i.e.,

$$V(s_t, \varepsilon_t) = \sup_{d_t \in \Pi(s_t)} \mathsf{E}\left\{ \sum_{j=t}^{\infty} \beta^{j-t} U(s_j, d_j, \varepsilon_j) | s_t, \varepsilon_t \right\}. \tag{6}$$

After the hospital chooses an action in period t, the system state evolves as follows. The number of ED patients of class i becomes $n^E_{i,t} - a_{i,t} - r_{i,t}$, and the number of patients in the ICU is $n^I_t + a_{l,t} + a_{h,t}$. We define an "intermediary" state $\varphi(s_t, d_t)$ after the action d_t is taken, which is given by

$$\varphi(s_t, d_t) = \left(n^E_{l,t} - a_{l,t} - r_{l,t}, n^E_{h,t} - a_{h,t} - r_{h,t}, n^I_t + a_{l,t} + a_{h,t}\right), \tag{7}$$

and describes the impact of action d_t on the system.

The system then evolves according to the following two steps. First, $A_{i,t}$ new patients of class i arrive to the ED, and E_t patients arrive to the ICU through non-ED channels–i.e., the external arrivals. If the ED or ICU is full, new arrivals cannot be accepted. Thus, the total accepted ED and ICU arrivals are given by

$$A^{acc}_{i,t} = \max\left\{A_{i,t}, Q_i - (n^E_{i,t} - a_{i,t} - r_{i,t})\right\}$$

and

$$E^{acc}_t = \max\left\{E_t, B - (n^I_t + a_{l,t} + a_{h,t})\right\},$$

respectively. Second, D_t patients leave the system as they complete their service in the ICU. This completes the system transition for period t.

The system state at the beginning of period $t+1$ is:

$$s_{t+1} = (n^E_{l,t+1}, n^E_{h,t+1}, n^I_{t+1}),$$

with

$$n^E_{i,t+1} = n^E_{i,t} - a_{i,t} - r_{i,t} + A^{acc}_{i,t} \text{ and } n^I_{t+1} = n^I_t + a_{l,t} + a_{h,t} + E^{acc}_t - D_t. \tag{8}$$

It is clear from the above description that the transition of s_t is Markovian, and its distribution only depends on s_t and d_t, but not ε_t. The timeline of system transition is summarized in Figure 2.

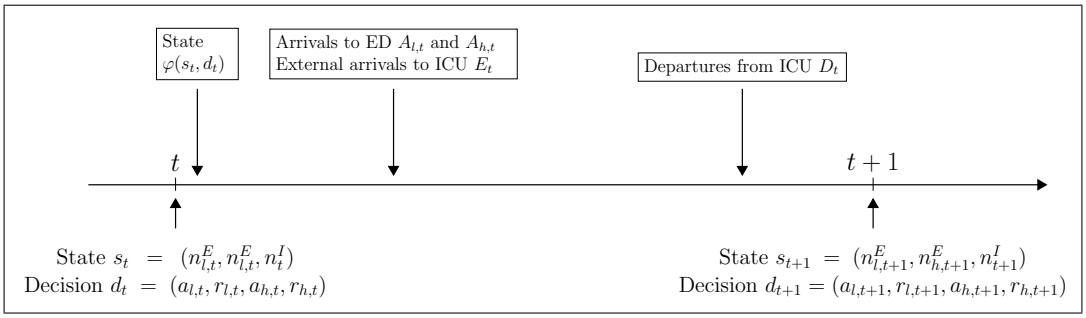
Both the current period's utility and how a specific action changes the system state, which in turn can impact future payoffs, may influence the hospitals's decision. Thus, the hospital chooses the action that maximizes

$$d_t = \arg\max_{d \in \Pi(s_t)} (-c(s_t, d) + \varepsilon_t(d) + \beta \mathsf{E}[V(s_{t+1}, \varepsilon_{t+1})]),$$

where the function $V(s, \varepsilon)$ is defined in (6). The last term in the right-hand side is the expectation of the future value function after the current action is taken. Thus, the optimal value function $V(s_t, \varepsilon_t)$ solves the following Bellman's equation

$$V(s_t, \varepsilon_t) = \max_{d \in \Pi(s_t)} (-c(s_t, d) + \varepsilon_t(d) + \beta \mathsf{E}[V(s_{t+1}, \varepsilon_{t+1})]),$$

Figure 2 **Timeline of system evolution: depiction of how the state evolves within a single time-slot.**



where the expectation is taken over both the system transition to s_{t+1} and the random component ε_{t+1}.

The above Bellman's equation is hard to evaluate due to the infinite state space associated with ε_t. Thus, we simplify the model by making the same conditional independence assumption (CI) as in Rust (1987).

Assumption 1 (CI) *The transition probabilities of the controlled process (s_t, ε_t) can be factored as*

$$\Pr(s_{t+1}, \varepsilon_{t+1} | s_t, \varepsilon_t, d_t) = q(\varepsilon_{t+1} | s_{t+1}) g(s_{t+1} | \varphi(s_t, d_t)), \tag{9}$$

where $\varphi(s_t, d_t)$ denotes the intermediate state (7) after action d_t is taken; the transition probability $g(s_{t+1}|\varphi(s_t, d_t))$ captures the random arrivals and departures shown by (8). Assumption (CI) states that s_{t+1} is sufficient to determine the distribution of ε_{t+1}. In other words, the random component $\{\varepsilon_t\}$ is superimposed on the state process $\{s_t\}$. Finally, we assume that the random component ε_t is independent and identically distributed (i.i.d.) and follows type I extreme value distribution for each action $d \in \Pi(s_t)$. Thus, the state s_t impacts the distribution of ε_t only through the number of admissible actions. As shown in Rust (1987), this assumption leads to a closed-form expression of the conditional choice probability for action d_t given state s_t, as denoted by $f(d_t|s_t)$.

Proposition 1 *With the above set-up, the conditional choice probability for action d_t given state s_t has the following closed-form representation:*

$$f(d_t | s_t) = \frac{\exp\big(-c(s_t, d_t) + \beta \tilde{V}(\varphi(s_t, d_t))\big)}{\sum_{d \in \Pi(s_t)} \exp\big(-c(s_t, d) + \beta \tilde{V}(\varphi(s_t, d))\big)}, \tag{10}$$

where function $\varphi(s_t, d_t)$ is given by (7). The function $\tilde{V}(s)$ is defined as

$$\tilde{V}(s) = \sum_{s'} \int_{\varepsilon'} V(s', \varepsilon')\, g(s'|s) q(\varepsilon'|s') d\varepsilon'. \tag{11}$$

The explicit expression for $g(s'|s)$, i.e. the transition probability to state s' given s (the system state after the action is taken but before the random arrivals and departures take place), is provided in Appendix A.1. The function $\tilde{V}(s)$ is the unique fixed point to the following functional equation

$$\tilde{V}(s) = \sum_{s'} \ln \left\{ \sum_{d' \in \Pi(s')} \exp\big(-c(s', d') + \beta \tilde{V}(\varphi(s', d'))\big) \right\} g(s'|s). \tag{12}$$

In the above proposition, we drop the dependence on model parameters to simplify the notation. As in most dynamic discrete choice models, the choice probability (10) has a closed-form logit representation and the value function $\tilde{V}(s)$ solves the functional equation (12). By (11), the new value function $\tilde{V}(s)$ represents the expected future utility given s, which represents the current state after the action has been taken. In the proposition below, we show it is monotonically non-increasing in the number of ICU patients.

Proposition 2 *For two intermediary states (after actions are taken) s and s' with $(n_i^E) = (n_i^E)'$ for $i \in \{l, h\}$ and $(n^I) \leq (n^I)'$, we have*

$$\tilde{V}(s) \geq \tilde{V}(s').$$

That is, for any given number of ED patients, the function $\tilde{V}(s)$ is monotonically non-increasing in the number of ICU patients.

PROOF: See Appendix B. □

This proposition shows that the intertemporal externalities indeed exist for the ICU admission decisions: As the hospital admits more patients, thereby increasing the ICU occupancy, the future expected utility decreases. In turn, this can result in a decrease in the likelihood of ICU admission as the ICU occupancy increases. Thus, the hospital must consider both current and future utilities when making ICU admission decisions.

4.2. Identification of discount factor

In this section, we discuss how our model is empirically identified from observed data. We first develop the general identification result for our model using recent developments in the econometric literature. Then, we use two simple examples to illustrate how some aspects of our model can be directly identified from observed choice probabilities.

In this study, we identify the discount factor and cost parameters *jointly* from observed data. Although the choice probability has a closed-form expression (10) in the dynamic choice model, it is not possible to identify the discount factor and costs parameters jointly without further restriction on the dynamic model (see, e.g. Lemma 3.3 in Rust (1994) or Proposition 2 in Magnac and Thesmar (2002)). Thus, most empirical studies assume the discount factor is known, and then estimate the cost parameters. However, the prespecified discount factors usually lack empirical support and economic justifications. Indeed, the implied discount rate can vary substantially across different settings (Frederick et al. 2002). In our study, identifying the discount factor from observed data is crucial to understanding how hospitals internalize the intertemporal externalities of the ICU admission decisions.

The reason that the discount factor and cost parameters can not be jointly identified is because there exists observationally equivalent structures – i.e., different combinations of discount factor and cost parameters –

that lead to the same choice probabilities for all states and actions. Thus, an agent's actions can be rationalized for different choices of discount factor. For example, in the ICU admission context, if we observe the hospital admits patients "aggressively"(e.g. with high probability), this may be because either the hospital is focused on the near-term – i.e., a small β – or the waiting and rerouting costs are large – i.e., high c_w and c_r. Without further restriction on model primitives (e.g., cost and utility), we cannot differentiate between such cases to identify the true discount factor.

There have been positive identification results for the discount factor in dynamic models. Magnac and Thesmar (2002) suggests that exclusion restriction can be used to identify the discount factor. The exclusion restriction they use is that there exists some state and action pairs for which the single period utilities are identical, but the future period utilities differ. This idea is further elaborated and applied in some empirical contexts (Dubé et al. (2014), Wang (2014), and Ching and Osborne (2017)). However, the exclusion restriction is abstract and hard to verify in applications (Abbring and Daljord (2019)), and all the subsequent literature focus on special cases.

To identify the discount factor in our parametric model, we leverage the recent identification results developed in Komarova et al. (2018). They prove the identification for the discount factor using an empirical model that is linear in the cost parameters conditional on the discount factor. In particular, they construct a one-dimensional criterion function that can be used for identification as well as estimation, which exploits the conditional linear structure and reduces the nonlinear problem to a one-dimensional grid search for the discount factor.

The main identification results of Komarova et al. (2018) proceed as follows. They consider an empirical model with linear structure. The choice and transition probabilities are nonparametrically identified. For a given value of the discount factor, they first construct estimates for the cost parameters following the standard two-step estimation procedure pioneered by Hotz and Miller (1993). The estimator minimizes the distance between the value functions observed from data and those directly implied by the empirical model. Then, they reduce the identification problem to a one-dimensional search for $\beta \in (0, 1)$: If there is a unique value of β, together with the corresponding cost estimates, that minimizes the distance objective function, then the model can be identified under some rank condition. This criterion also provides a natural way to estimate the model with observed data.

We now show how the identification results in Komarova et al. (2018) are applied in our setting. First, it is straightforward that our model satisfies the basic assumptions in Komarova et al. (2018), i.e., additive separability of utility, conditional independence of transition, and finite state space. Additionally, by (3) and (4), it is also clear that the deterministic part of the per-period utility (i.e., $-c(s_t, d_t)$) is linear in the cost parameters $c_{w,l}$, $c_{r,l}$, $c_{w,h}$, and $c_{r,h}$. Thus, the linear-in-parameter assumption is also satisfied in our setting. More details of the assumptions in Komarova et al. (2018) and how they apply to our setting are included in Appendix A.2. We define the base action as admitting every ED patient whenever there are ICU beds

available. As we assume admission costs are zero for both classes of patients, the base action brings zero cost for all states as long as it is admissible. Thus, we can apply Theorem 1 of Komarova et al. (2018) in our setting to identify the discount factor and cost parameters jointly in our dynamic model.

While Komarova et al. (2018) assumes the same admissible action set for all states, our setting has state dependent action sets. This creates additional challenges in the identification of the discount factor as the denominator for the choice probability in (10) is also state-dependent. However, we note that given each state, the admissible action set can be fully determined by (2). Thus, it is easy to verify the derivation leading to Theorem 1 in Komarova et al. (2018) still applies to our setting, where we plug in the admissible action set for each state according to (2). Additionally, in next section, we use two simple examples to show how the state-dependence property of the admissible action set can be handled by constructing proper state-action pairs that identify certain aspects of our model.

We can construct a one-dimensional criterion similar to the one in Komarova et al. (2018) based on the maximum likelihood estimator: For each candidate β, we estimate the cost parameters that maximize the choice likelihood. Then we conduct a one-dimensional search over $\beta \in (0, 1)$. The model can be identified if there is a unique β (together with the cost estimates) that maximizes the likelihood, and that the rank condition in Theorem 1 of Komarova et al. (2018) is satisfied. Using data from 22 hospitals, we find the optimal discount factor that maximizes the likelihood function is unique in all circumstances. Moreover, in most cases, the likelihood function monotonically decreases as the discount factor moves away from the optimal level, which further strengthens the identification of the discount factor. More details of the algorithmic approach can be found in the next section. We provide some illustrative examples of the likelihood versus discount factor in Figure 7 of Appendix A.4. We also show by simulation that the rank condition is satisfied. The results suggest that the discount factor and cost parameters can be jointly identified in our empirical model.

4.2.1. Two illustrative examples for identification In this section, we use two simple examples to illustrate how some aspects of our dynamic model can be directly identified from observed choice probabilities. While the formal identification results rely on the one-dimensional criterion established above, the two examples below provide some insights about how the observed choice probabilities can be used for identification. Taking log on both sides of the choice probability (10), we have

$$\ln f(d_t|s_t) = -c(s_t, d_t) + \beta \tilde{V}(\varphi(s_t, d_t)) - C(s_t), \tag{13}$$

where function $C(s)$ is given by

$$C(s) = \ln \left\{ \sum_{d \in \Pi(s)} \exp\left(-c(s, d) + \beta \tilde{V}(\varphi(s, d)) \right) \right\}.$$

We see the log choice probability can be decomposed to three terms in (13): the first term $-c(s_t, d_t)$ is the negative of the per-period cost, which directly depends on the cost parameters; the second term $\beta\tilde{V}(\varphi(s_t, d_t))$ is related to the discount factor and value function, which captures the impact of the decision on future utility via the change of system state; the last term $C(s_t)$ is a function of system state s_t, but it does not depend on the decision d_t. Note the discount factor and cost parameters are implicitly captured in $\tilde{V}(\cdot)$ and $C(\cdot)$.

As shown by (13), the discount factor and cost parameters are captured in the choice probabilities in a complicated way. However, we will illustrate that certain aspects of the model can be directly identified by constructing appropriate *state and action pairs*. In the first example, we show the rerouting cost difference, i.e., $c_{r,h} - c_{r,l}$, can be identified by cancelling out the second and third terms related to $\beta\tilde{V}(\cdot)$ and $C(\cdot)$ in (13). In the second example, we show certain linear combinations of $\beta\tilde{V}(\cdot)$ can be identified by removing the first and third terms related to $c(s_t, d_t)$ and $C(\cdot)$ in (13),

In the first example, we consider two states $s_1 = (1, 0, 0)$ and $s_2 = (0, 1, 0)$. That is, there is one patient in the ED from the low and high severity class in s_1 and s_2 respectively, and the ICU in both states is assumed to be empty. We denote the probabilities of admitting and rerouting the patient under the two states by $\Pr(a|s_1)$, $\Pr(r|s_2)$, $\Pr(a|s_1)$, and $\Pr(r|s_2)$. Following Proposition 1, we can compute the log of these choice probabilities as

$$\ln \Pr(a|s_1) = \beta\tilde{V}(s_a) - C(s_1), \tag{14a}$$

$$\ln \Pr(r|s_1) = -c_{r,l} + \beta\tilde{V}(s_r) - C(s_1), \tag{14b}$$

and

$$\ln \Pr(a|s_2) = \beta\tilde{V}(s_a) - C(s_2), \tag{14c}$$

$$\ln \Pr(r|s_2) = -c_{r,h} + \beta\tilde{V}(s_r) - C(s_2), \tag{14d}$$

where $s_a = (0, 0, 1)$ denotes the system state after the patient is admitted to the ICU, and $s_r = (0, 0, 0)$ denotes the system state after the patient is rerouted to the ward. Note the choice probabilities of admission decisions in (14a) and (14c) used the assumption that the per-period cost of admitting a patient to the ICU is zero for both classes of patients.

We show the difference in rerouting costs, i.e., $c_{r,h} - c_{r,l}$, can be identified from the four choice probabilities (14a) – (14d). First, we compute the differences in choice probabilities across the two states for admission and rerouting decisions respectively. This leads to

$$\ln \Pr(a|s_1) - \ln \Pr(a|s_2) = C(s_2) - C(s_1). \tag{15a}$$

and

$$\ln \Pr(r|s_1) - \ln \Pr(r|s_2) = c_{r,h} - c_{r,l} + C(s_2) - C(s_1). \tag{15b}$$

Here we see the terms involving the value function $\tilde{V}(\cdot)$ in (14a) – (14d) are cancelled out in the differences. Further subtracting (15a) from (15b), we can get rid of the terms related to state-dependent function $C(\cdot)$ and identify the rerouting cost difference as

$$\begin{aligned} c_{r,h} - c_{r,l} &= \ln P(r|s_1) - \ln \Pr(r|s_2) - (\ln \Pr(a|s_1) - \ln \Pr(a|s_2)) \\ &= \ln\left(\frac{\Pr(r|s_1)}{\Pr(a|s_1)}\right) - \ln\left(\frac{\Pr(r|s_2)}{\Pr(a|s_2)}\right). \end{aligned} \tag{16}$$

Thus, the difference in rerouting costs can be directly identified by the difference in log ratios of rerouting to admission probabilities. This result can be interpreted as follows. The ratio of rerouting to admission probability is negatively related to the relative cost of rerouting the patient to the ward compared with admitting to the ICU. As the admission cost is assumed to be zero for both patient classes, higher rerouting to admission ratio implies a smaller rerouting cost. Thus, the difference in the log ratios can be used to identify the difference in rerouting costs of the two classes of patients. A larger difference on the right-hand side of (16) suggests the patients from the high severity class are less likely to be rerouted compared with those from the low severity class, which implies a larger difference in their rerouting costs.

The steps in (15a), (15b), and (16) show how well-constructed state and action pairs can be used to disentangle the complex structure in the choice probability expression (13). By taking differences across the two states for admission or rerouting decisions in (15a) and (15b) respectively, we remove the term $\beta\tilde{V}(\cdot)$ that is related to future payoff. Then, by taking difference between the two decisions in (16), we further remove the state-dependent terms $C(\cdot)$ to expose the cost parameters.

In the second example, we consider two states with $s_1 = (0, 1, 1)$ and $s_2 = (0, 1, 0)$. That is, there is one patient in the ICU in the first state, and one high severity patient in the ED in both states.

$$\ln \Pr(a|s_1) = \beta\tilde{V}_2 - C(s_1), \tag{17a}$$

$$\ln \Pr(r|s_1) = -c_{r,h} + \beta\tilde{V}_1 - C(s_1), \tag{17b}$$

and

$$\ln \Pr(a|s_2) = \beta\tilde{V}_1 - C(s_2), \tag{17c}$$

$$\ln \Pr(r|s_2) = -c_{r,h} + \beta\tilde{V}_0 - C(s_2), \tag{17d}$$

where $\tilde{V}_k = \tilde{V}((0, 0, k))$ denotes the value function of the state with k patients in the ICU and no patients in the ED. Computing the difference in log probabilities across the two states for admission and rerouting decisions respectively, we get

$$\ln \Pr(a|s_1) - \ln \Pr(a|s_2) = \beta(\tilde{V}_2 - \tilde{V}_1) + C(s_2) - C(s_1), \tag{18a}$$

$$\ln \Pr(r|s_1) - \ln \Pr(r|s_2) = \beta(\tilde{V}_1 - \tilde{V}_0) + C(s_2) - C(s_1), \tag{18b}$$

Then we can remove the terms related to state-dependent function $C(\cdot)$ by subtracting the above two equations, which leads to

$$\beta\left[(\tilde{V}_2-\tilde{V}_1)-(\tilde{V}_1-\tilde{V}_0)\right]=\ln\left(\frac{\Pr(a|s_1)}{\Pr(r|s_1)}\right)-\ln\left(\frac{\Pr(a|s_2)}{\Pr(r|s_2)}\right). \tag{19}$$

The left-hand side is a linear combination of value functions multiplied by the discount factor. Specifically, $\tilde{V}_2-\tilde{V}_1$ (resp. $\tilde{V}_1-\tilde{V}_0$) measures the impact on the future payoff from admitting one more patient when the ICU currently has one (resp.zero) patient. Thus, the left-hand side of (19) actually measures the change in the impact on the future payoff when adding one more patient to the ICU in its current state. Accordingly, it can be identified by the change in log ratios of admission to rerouting probability as the ICU state moves from s_1 to s_2, which have one and zero patients respectively. Again, this well-constructed state and action pair enable us to unpack the terms in the log choice probability (13). By taking differences across the two states in (18a) and (18b), we get rid of the terms related to the per-period costs. Then, by taking differences between the two actions in (19), we further remove the terms related to $C(\cdot)$ to get the desired result of the value function.

4.3. Algorithmic Approach

In this section, we document the details for how the dynamic discrete choice model is estimated from data. We employ the nested fixed-point algorithm in Rust (1987) to estimate the cost parameters (conditioning on discount factor) by maximizing the likelihood of observed choices. We note that the maximum likelihood estimator used by us and the ordinary least-squares estimator in Komarova et al. (2018) are asymptotically equivalent (see relevant proof in Miessi Sanches et al. (2016) and Pesendorfer and Schmidt-Dengler (2008)).

First, the arrival and departure rates, as well as the ED and ICU capacities, are estimated directly from data – outside of the structural model. We estimate the ED arrival rates $\lambda_{Q,i}$ and maximum arrival number M_{A_i} for $i\in\{l,h\}$ using the average and maximum number of arrivals to the ED for the two classes in each period. We estimate the ICU external arrival rate λ_E using the average number of patients admitted to ICU in each time slot who are not included in our low and high severity ED groups. The departure probability μ_I is estimated as the ratio of total number of departures to the total periods of ICU stay across all ICU patients. The ICU capacity B is set to be the maximum number of patients in the ICU observed from data. It includes both medical and surgical, emergency and elective patients, to reflect the true maximum ICU occupancy. We also tested other choices of B to show the robustness of our estimation results.

Note that our data captures the number of patients admitted to the hospital from the ED, but does not include any patients who are discharged from the ED (e.g. to home or to a skilled nursing facility), but who inevitably utilize ED resources. Thus, it is difficult to accurately determine the maximum number of admitted patients allowable in the ED, i.e. the ED capacity in our model. Given this challenge, we set ED capacities Q_i using the following heuristic:

$$Q_i=M_{Q_i}+\left\lfloor\sqrt{M_{A_i}}\right\rfloor, \tag{20}$$

where M_{Q_i} is the maximum number of patients in the ED observed in the data; M_{A_i} is the maximum number of arrivals in each period; and $\lfloor \cdot \rfloor$ denotes the floor function. We introduce the square root term $\sqrt{M_{A_i}}$ as a "safety buffer" to ensure we have ample ED capacity to avoid balking upon arrival to the ED, as patients are rarely turned away from the hospital at this stage. For appropriately loaded queueing systems, it is well known that stochastic fluctuations of the queue length are on the order of the square root of the average offered load (see e.g., Halfin and Whitt (1981)). In our setting, we could approximate the average offered load by the arrival rates $\lambda_{Q,i}$ since the ED waiting time is generally very short. We take a more conservative approach and use the square root of the maximum number of arrivals M_{A_i} as the "safety buffer". We verify by simulation that the ED rarely reaches its full capacity (20) in our structural model; thus, while the ED has a finite capacity, patients rarely balk. We also find that the choice probabilities are very robust to alternative specifications of the ED capacity, which is not the case when varying the ICU capacity B. This suggests that our structural model primarily captures the interplay between ICU congestion and the importance the decision-making team places on intertemporal externalities when making admission decisions.

The remaining parameters – the discount factor, waiting and rerouting costs for the two classes, i.e., $\theta = \{\beta, c_{r,l}, c_{r,h}, c_{w,l}, c_{w,h}\}$ – are estimated within the structural model using the observed states and actions. Given the observed state and action sequences $\{s_t, d_t\}$ for $t = 1, 2, ..., T$, the likelihood for a fixed set of parameters, θ, is given by

$$l^f(s_1, ..., s_T, d_1, ..., d_T|\theta) = \prod_{t=1}^{T} f(d_t|s_t, \theta) g(s_{t+1}|\varphi(s_t, d_t)), \tag{21}$$

where $f(d_t|s_t, \theta)$ denotes the choice probability in (10) given parameter θ. The state transition probability $g(s_{t+1}|\varphi(s_t, d_t))$ is explicitly given in the Appendix A.1. The likelihood l^f can be decomposed into two parts:

$$l^f(s_1, ..., s_T, d_1, ..., d_T|\theta) = l^d(s_1, ..., s_T, d_1, ..., d_T|\theta) \cdot l^s(s_1, ..., s_T, d_1, ..., d_T),$$

where l^d is the part of l^f associated with the choice probabilities, given by

$$l^d(s_1, ..., s_T, d_1, ..., d_T|\theta) = \prod_{t=1}^{T} f(d_t|s_t, \theta), \tag{22}$$

and l^s is the part of l^f from the state transition, i.e.,

$$l^s(s_1, ..., s_T, d_1, ..., d_T|\theta) = \prod_{t=1}^{T} g(s_{t+1}|\varphi(s_t, d_t)).$$

We see that the structural parameter θ is only involved in the likelihood function l^d for the choice probabilities, but not the likelihood function l^s for the state transitions, which only depends on the arrival and departure rates.

Our estimation approach is based on the nested fixed point algorithm in Rust (1987). The estimation procedure consists of two loops: The "inner" loop computes the function $\tilde{V}$ for a fixed θ, and the "outer" loop searches for the value of θ that maximizes the log-likelihood $\ln l^f$ in (21). Since the partial likelihood l^d in (22) is the only part in l^f that involves θ, maximizing the full log-likelihood is equivalent to finding the value of θ that maximizes $\ln l^d$. In the "inner" loop, the unknown function $\tilde{V}$ is computed by value iteration on the functional equation (12). In the "outer" loop, we use a gradient descent algorithm to find the optimal parameter θ. To reduce the computational burden, we restrict the potential values of the discount factor to a discrete grid $\beta = \{0.1, 0.2, ..., 0.9\}$. While coarse, this discrete grid is granular enough to measure how the hospital internalizes the intertemporal externalities, and enables us to determine whether they are more cognizant of the near versus longer-term when making decisions. To summarize, for each candidate value of β, we estimate the cost parameters $\{c_{r,l}, c_{r,h}, c_{w,l}, c_{w,h}\}$ that maximizes the log-likelihood $\ln l^d$. Then, we choose the discount factor and its associated cost estimates that lead to the largest likelihood among all candidate β.

With the estimated $\hat{\theta}$, the standard deviation of the parameters are computed as

$$\text{Sd}(\hat{\theta}_i) = 1/\sqrt{\mathcal{I}(\hat{\theta})_{ii}},$$

where $\mathcal{I}(\hat{\theta})$ is the Fisher's information matrix

$$\mathcal{I}(\hat{\theta}) = \mathsf{E}\left[\frac{\partial \ln l^d(\hat{\theta})}{\partial \hat{\theta}}\left(\frac{\partial \ln l^d(\hat{\theta})}{\partial \hat{\theta}}\right)^{\top}\right].$$

To examine the proportion of variation explained by our structural model, we compute the McFadden's pseudo R-squared as

$$\text{Pseudo } R^2 = 1 - \frac{\ln l^d(\hat{\theta})}{\ln l^{null}}, \tag{23}$$

where l^{null} is the "null" likelihood from a multinomial logistic regression model with only an intercept term and hospital fixed-effects, i.e., the action probabilities do not depend on system states (See Appendix A.3 for details of the null model).

5. Estimation Results

We present the estimation results from the structural model in this section. Before we discuss the detailed results, we first describe how we apply the structural model to the data in our study setting and provide some preliminary findings using the system summary statistics calculated from the data and the parameters estimated outside the structural model.

We define each period to be a two hour time interval. Recall that only 12% of all ED patients are admitted to the ICU; there are some patients (e.g. those with very low LAPS2 scores) who are likely to be 'rerouted' to the ward regardless of ICU bed availability. Thus, in order to understand the impact of intertemporal

externalities on the ICU admission decision, we must identify a group of high severity patients with sufficiently high likelihood of ICU admission. Additionally, we require enough observations in the high severity class to effectively estimate the rerouting and waiting costs, $c_{r,h}$ and $c_{w,h}$, as well as to be able to identify the discount factor. We partition the ED patients into two classes by their LAPS2 score, as it has the highest correlation with the ICU admission decision among all severity scores. We define the low severity class as patients with LAPS2 score in the range of $[0, 113]$, corresponding to those below the 85th percentile of the LAPS2 score distribution, and the high severity class as patients with LAPS2 score in the range $(113, 294]$, corresponding to those above the 85th percentile. We note that this means any medical patient admitted via the ED is included in our structural model as either a low or high severity patient. While we expect the ICU admission rate for low severity patients to be lower, some of these patients will be admitted to the ICU possibly due to factors that are unobservable in the data, thereby introducing intertemporal externalities on the system. Indeed, we find that while 8% of low severity patients are admitted to the ICU (compared to 34% of high severity patients), they contribute more (58%) to the total number of medical patients admitted from ED than the high severity class (42%). This highlights the importance of including the costs of both severity classes in our model, instead of just the high severity class. If we ignored the low severity patients, we would significantly underestimate the impact of the ICU admission decision on future patients.

Recall that our structural model considers the admission decisions observed at each hospital. Within each hospital, multiple physicians and administrators cover the ED and ICU over the study period. Due to limitations in data availability, we only know the hospital where the decision is made, but have no information on any individuals or specific system constraints which potentially impact the decision. Thus, we can only estimate our model at the hospital level even though there may be many different factors and systems involved in the decision making process in practice. Therefore, the estimation results will reflect the 'average' behavior within each hospital. We note that the hospital level result will tend to underestimate the actual variation in the systems' behaviors, as it ignores the potential heterogeneity among systems (e.g. night versus day, weekday versus weekend, ability to flex beds, etc.) within the same hospital. Thus, if we see large heterogeneity across hospitals, this would imply there is likely even larger heterogeneity when considering different system level factors, which we do not observe in the data.

In Table 4, we provide the system summary statistics of each hospital, including the ED and ICU capacities, average ICU occupancy, arrival and departure rates, and overall ICU admission probabilities for two patient classes. The ICU admission probability is the proportion of ED patients who are eventually admitted to the ICU regardless of their waiting time. As we already restrict our final study cohort to the patients whose next unit is either the ICU or ward (including the TCU if the hospital has one), the admission probability can be computed by $N_{ICU}/(N_{ICU} + N_{ward})$, where N_{ICU} (N_{ward}) denotes the total number of patients that are admitted (rerouted). All the statistics in Table 4 are estimated directly from data outside of the structural model.

Table 4 System summary statistics by hospital

Hosp	Q_l	Q_h	B	ICUOccu	$\lambda_{Q,l}$	$\lambda_{Q,h}$	λ_E	μ_I	$\Pr(a_l)$	$\Pr(a_h)$
1	11	5	21	0.67	1.230	0.221	0.252	0.035	0.12	0.41
2	14	7	26	0.76	1.429	0.227	0.268	0.026	0.11	0.36
3	8	4	12	0.49	0.859	0.146	0.101	0.030	0.05	0.22
4	13	7	31	0.71	1.529	0.327	0.519	0.031	0.05	0.29
5	7	3	11	0.65	0.633	0.098	0.108	0.030	0.11	0.41
6	9	5	21	0.58	1.123	0.208	0.278	0.036	0.08	0.36
7	7	4	11	0.71	0.583	0.111	0.188	0.030	0.05	0.18
8	10	6	16	0.67	0.868	0.185	0.158	0.033	0.14	0.40
9	15	5	22	0.71	1.807	0.264	0.354	0.036	0.07	0.31
10	8	4	12	0.52	0.714	0.092	0.228	0.048	0.06	0.28
11	6	4	7	0.55	0.284	0.045	0.056	0.029	0.12	0.46
12	12	6	24	0.69	1.256	0.173	0.299	0.026	0.06	0.33
13	9	4	16	0.50	0.817	0.182	0.110	0.028	0.07	0.30
14	11	6	36	0.72	1.332	0.257	0.493	0.026	0.07	0.33
15	8	4	16	0.43	0.897	0.111	0.114	0.031	0.07	0.31
16	8	4	13	0.44	0.752	0.133	0.108	0.033	0.06	0.26
17	7	4	9	0.62	0.418	0.099	0.066	0.024	0.09	0.32
18	9	6	32	0.58	0.796	0.122	0.557	0.034	0.06	0.26
19	8	4	25	0.46	0.966	0.170	0.196	0.031	0.10	0.37
20	8	4	11	0.34	0.545	0.087	0.065	0.035	0.07	0.29
21	8	4	13	0.68	0.451	0.076	0.139	0.025	0.10	0.44
22	8	4	16	0.62	0.422	0.090	0.182	0.028	0.15	0.45

System summary statistics for each hospital: Q_i for $i \in \{l, h\}$ is the ED capacity for the two classes of patients; B is the ICU capacity; ICUOccu is the average ICU occupancy level; $\lambda_{Q,i}$ for $i \in \{l, h\}$ is the ED arrival rate; λ_E is the external arrival rate to ICU; μ_I is the ICU departure rate; $\Pr(a_i)$ for $i \in \{l, h\}$ is the overall admission probability for the ED patients.

We note the following observations from Table 4. First, the ICUs in the hospitals are generally congested. The average ICU occupancy in most hospitals is higher than 50%. For some of them (e.g., hospitals 2 and 14), this number is even higher than 70%. Second, the ICU admission probability for the high severity class is above 30% for most hospitals, and is usually three or four times larger than that for the low severity class. This implies the costs parameters of the two classes should be very different, which is indeed captured in our structural model. Finally, we see the hospitals in KPNC are very different regarding their sizes, work loads, and admission behaviors. For example, large hospitals have more than 30 beds in their ICU, while small hospitals have fewer than 10. Additionally, the ICU admission probabilities can be very different across hospitals even for the same severity class of patients. With such large heterogeneity in the system statistics, hospitals may also behave very differently when they make admission decisions. In the next section, we quantify this heterogeneity with empirical evidence from our structural model.

5.1. Intertemporal Externalities

5.1.1. Main Estimation Results In this section, we provide the estimation results of our structural model, i.e., the estimated discount factor and costs. Before estimating the model for each hospital separately, we first show the estimation results for all hospitals combined. That is, we estimate one set of parameters that maximizes the sum of log-likelihood from all hospitals. The results are summarized in the table below.

Table 5 Estimation results of structural model: All hospitals combined ($N = 154,140$ hospital-periods)

Discount factor	Low Severity		High Severity		
$\hat{\beta}$	$\hat{c}_{w,l}$	$\hat{c}_{r,l}$	$\hat{c}_{w,h}$	$\hat{c}_{r,h}$	R^2
0.3^{***}	0.071^{***}	-1.950^{***}	0.932^{***}	-0.671^{***}	0.18
(0.003)	(0.007)	(0.007)	(0.015)	(0.012)	

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

The first and second rows report the estimated parameters and their standard errors in parentheses. The last column provides the McFadden's pseudo R^2. The estimated discount factor is $\hat{\beta} = 0.3$, and is significantly different from adjacent levels 0.2 and 0.4 at the 1% level. The estimated waiting costs $\hat{c}_{w,l}$ and $\hat{c}_{w,h}$ are both significantly positive, while the rerouting costs $\hat{c}_{r,l}$ and $\hat{c}_{r,h}$ are both significantly negative. The McFadden's pseudo R^2 from the structural model is 0.18, which is higher than the level ($R^2 = 0.11$) from the multinomial logit regression in Table 2.

At first glance, the estimated $\hat{\beta}$ is quite surprising. In most of the empirical literature, the discount factor is usually assumed to be relatively large, e.g., 0.90 or 0.95. However, we see here the estimated $\hat{\beta}$ is much smaller than these levels. This provides additional evidence that the level of discount factor may vary dramatically in different empirical settings (Frederick et al. (2002)). Thus, it is crucial to identify the discount factor using real data instead of assuming a pre-specified value. In our model, the relatively small value of $\hat{\beta}$ implies the hospitals are *not* very forward-looking when making ICU admission decisions. Given each period in our model is a two hour interval, the result suggests that the hospitals barely consider the impact of their decisions on the system beyond the next six hours (after three 2-hour periods, $0.3^3 \approx 0.03$). This interpretation appears to be somewhat consistent with the 12-hour shift schedules at KPNC. That is, the hospitals indeed account for the future when making decisions, but they mostly internalize the impact of their decisions on the system state within their own shifts rather than considering the longer-term.

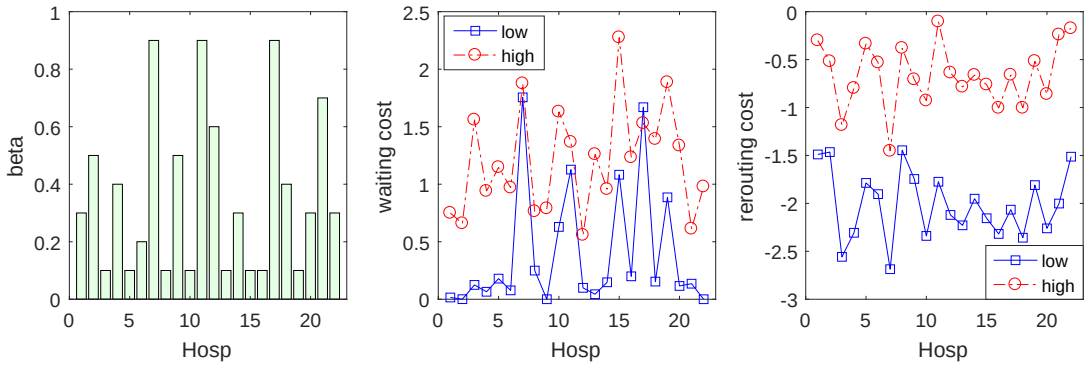
Unlike the waiting costs, which are constrained to be non-negative, the rerouting costs are negative and significant for both the low and high severity classes. The negative rerouting costs suggest that it is more desirable to reroute a large fraction of patients. Indeed, this seems reasonable for low severity patients for whom ICU care is expensive, and (likely) unnecessary. This turns out to also be the case for some of the patients in the high severity group, which includes all patients with a LAPS2 score higher than the 85th percentile. This is also supported by the observed admission probability $\Pr(a_h)$ in Table 4: The admission

probability for the high severity class is less than 50% in all hospitals, which means that the majority of these patients are rerouted to the ward.

Finally, both the waiting and rerouting costs are significantly higher for the high severity class than for the low severity class. As expected, it is more costly (financially, operationally, and clinically) for the high severity patients to wait in the ED, i.e., $\hat{c}_{w,h} > \hat{c}_{w,l} > 0$, and the low severity class on average benefits more from being rerouted to the ward, i.e., $\hat{c}_{r,l} < \hat{c}_{r,h} < 0$. The apparent differences in cost parameters highlights the importance of differentiating between the two severity classes.

The estimated $\hat{\beta} = 0.3$ represents the average behavior across all hospitals; however, it does not provide information about potential differences across hospitals. If all hospitals behave quite similarly when internalizing the intertemporal externalities on admission decisions, most of them should have a discount factor close to the average level 0.3. On the other hand, if there is large heterogeneity in the discount factors, e.g., some hospitals with $\hat{\beta} = 0.9$ and some with $\hat{\beta} = 0.1$, this would imply that some hospitals are quite concerned with the longer-term while others focus more on the near-term. To address this question, we estimate the structural parameters $\theta = \{\beta, c_{r,l}, c_{r,h}, c_{w,l}, c_{w,h}\}$ for each hospital individually. The results are summarized in Table 6. The second column reports the number of periods (two-hour snapshots) in each hospital, and the last column provides the McFadden's pseudo R^2. We also plot the estimated discount factors and costs in Figure 3.

Figure 3 Comparison of estimated discount factors and costs across the 22 individual hospitals.



Note: The left panel shows the estimated discount factor, the middle and right panels show the estimated waiting and rerouting costs, respectively, for the low (blue solid line) and high (red dotted line) severity patients.

First, we see substantial heterogeneity in the estimated discount factors across hospitals. In particular, we have 13 out of 22 hospitals with relatively small estimated discount factors $\hat{\beta} \in \{0.1, 0.2, 0.3\}$, five with medium discount factors $\hat{\beta} \in \{0.4, 0.5, 0.6\}$, and the other four with relatively large discount factors $\hat{\beta} \in \{0.7, 0.8, 0.9\}$. All $\hat{\beta}$ estimates are significantly different from adjacent levels. Thus, ICU admission dynamics are very different across hospitals: Some of them have relatively small, near-term discount factors

Table 6 Estimation results of structural model by individual hospital

Hosp	Num. of periods	Discount Factor	Low Severity		High Severity		
		$\hat{\beta}$	$\hat{c}_{w,l}$	$\hat{c}_{r,l}$	$\hat{c}_{w,h}$	$\hat{c}_{r,h}$	R^2
1	8,016	0.3^{***}	0.015	-1.490^{***}	0.749^{***}	-0.301^{***}	0.19
		(0.012)	(0.021)	(0.022)	(0.048)	(0.043)	
2	6,012	0.5^{***}	0.001	-1.465^{***}	0.659^{***}	-0.519^{***}	0.23
		(0.010)	(0.017)	(0.024)	(0.047)	(0.051)	
3	8,016	0.1^{***}	0.124^{*}	-2.558^{***}	1.560^{***}	-1.185^{***}	0.08
		(0.019)	(0.051)	(0.040)	(0.128)	(0.065)	
4	8,016	0.4^{***}	0.065^{**}	-2.307^{***}	0.940^{***}	-0.798^{***}	0.15
		(0.010)	(0.024)	(0.029)	(0.046)	(0.039)	
5	6,924	0.1^{***}	0.179^{***}	-1.786^{***}	1.147^{***}	-0.336^{***}	0.10
		(0.023)	(0.046)	(0.037)	(0.107)	(0.072)	
6	8,016	0.2^{***}	0.077^{*}	-1.904^{***}	0.970^{***}	-0.531^{***}	0.14
		(0.014)	(0.030)	(0.027)	(0.063)	(0.046)	
7	6,948	0.9^{***}	1.755^{***}	-2.687^{***}	1.876^{***}	-1.453^{***}	0.10
		(0.016)	(0.047)	(0.075)	(0.093)	(0.095)	
8	7,848	0.1^{***}	0.249^{***}	-1.446^{***}	0.764^{***}	-0.380^{***}	0.15
		(0.020)	(0.034)	(0.026)	(0.062)	(0.048)	
9	6,180	0.5^{***}	0.000	-1.745^{***}	0.789^{***}	-0.707^{***}	0.18
		(0.009)	(0.017)	(0.025)	(0.048)	(0.048)	
10	7,320	0.1^{***}	0.629^{***}	-2.339^{***}	1.630^{***}	-0.929^{***}	0.07
		(0.027)	(0.067)	(0.046)	(0.160)	(0.081)	
11	8,016	0.9^{***}	1.127^{***}	-1.774^{***}	1.366^{***}	-0.104	0.08
		(0.017)	(0.036)	(0.064)	(0.088)	(0.106)	
12	4,668	0.6^{***}	0.099^{***}	-2.121^{***}	0.558^{***}	-0.637^{***}	0.22
		(0.010)	(0.023)	(0.040)	(0.052)	(0.068)	
13	8,016	0.1^{***}	0.044	-2.229^{***}	1.260^{***}	-0.785^{***}	0.13
		(0.018)	(0.043)	(0.035)	(0.088)	(0.052)	
14	8,016	0.3^{***}	0.147^{***}	-1.951^{***}	0.956^{***}	-0.662^{***}	0.19
		(0.012)	(0.027)	(0.027)	(0.054)	(0.042)	
15	6,912	0.1^{**}	1.082^{***}	-2.155^{***}	2.275^{***}	-0.762^{***}	0.10
		(0.033)	(0.074)	(0.042)	(0.195)	(0.075)	
16	8,016	0.1^{***}	0.199^{***}	-2.319^{***}	1.233^{***}	-1.005^{***}	0.11
		(0.020)	(0.050)	(0.039)	(0.108)	(0.063)	
17	6,588	0.9^{***}	1.669^{***}	-2.065^{***}	1.530^{***}	-0.660^{***}	0.09
		(0.021)	(0.049)	(0.067)	(0.077)	(0.085)	
18	8,016	0.4^{***}	0.154^{***}	-2.358^{***}	1.391^{***}	-1.005^{***}	0.04
		(0.013)	(0.033)	(0.039)	(0.099)	(0.070)	
19	6,576	0.1^{***}	0.886^{***}	-1.809^{***}	1.885^{***}	-0.517^{***}	0.12
		(0.029)	(0.057)	(0.035)	(0.126)	(0.058)	
20	8,004	0.3^{***}	0.116^{**}	-2.260^{***}	1.335^{***}	-0.862^{***}	0.09
		(0.017)	(0.041)	(0.042)	(0.115)	(0.078)	
21	4,008	0.7^{***}	0.135^{***}	-2.002^{***}	0.612^{***}	-0.240^{*}	0.12
		(0.012)	(0.030)	(0.064)	(0.071)	(0.111)	
22	4,008	0.3^{***}	0.000	-1.513^{***}	0.979^{***}	-0.175	0.09
		(0.025)	(0.047)	(0.051)	(0.116)	(0.099)	

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

– focusing primarily on the individual patient in front of them – while others have relatively large, longer-term discount factors – accounting for the impact of their current decisions on the ability to treat other patients later. Such heterogeneity in the discount factor reflects the behavioral variation across hospitals in different hospitals.

As we have discussed before, there is much debate on how much hospitals internalize the intertemporal externalities when making admission decisions, and supporting evidence exists for both near versus longer term aspects of their behaviors. On one hand, physicians are trained to provide timely and appropriate care to their patients. On the other hand, hospitals need to manage the occupancy level in the often congested ICU to reserve enough capacity for future, perhaps more severe patients. Moreover, effectively accounting for the future can be difficult as members of the hospitals (e.g. physicians and administrators) are rarely trained to think about the complex system dynamics that arise in hospitals. Our findings show that both immediate and future considerations can influence the hospital's behavior, and the overall effect can vary substantially across hospitals. This finding reveals an important behavioral perspective of the decision-making process for ICU admission and provides additional explanation for observed practice variation, which has received much attention from medical professionals in recent literature (Westert et al. 2018, Corallo et al. 2014). Indeed, a number of studies suggest that identifying the patients that will benefit from ICU care is highly subjective and depends on a physician's own training and experience (Fisher et al. (2004), Mullan (2004), O'Connor et al. (2004), Chen et al. (2012)); our work suggests that some of the variation observed in practice may also be due to variations in how hospitals internalize intertemporal externalities.

The pattern of estimated rerouting and waiting costs in Table 6 are very similar to that for all hospitals combined. Most of the waiting costs are significant and positive, including 17 out of 22 estimates for $\hat{c}_{w,l}$ and all estimates for $\hat{c}_{w,h}$. The rerouting costs are all significant and negative, except for $\hat{c}_{r,h}$ of Hospital 11. In most cases, the rerouting and waiting costs for high severity class are significantly higher than those for low severity class (except for waiting costs in Hospitals 8 and 11). The costs estimates vary substantially across hospitals. Recall these parameters represent the average costs measured relative to the admission decision in each hospital. Such variation suggests that there is large heterogeneity across hospitals in their medical resources and the degree of severity of their patient population.

With the estimated discount factors and costs parameters, it is important to show our structural model can indeed capture the intertemporal externalities on ICU admission decisions. We compute the admission probability at different ICU states using our estimated structural model and test whether it decreases as the ICU gets more congested. As an illustrative example, we consider a representative ED state with $n^E_{l,t} = \lfloor Q_l/2 \rfloor$ and $n^E_{h,t} = 1$, i.e., several low severity class patients and one high severity patient[2]. We focus on the admission decision for the high severity class patient, who is more likely to be admitted to the ICU. We

[2] We find that the admission probability is practically insensitive to the particular choice of $n^E_{l,t}$.

compute the reduction in her admission probability as the ICU occupancy moves from half full to almost full; that is,

$$\text{AdmDrop} = \Pr\left(a_{h,t}=1|n_t^I=\lfloor B/2\rfloor\right)-\Pr\left(a_{h,t}=1|n_t^I=B-1\right). \tag{24}$$

We also compute the relative drop as

$$\text{Rel.AdmDrop} = \frac{\Pr\left(a_{h,t}=1|n_t^I=\lfloor B/2\rfloor\right)-\Pr\left(a_{h,t}=1|n_t^I=B-1\right)}{\Pr\left(a_{h,t}=1|n_t^I=\lfloor B/2\rfloor\right)}. \tag{25}$$

The results are summarized in Table 7 below.

Table 7 Drop in admission probability as ICU gets congested: e.g. increasing from 50% occupancy to having only 1 available bed

Hosp	$\hat{\beta}$	AdmDrop	Rel. AdmDrop
1	0.3	0.047	0.142
2	0.5	0.047	0.170
3	0.1	0.013	0.060
4	0.4	0.021	0.082
5	0.1	0.036	0.099
6	0.2	0.031	0.099
7	0.9	0.017	0.105
8	0.1	0.047	0.139
9	0.5	0.036	0.140
10	0.1	0.018	0.068
11	0.9	0.055	0.150
12	0.6	0.025	0.104
13	0.1	0.021	0.075
14	0.3	0.029	0.101
15	0.1	0.024	0.078
16	0.1	0.018	0.072
17	0.9	0.047	0.171
18	0.4	0.018	0.074
19	0.1	0.035	0.100
20	0.3	0.021	0.081
21	0.7	0.030	0.106
22	0.3	0.047	0.127

The estimated discount factor and drop in admission probability: AdmDrop and Rel.AdmDrop denote the absolute and relative admission probability drop given in (24) and (25), respectively.

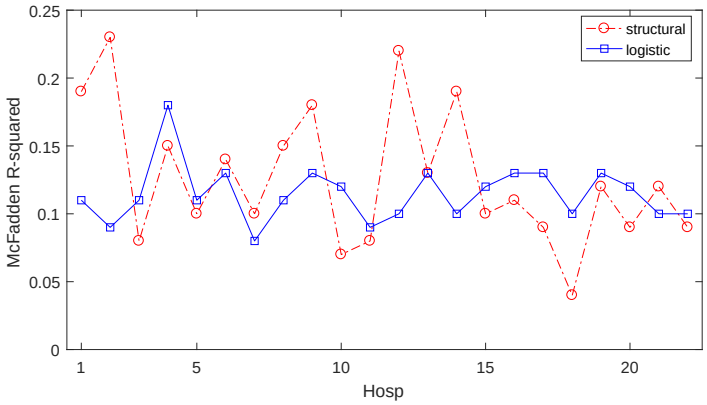
We can see the ICU state has a substantial impact on the admission decision. On average, the hospitals see a 11% relative drop in the admission probability as the ICU becomes almost full. For some hospitals, the drop can be as large as 17%. This effect can be larger when the ED is more congested with more high severity patients, as more ICU beds can be potentially occupied by these patients, impacting future ICU bed availability. For example, if we have two high severity class patients in the ED, the average and maximum

reduction in probability of admitting both of them (as n_t^I changes from $\lfloor B/2 \rfloor$ to $B-2$) increase to 15% and 25%, respectively. We also find the relative drop and estimated $\hat{\beta}$ are highly correlated. Their correlation is 0.557, which is statistically significant at the 1% level. This can be interpreted as follows: In the hospitals with larger discount factors, the system is more forward-looking in making admission decisions, thus the admission probabilities in these hospitals are more sensitive to the ICU occupancy level, leading to larger drops as the ICU become congested. This relationship confirms that the discount factors in our structural model indeed capture the differences in the length of the horizon (near versus longer term) for which the hospital considers when making admission decisions.

5.1.2. Goodness-of-Fit for Structural Model While the structural model can help bring behavioral insights into the hospitals' decisions, it is important to check whether the estimated model fits the data well. In this section, we show our structural model provides good estimates to both the hospitals' decisions and the system states.

First, we compare the explanatory power for the hospitals' decisions, as measured by the McFadden's pseudo R^2 in (23), from our structural model with that from the reduced-form multinomial logistic regression model in Section 3. The results are shown in Figure 4 below.

Figure 4 Comparison of McFadden's pseudo R^2 from structural and multinomial models



Note: The red dotted and blue solid lines report the McFadden's pseudo R^2 for each hospital from the structural model and the multinomial model, respectively.

We see our structural model has higher or comparable McFadden's pseudo R^2 to the comprehensive multinomial logistic model for most hospitals. The average McFadden's pseudo R^2 is also higher in the structural model (0.13 versus 0.11) than the combined hospital multinomial logistic model. The R^2 values are not very high in both models. This is not too surprising as the hospitals consider many factors that are not recorded in the data when making admission decisions; that is, the decisions appear very "noisy" in the data. We note that the multinomial model contains most variables one could expect to influence admission

decisions, including patient's characteristics (gender, age, and three severity scores), system states (ED and ICU occupancy, departures in previous period, and average ICU severity level), as well as seasonality fixed effects related to day/night and weekday/weekend. On the other hand, our structural model is quite parsimonious with only five free parameters (discount factor and costs parameters), but is still able to explain similar (or more) variation in the data. Thus, our structural model appears to have reasonable explanatory power in capturing the hospitals' decisions. Of course, we acknowledge that there is still quite a bit of variation that our data cannot capture (e.g. perhaps due to availability of system-level interventions to increase capacity or stabilize and treat patients outside of the ICU.

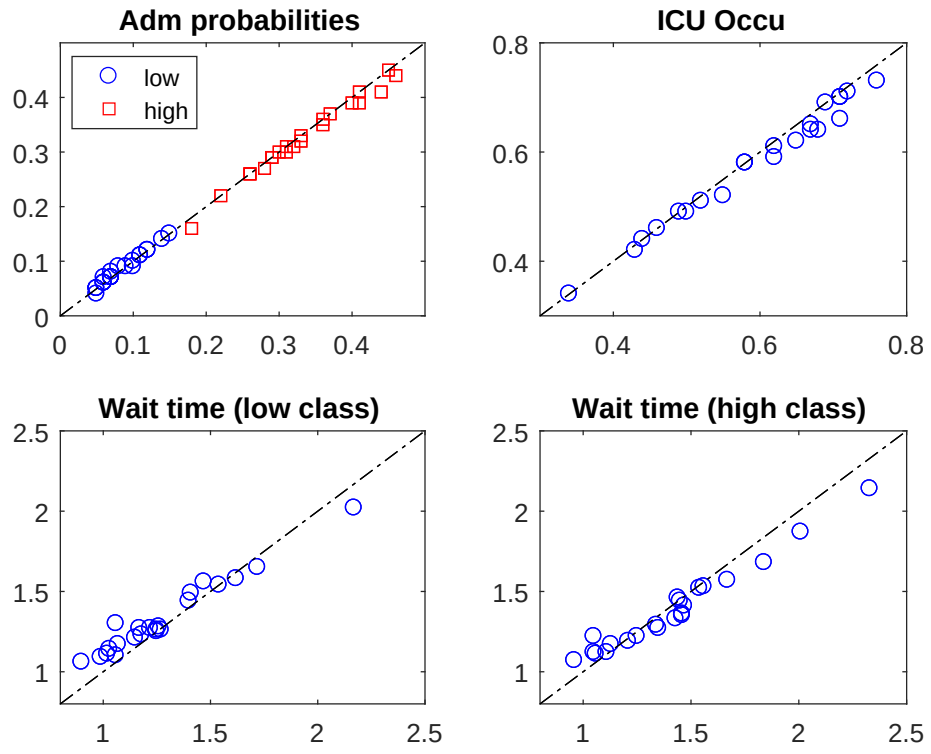
Next, we show by simulation that our structural model produces system statistics close to those observed in the data. As the arrival and departure rates are directly calibrated from data, we expect the average number of arrivals and departures in each period of our structural model to be close to that observed in the data. Therefore, we focus on other important system statistics including the average ICU occupancy, overall proportion admitted to the ICU (i.e., $N_{ICU}/(N_{ICU} + N_{ward})$), as well as the ED waiting times of each patient class. The statistics estimated from our structural model are averaged over 100 simulation runs. Each run contains the same number of periods as in the data plus an additional three-month warm up period to allow the system to reach steady-state. The warm up period is dropped when computing the system statistics. The comparison of the system statistics is shown in Figure 5.

In each panel, each point represents a hospital in our study; its x-coordinate (y-coordinate) corresponds to the observed (simulated) value of the system statistic. We plot the 45-degree line in each panel, which represents a perfect fit. As we can see, most points fall close to the identity line, implying that our estimated structural model produces system statistics that are close to the observed data. Although our structural model is trained to fit the choice probabilities of the hospital's actions, it also leads to system dynamics that fit a number of observed metrics very well. This further supports its effectiveness in modeling the admission process for ED patients.

5.1.3. Heterogeneity in Discount Factors We have seen in the previous section there is large heterogeneity in the discount factors across hospitals, i.e., some hospitals appear to be relatively focused on the near-term in making admission decisions, while others appear to consider more longer-term dynamics. Such heterogeneity in the perceived discout factor is important for understanding system-level admission decisions. There are many potential reasons for this heterogeneity, such as the different physicians and administrators in the hospital as well as system-level factors (e.g. the ability to flex capacity and/or the amount of demand from other sources) which differentially influence the occupancy challenges at each hospital. In this section, we look into the heterogeneity in discount factors in more depth. We identify possible reasons that explain the heterogeneity, and examine how the heterogeneity impacts system performance.

We start by computing the correlations between the estimated discount factors and other system statistics across the 22 hospitals. The results are shown in Table 8. We show the correlations between $\hat{\beta}$ and the

Figure 5 Comparison of system statistics from structural model and real data



Note: The figure compares the system statistics simulated from the structural model (y-coordinate) and observed from the real data (x-coordinate).

observed ICU departure rate μ_I, average ICU occupancy level, average ED waiting time EDWait_i for $i \in \{l, h\}$, and the increase in waiting time for admitted patients due to ICU congestion, i.e., $\Delta\text{EDWait\_Adm}_i$. This increase is defined as the difference between average ED waiting time of patients admitted to the ICU when the occupancy level is below the 70th percentile versus when it is above the 95th percentile.

Table 8 Correlations between estimated discount factor, $\hat{\beta}$, and system statistics from data

μ_I	ICUOccu	EDWait_l	EDWait_h	$\Delta\text{EDWait\_Adm}_l$ with ICU congestion	$\Delta\text{EDWait\_Adm}_h$ with ICU congestion
-0.428^{*}	0.445^{*}	0.545^{**}	0.581^{**}	0.475^{*}	0.566^{**}

$^{*}p < 0.05$, $^{**}p < 0.01$ and $^{***}p < 0.001$. Correlations between estimated discount factor and the observed ICU departure rate (μ_I), ICU occupancy (ICUOccu), average ED waiting time (EDWait_i), and increase in ED waiting time for admitted patients under ICU congestion ($\Delta\text{EDWait\_Adm}_i$).

It is not obvious ex-ante in which direction ICU congestion impacts the hospitals' admission behaviors. On one hand, when the ICU is busy, hospitals have to be more judicious when making bed allocation decisions as access issues could substantially jeopardize quality of care, which suggests they should be

more forward-looking when making decisions. For example, if the hospital expects the ICU will be highly congested, they may choose to save some ICU beds for future (potentially sicker) patients by delaying the admission of current ED patients. On the other hand, busier ICUs also tends to have more external arrivals, such as non-ED surgical patients. This introduces increased competing demand for ICU beds from external arrivals, which may reduce the hospital's motivation or ability to save ICU beds for future ED patients (the cohort for which our model captures) as the beds will likely be utilized by external arrivals. Our results suggest that, on average, the first effect dominates the second: hospitals with lower ICU departure rates or higher ICU occupancy levels tend to have larger discount factors and these hospitals tend to be more congested. One possible explanation is that the majority of the external arrivals ($> 70\%$) are surgical patients whose ICU beds are often reserved in advance to accommodate their post-surgery recovery, so hospitals may not necessarily account for the competing demand from external arrivals when making ICU admission decisions.

We also find that the discount factors are positively correlated with the average ED waiting time of both classes. This makes sense as hospitals who account more for the longer-term are more likely to save beds for future patients, thereby increasing the waiting time of some ED patients.

The discount factors are also positively correlated with the increase in ED waiting time for admitted patients when the ICU is congested. This is not surprising because hospitals with longer-term discounting behaviors tend to be more sensitive to ICU congestion when making admission decisions and are more likely to delay patients when the unit is congested. These results are consistent with the relationship between $\hat{\beta}$ and the relative admission probability drop (25) discussed at the end of Section 5.1.1. These results lend further support that our structural model is able to capture the intertemporal externalities on admission decisions.

Finally, we calculate the correlation between the estimated $\hat{\beta}$ from the structural model and the estimated coefficient γ_{ICU} from the multinomial logit model (1) for the admission decision, i.e., with $d_{it} = Adm_{it}$ on the left-hand side. Recall γ_{ICU} measures the association of the ICU occupancy level (measured by percentile rank) with the admission decision. We saw earlier that it is negative in most hospitals, suggesting the likelihood of ICU admission is lower when the ICU becomes more congested. For hospitals with larger $\hat{\beta}$, their admission probabilities tend to be more sensitive to ICU congestion, leading to more negative values of γ_{ICU}. Thus, we expect to see a negative correlation between $\hat{\beta}$ and γ_{ICU} across hospitals. This is indeed the case:

$$\text{Corr}(\hat{\beta}, \gamma_{ICU}) = -0.771^{***},$$

which is significant at the 0.1% level. The result also holds when we consider the coefficient γ_{ICU} in the binary logit model with only admission and rerouting decisions: The correlation in this case is -0.762, which is also significant at the 0.1% level. The significant and negative relationship between $\hat{\beta}$ and γ_{ICU}

confirms that the heterogeneity observed in our structural model is consistent with the reduced-form evidence: The hospitals who are identified as being more longer-term focused in our model indeed react more to the ICU occupancy level when admitting patients. This provides additional support for our structural model as an effective approach to estimate how hospitals internalize the intertemporal externalities on their admission decisions.

5.2. Counterfactual Simulations

One of the most important advantages of using the structural estimation approach is the ability to conduct counterfactual studies via simulation. This allows us to evaluate the effect of different policies or interventions that can not be directly observed in the data or through reduced-form regression analyses. We consider two counterfactual studies, which focus on potential changes in ICU resources (e.g. beds plus necessary staff) and the hospital's admission behaviors (e.g. how it reduces admission probabilities due to ICU congestion). The counterfactual results can provide insights into the impact various interventions may have on medical and economic outcomes and system performance metrics.

In the first counterfactual, we measure the impact of adding one bed in ICU while assuming the hospitals' behaviors remain unchanged. That is, we increase the ICU capacity by one in each hospital, and use the estimated structural model (discount factors and costs parameters) to predict the hospitals' decisions. This enables us to quantify the change in system statistics from adding one bed in ICU. While increasing ICU capacity naturally reduces ICU congestion, such a change requires substantial capital investments (e.g. up to $1.1 million per year estimated from the $4302 daily expense in Franzini et al. (2011)). In the second study, we keep the ICU capacity unchanged and consider the hypothetical situation where the hospital is made to incorporate longer-term impacts of their decisions when making their admission decisions. In particular, we quantify the impact on system statistics as the hospitals increase their discount factors from the current estimated levels to $\beta = 0.9$ without introducing any change in the medical resources they have, e.g., adding ICU beds. While it is not immediately clear whether it would be possible to alter the hospital's behavior in a manner which is consistent with a larger discount factor, this counterfactual experiment reveals the potential impact of such a change, which we can measure relative to the impact of adding an ICU bed. In both counterfactual studies, the statistics are computed from the average of 100 simulation runs. Each run has the same number of periods as in the data plus a three-month warm up period.

5.2.1. Adding One ICU Bed We start by studying the impact of adding one more bed in ICU. In particular, we focus on the impact on ICU congestion when keeping the arrival and service rates fixed. Due to the size of the ICUs in our study, the addition of one bed has very limited impact on the average ICU occupancy level. Thus, we focus on two statistics that capture "extremely" congested ICU states. The first is the probability of high ICU congestion, which is defined by

$$\Pr(\text{HighCgstn}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left\{n_t^I \geq B - 1\right\}; \tag{26}$$

i.e, the proportion of periods with only one or no empty bed(s) in the ICU. High ICU congestion is known to be associated with worse medical outcomes such as higher mortality, longer hospital length-of-stay, and higher risk from postoperative complications (Gattinoni et al. (2004), Hugonnet et al. (2007), and Gabler et al. (2013)); this may be due, for example, to reduced likelihood of ICU admission (Kim et al. (2015)) or increased likelihood of demand-driven discharges (Kc and Terwiesch (2012)).

We also examine the probability of the ICU being full. This impacts the likelihood of external arrivals balking upon arrival because there are no ICU beds available.

$$\Pr(\text{Balk}|\text{External Arrival}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left\{n_t^I = B\right\}. \tag{27}$$

We estimate the absolute and relative impact of adding one ICU bed on the probability of high ICU congestion and the probability of external arrivals balking. Recall that each period in our structural model corresponds to a two hour interval. We provide a conservative estimate for the number of ICU patients who spend their ICU stay under highly congested states by

$$\text{Pats HighCgstn} = \Pr(\text{HighCgstn}) \times 365 \times 12 \times \mu_I \times (B-1).$$

The first three components provide an estimate for the total number of high congestion periods in a year. We divide this by the average LOS of each patient, $1/\mu_I$ periods, to estimate the number of patients each bed can serve during the high congestion periods in a year. Since there is at most one bed available during the high congestion periods, putting everything together gives an estimate for the number of patients that are exposed to high congestion periods. Using a similar argument, the number of external arrivals who balk is estimated by

$$\text{Pats Balk} = \Pr(\text{Balk}|\text{External Arrival}) \times 365 \times 12 \times \lambda_E,$$

where λ_E is the external arrival rate in each period. The results are shown in Table 9 below.

From Table 9, we see adding one bed in the ICU indeed leads to substantial reductions in ICU congestion. The relative drops in high ICU congestion and balking probabilities are greater than 30% in most cases. While not reported here, the standard deviations computed from simulation confirm all the differences are statistically significant at 1% level. For some hospitals, such reduction translates to 20 fewer days and 50 fewer patients exposed to high ICU congestion, as well as 10 more external arrivals admitted in each year. As the patients who require ICU care are usually very severe, the effects observed above have the potential to lead to large improvements for their medical outcomes.

Next, we look into how the benefit of adding one bed in ICU varies across hospitals. Table 10 reports the correlations of the three measures (ΔDays HighCgstn, ΔPats HighCgstn, and ΔPats Balking in Table 9) with the average ICU occupancy levels and estimated discount factors across 22 hospitals. All three effect measures are positively correlated with both the ICU occupancy levels and discount factors. The results

Table 9 Counterfactual estimates of impact when adding one bed in ICU

Hosp	$\hat{\beta}$	Δ Pr(HighCgstn) (in % points)	Rel Δ Pr(HighCgstn) (in %)	Δ Pr(Balk) (in % points)	Rel Δ Pr(Balk) (in %)	ΔDays HighCgstn (in # days)	ΔPats HighCgstn (in # patients)	ΔPats Balk (in # patients)
1	0.3	1.72	0.35	0.61	0.39	6.27	52.26	6.73
2	0.5	1.86	0.25	0.72	0.27	6.78	52.78	8.43
3	0.1	1.66	0.54	0.48	0.56	6.06	23.82	2.11
4	0.4	1.16	0.31	0.38	0.31	4.22	47.08	8.63
5	0.1	5.13	0.38	1.87	0.39	18.74	68.04	8.84
6	0.2	0.75	0.42	0.20	0.37	2.73	23.77	2.42
7	0.9	6.36	0.35	2.57	0.38	23.23	84.04	21.20
8	0.1	2.99	0.38	1.14	0.42	10.90	64.88	7.92
9	0.5	2.36	0.32	0.88	0.35	8.60	78.61	13.57
10	0.1	2.25	0.53	0.76	0.60	8.23	51.80	7.56
11	0.9	7.39	0.49	2.70	0.54	26.98	56.65	6.61
12	0.6	1.66	0.30	0.59	0.31	6.06	43.17	7.67
13	0.1	0.64	0.51	0.20	0.57	2.34	11.65	0.95
14	0.3	0.65	0.24	0.24	0.28	2.36	25.90	5.26
15	0.1	0.23	0.60	0.06	0.61	0.85	4.77	0.30
16	0.1	0.69	0.56	0.20	0.56	2.52	12.07	0.95
17	0.9	5.90	0.40	2.28	0.45	21.54	50.58	6.60
18	0.4	0.18	0.44	0.05	0.42	0.65	8.24	1.16
19	0.1	0.02	0.41	0.01	0.56	0.08	0.75	0.09
20	0.3	0.26	0.63	0.05	0.52	0.94	3.91	0.15
21	0.7	4.55	0.40	1.97	0.47	16.61	59.02	11.97
22	0.3	2.26	0.39	0.95	0.46	8.27	42.31	7.59

Note: Counterfactual simulation result from adding one ICU bed: The third and fourth columns report the absolute and relative drops in high congestion probability in (26); the fifth and sixth columns report the absolute and relative drops in balking probability in (27). The last three columns report the equivalent numbers of days and patients affected in a year.

Table 10 Correlations between impact of one bed on ICU congestion and ICU occupancy or $\hat{\beta}$

	ΔDays HighCgstn	ΔPats HighCgstn	ΔPats Balk
ICUOccu	0.354	0.737^{**}	0.732^{**}
$\hat{\beta}$	0.700^{***}	0.520^{*}	0.602^{**}

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

suggest that the benefit from adding one ICU bed is more significant for ICUs with higher occupancy and/or with hospitals that are more longer-term looking.

We further explore this relationship with the following simple linear regression to examine how the average ICU occupancy level (measured in percentage) and the discount factor impact the benefit from adding an ICU bed:

$$\text{Effect}_i = \alpha + \gamma_I \times \text{ICUOccu}_i + \gamma_b \times \hat{\beta}_i + \varepsilon_i, \tag{28}$$

where the dependent variable Effect_i is set as ΔDays HighCgstn, ΔPats HighCgstn, and ΔPats Balking, respectively. The coefficients γ_I and γ_b are summarized in Table 11 below. The results can be interpreted as follows: An additional bed can bring greater benefit for ICUs with higher occupancy levels (this is the classic non-linear effect of load seen in queueing systems). Additionally, when the hospital has a larger discount factor, their admission decisions are more sensitive to the ICU state; therefore, they are more likely

to save beds for future patients when the ICU occupancy is already high. Such foresight enables them to use the additional bed more efficiently to manage ICU congestion.

Table 11 Coefficients in regression (28) for impact of one bed on ICU congestion with ICU occupancy and $\hat{\beta}$ as explanatory variables

	ΔDays HighCgstn	ΔPats HighCgstn	ΔPats Balk
ICUOccu	3.66	140.14^{**}	26.22^{**}
$\hat{\beta}$	19.14^{**}	21.48	6.31^{*}

Note: $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Each column corresponds to a separate regression in (28), with the dependent variable being reduction in days of highly congested state, drop in patients affected by highly congested state, or drop in external arrivals who balk, respectively.

5.2.2. Increasing the Discount Factor Adding an ICU bed can significantly reduce congestion, but also requires substantial financial investments. We now consider the potential impact of modifying the hospital's behavior. As discussed earlier, hospitals are typically not trained to manage their patients with a system-level view; our estimation results suggest they tend to be more near-term focused when making ICU admission decisions. That said, we found that some hospitals were more forward looking. Hence, we wish to estimate what might happen if the hospitals were provided with the right support and information to enable them to account more for the longer-term when making admissions decisions. While it might not be possible, in practice, to alter the discount factor of all hospitals, this experiment provides some insight into what might be possible if the discount factor could be nudged a bit higher as well as what could happen if the hospitals had more sophisticated strategies to deal with congestion.

The specific counterfactual which we consider is to keep the ICU capacity unchanged, and increase the discount factors from the current estimated levels to 0.9. Practically, this corresponds to the hospital caring about the decision costs and system states over the next one to two days ($0.9^{12} \approx 0.28$), rather than the next six hours, which was the average behavior implied by Table 5. Note that for some hospitals, the estimated $\hat{\beta}$ is already 0.9, so this counterfactual has no impact on the system dynamics. We study how increasing β impacts ICU congestion and ED waiting time for the two classes of patients. The results are summarized in Table 12.

We see increasing β *alone* can have substantial effect on reducing ICU congestion. For some hospitals, the relative drops in the probability of high ICU congestion and balking are more than 20% as the hospital shifts towards longer-term discounting. This translates to significant improvement in terms of the frequency of high ICU congestion as well balking of external arrivals. For example, when Hospital 2 increases the discount factor from 0.5 to 0.9 when making admission decisions, there would be on average six fewer days and 34 fewer patients under high ICU congestion, as well as eight more external arrivals admitted in

Table 12 Counterfactual estimates of impact when β increases from the estimated $\hat{\beta}$ to 0.9

Hosp	$\hat{\beta}$	Δ Pr(HighCgstn) (in % points)	Rel Δ Pr(HighCgstn) (in %)	Δ Pr(Balk) (in % points)	Rel Δ Pr(Balk) (in %)	ΔDays HighCgstn (in # days)	ΔPats HighCgstn (in # patients)	ΔPats Balk (in # patients)
1	0.3	1.06	0.22	0.42	0.27	3.85	32.12	4.67
2	0.5	1.59	0.21	0.65	0.25	5.79	45.04	7.59
3	0.1	0.38	0.12	0.11	0.13	1.38	5.41	0.48
4	0.4	0.34	0.09	0.11	0.09	1.25	13.98	2.45
5	0.1	1.70	0.13	0.71	0.15	6.21	22.56	3.38
6	0.2	0.41	0.23	0.14	0.26	1.49	12.98	1.69
7	0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.1	1.72	0.22	0.74	0.27	6.27	37.32	5.15
9	0.5	1.51	0.20	0.58	0.23	5.50	50.25	8.91
10	0.1	0.26	0.06	0.06	0.04	0.96	6.04	0.57
11	0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.6	0.62	0.11	0.28	0.15	2.25	16.00	3.70
13	0.1	0.16	0.12	0.02	0.05	0.57	2.82	0.08
14	0.3	0.43	0.16	0.14	0.16	1.56	17.11	3.02
15	0.1	0.06	0.15	0.01	0.07	0.21	1.18	0.03
16	0.1	0.26	0.21	0.08	0.23	0.94	4.48	0.39
17	0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	0.4	0.03	0.08	0.01	0.06	0.12	1.51	0.18
19	0.1	0.01	0.26	0.00	0.00	0.05	0.47	0.00
20	0.3	0.01	0.02	0.01	0.10	0.02	0.10	0.03
21	0.7	0.12	0.01	0.35	0.08	0.45	1.59	2.16
22	0.3	0.34	0.06	0.15	0.07	1.24	6.35	1.22

Note: Counterfactual simulation result from increasing discount factor from current estimated level ($\hat{\beta}$) to 0.9: The third and fourth columns report the absolute and relative drops in high congestion probability in (26); the fifth and sixth columns report the absolute and relative drops in balking probability in (27). The last three columns report the equivalent numbers of days and patients affected in a year.

each year. Comparing the corresponding results for Hospitals 1, 2, 8, and 9 in Tables 9 and 12, we notice that the impact of increasing β can be comparable to adding one ICU bed. This highlights the importance of understanding how the hospitals internalize the intertemporal externalities on ICU admission decisions, and reveals a potential approach for hospitals to reduce ICU congestion, i.e., providing their hospitals more tools and skills to manage and react to congestion.

While increasing β helps to reduce ICU congestion, it may also lead to longer ED waiting time as hospitals would reduce the ICU admission probability as the unit get more congested. This potentially undesirable effect can also be quantified by our structural model via simulation. In the table below, we show the increase in average ED waiting time for the two classes of patients as β increases to 0.9, denoted by ΔEDWait_l and ΔEDWait_h respectively. We also compute the difference in average ED waiting time between the two classes at the estimated $\hat{\beta}$ and $\beta = 0.9$.

In the third and fourth columns in Table 13, we see a larger β indeed increases the average ED waiting time for both classes of patients. For example, as β increases from 0.3 to 0.9 in Hospital 1, the average ED waiting time increases by two hours and 36 minutes (0.6 hour) for low and high severity classes, respectively. However, the magnitude of the impact is very different for the two classes. Comparing ΔEDWait_l and ΔEDWait_h, we see the increase in ED waiting time is much more significant for patients in the low severity class than for those in the high severity class. This can be interpreted as follows: As β increases,

Table 13 Relationship of ED waiting time with β (measured in hours)

Hosp	Estimated $\hat{\beta}$	$\text{EDWait}_l - \text{EDWait}_h$		Change β from $\hat{\beta} \to 0.9$	
		$\beta = \hat{\beta}$	$\beta = 0.9$	ΔEDWait_l	ΔEDWait_h
1	0.3	−0.09	1.33	2.05	0.63
2	0.5	−0.03	1.19	1.84	0.63
3	0.1	0.01	1.31	1.59	0.29
4	0.4	−0.10	0.81	1.33	0.42
5	0.1	−0.05	1.26	1.75	0.44
6	0.2	−0.08	1.11	1.70	0.50
7	0.9	−0.06	−0.06	0.00	0.00
8	0.1	−0.15	0.46	1.36	0.76
9	0.5	−0.04	1.02	1.55	0.49
10	0.1	−0.03	0.43	0.73	0.26
11	0.9	0.01	0.01	0.00	0.00
12	0.6	−0.33	0.10	1.15	0.71
13	0.1	−0.02	1.67	2.08	0.38
14	0.3	−0.09	0.71	1.26	0.46
15	0.1	−0.01	0.24	0.37	0.11
16	0.1	−0.03	1.02	1.48	0.44
17	0.9	−0.17	−0.17	0.00	0.00
18	0.4	0.08	1.22	1.42	0.28
19	0.1	−0.02	0.26	0.45	0.16
20	0.3	0.06	1.63	1.90	0.33
21	0.7	−0.13	0.71	1.33	0.49
22	0.3	0.10	2.86	3.23	0.47

Note: This table reports the impact on ED waiting time from increasing discount factor from current estimated level to 0.9: The third and fourth columns report the difference in ED waiting time between the two classes ($\text{EDWait}_l - \text{EDWait}_h$) at the estimated $\hat{\beta}$ and $\beta = 0.9$; the last two columns (ΔEDWait_l and ΔEDWait_h) report the increase in ED waiting time for the two classes of patients when β increases to 0.9.

hospitals tend to reduce the probability of ICU admission which increases waiting times. When they do so, they generally prefer to delay the admissions of patients from the low severity class rather than those from high severity class, as the latter on average have higher waiting costs and are more likely to eventually be admitted to the ICU.

Next, we look into the difference in ED waiting times of the two classes, EDWait_l - EDWait_h, evaluated at the estimated $\hat{\beta}$ and $\beta = 0.9$. The average ED waiting times for the two classes are very close at the estimated $\hat{\beta}$. As we have shown in Section 5.1.2, the average waiting time produced by our structural model approximates the observed data very well. Indeed, the observed average ED waiting times for the two classes is 1.27 hours (low severity) and 1.42 hours (high severity) respectively. However, when β increases to 0.9, the average ED waiting time for the low severity class becomes significantly longer than that for the high severity class in most hospitals. As the hospital accounts more for the longer-term (i.e. the discount factor increases), they tend to actively differentiate across the two classes by admitting high severity patients more

quickly and having low severity patients wait longer. Such behavior may partially offset the negative impact of longer ED waiting time, by disproportionately impacting the less severe patients.

We also find there is little impact of β on the overall admission probability, i.e., the proportion of patients eventually admitted to ICU. This is because the overall admission probability primarily depends on the rerouting costs and system workload. The rerouting costs measure the average relative effect of rerouting to the ward compared with admitting to the ICU, and the system workload determines the long-run bed availability in the ICU. As both are unaffected by changes in β, the overall admission probability also is practically unchanged as β varies. Thus, the primary impact of increasing β is to "delay" admissions to periods when the ICU is less congested. This effectively reduces the likelihood of high ICU congestion states but increases the ED waiting time. In a way, increasing β smooths the load on the ICU.

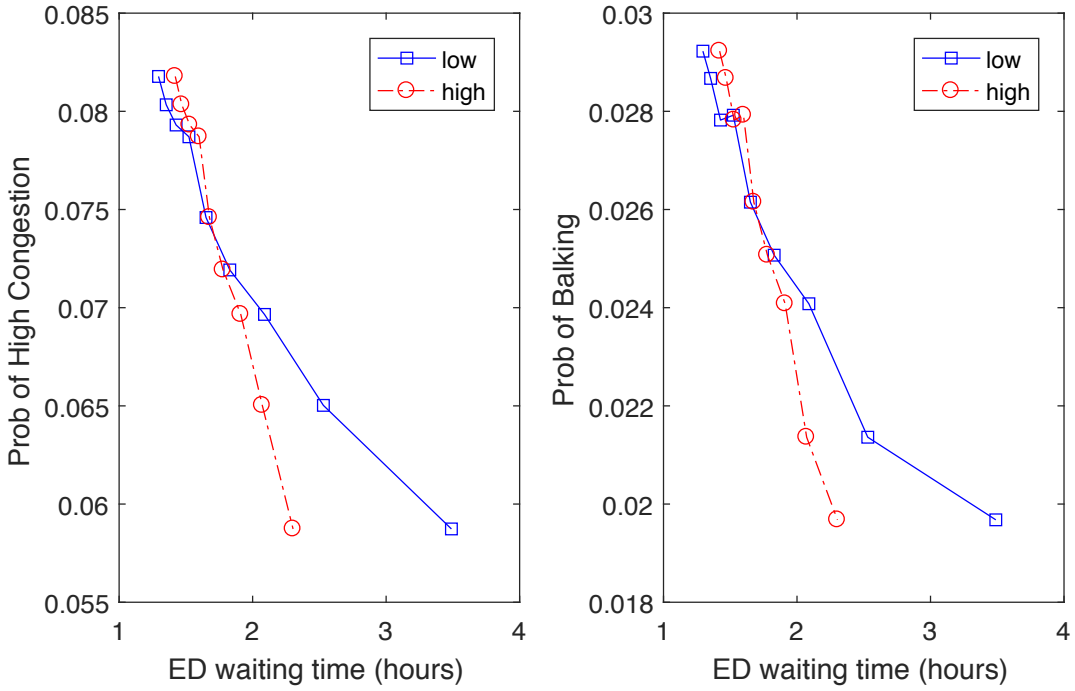
We use Hospital 2 as an example to further illustrate the trade-off between reduction in ICU congestion states and longer ED waiting time. In Figure 6, we plot the probabilities of high ICU congestion and balking versus the average ED waiting time for the two patient classes, for each $\beta \in \{0.1, 0.2, \ldots, 0.9\}$. The points on each line from left to right represent the results with $\beta = 0.1, 0.2, \ldots, 0.9$ respectively. The downward patterns again illustrate the trade-off of increasing β: Larger β decreases the probabilities of high ICU congestion and balking, but increases ED waiting time for both classes. Moreover, the increase in ED waiting time is more significant for the low severity class, and the two classes are differentiated by their average waiting time when β is large.

6. Conclusions

Clinical practice aims to provide the best care possible for each individual patient. That said, it has also been well documented that ICU admission behaviors are impacted by ICU congestion. In this work, we aim to understand if some of this practice variation could be explained by perceived discounting behavior. To the best of our knowledge, our work is the first to study how hospitals internalize intertemporal externalities – i.e., the admission decision for a current patient impacts ICU congestion and possible other, future patients – when making the dynamic ICU admission decision. While, on average, the hospitals appear to be more focused on the near-term when making ICU admission decisions, we find that there is *large heterogeneity* in the degree of forward-looking behavior across hospitals. This suggests that some of the observed practice variation in how hospitals alter admission behaviors in response to ICU congestion may be partially explained by the fact that hospitals appear to internalize the intertemporal externalities very differently.

We use counterfactual simulations to show that if it were possible to increase the hospital's degree of forward-looking behavior *alone*, this can have a substantial impact on ICU congestion and patient flow. For some hospitals, the effect of this change is comparable to increasing the ICU capacity, which may be very financially and space expensive. That said, it may not be straightforward whether providing more information about the impact of ICU congestion and how it can influence patient flow would in fact result in

Figure 6 **Counterfactual statistics for Hospital 2 with $\beta = 0.1, 0.2, \ldots, 0.9$ (from left to right)**



Note: The left (resp. right) panel shows the high-congestion probability (resp. balking probability) versus ED waiting time for the low (blue solid line) and high (red dotted line) severity patients at Hospital 2 with $\beta = 0.1, 0.2, \ldots, 0.9$ (from left to right).

a change in hospital behavior as there may be system-level factors (such as float staff availability) which cannot be readily altered with these interventions. Additionally, we do not claim that the hospital is solving the proposed dynamic optimization problem, rather we posit that it serves as a reasonable way to approximate their decision making process. Despite these practical challenges in altering hospital behavior, our work suggests that how hospitals respond to staffing and capacity shortages could benefit from a better understanding of their forward-thinking behavior. This would also help the hospital better understand the potential impact different strategies – e.g. better forecasting capabilities of demand and/or patient severity – could have on patient flow.

Our study has several limitations that may shed light on future research directions. First, the data we use have no direct information on the ICU admission decision process – we can only observe the resulting outcome of where patients are admitted. Thus, while we posit a structural model to capture various features which influence where patients are admitted and when, there are many factors which are not explicitly included in the model nor do we have data to understand their impact on the decisions. Second, the data is also limited to hospitals within one healthcare system. It is possible that different payment models may impact the manner in which hospitals internalize intertemporal externalities. For instance, it is possible that under a purely Fee-For-Service system (rather than the capitated system of KPNC), hospitals may be even more focused on the near-term. If this is the case, our counterfactual results suggest that other hospital

systems may benefit even more from shifting to be more longer-term focused. Finally, all the ICUs in our study cohort are closed, so that the attending intensivist has final say on which patients are admitted. It is not clear how the hospital would internalize the intertemporal externalities in an open ICU.

In order to focus on the identification of the intertemporal externalities – the discount factor – we propose a parsimonious structural model. As ICU patient flow and admission decisions are quite complex, it would be impossible to capture all features and still have a tractable model and so we made a number of simplifications in our model. For example, we assume constant patient arrival rates for ED patients, a homogeneous constant and state-independent departure rate for all ICU patients, and we do not account for possible ICU readmissions. As analysis of queueing systems with such features is an active area of research which often utilizes approximation approaches (e.g. fluid and/or infinite server models) and/or asymptotic regimes, it would be interesting to see whether it would be possible to incorporate such features into the structural model. This might require assuming that the hospital must heuristically solve the dynamic optimization problem. Finally, this study focuses on one aspect of the hospital's behavior in ICU admission decisions – the degree of internalization of the intertemporal externalities. However, there is evidence that there are other adaptive behaviors – such as demand driven discharges (Kc and Terwiesch 2012) – the hospital may utilize when managing ICU beds. Such behavior may also affect or be affected by the discount factor of the hospital. Thus, a potential future direction of research is to study the joint impact of these various behavioral aspects of the hospital's decision-making on ICU capacity management.

Acknowledgments

We thank Marla Gardner and John Greene 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.

References

Abbring, Jaap H, Øystein Daljord. 2019. Identifying the discount factor in dynamic discrete choice models. *Becker Friedman Institute for Research in Economics Working Paper* (2017-17).

Akşin, Zeynep, Barış Ata, Seyed Morteza Emadi, Che-Lin Su. 2013. Structural estimation of callers' delay sensitivity in call centers. *Management Science* **59**(12) 2727–2746.

Akşin, Zeynep, Baris Ata, Seyed Morteza Emadi, Che-Lin Su. 2016. Impact of delay announcements in call centers: An empirical approach. *Operations Research* **65**(1) 242–265.

Allon, Gad, Sarang Deo, Wuqin Lin. 2013. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations research* **61**(3) 544–562.

Bajari, Patrick, C Lanier Benkard, Jonathan Levin. 2007. Estimating dynamic models of imperfect competition. *Econometrica* **75**(5) 1331–1370.

Batt, Robert J, Diwas S Kc, Bradley R Staats, Brian W Patterson. 2019. The effects of discrete work shifts on a nonterminating service system. *Production and Operations Management* **28**(6) 1528–1544.

Batt, Robert J, Christian Terwiesch. 2016. Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science* **63**(11) 3531–3551.

Bray, Robert L, Yuliang Yao, Yongrui Duan, Jiazhen Huo. 2019. Ration gaming and the bullwhip effect. *Operations Research* **67**(2) 453–467.

Chan, Carri W, Vivek F Farias, Gabriel J Escobar. 2016. The impact of delays on service times in the intensive care unit. *Management Science* **63**(7) 2049–2072.

Chen, Lena M, Marta Render, Anne Sales, Edward H Kennedy, Wyndy Wiitala, Timothy P Hofer. 2012. Intensive care unit admitting patterns in the veterans affairs health care system. *Archives of internal medicine* **172**(16) 1220–1226.

Ching, Andrew T, Matthew Osborne. 2017. Identification and estimation of forward-looking behavior: The case of consumer stockpiling. *Rotman School of Management Working Paper* (2594032).

Coopersmith, Craig M, Hannah Wunsch, Mitchell P Fink, Walter T Linde-Zwirble, Keith M Olsen, Marilyn S Sommers, Kanwaljeet JS Anand, Kathryn M Tchorz, Derek C Angus, Clifford S Deutschman. 2012. A comparison of critical care research funding and the financial burden of critical illness in the united states. *Critical care medicine* **40**(4) 1072–1079.

Corallo, Ashley N, Ruth Croxford, David C Goodman, Elisabeth L Bryan, Divya Srivastava, Therese A Stukel. 2014. A systematic review of medical practice variation in oecd countries. *Health Policy* **114**(1) 5–14.

Dong, Jing, Elad Yom-Tov, Galit B Yom-Tov. 2018. The impact of delay announcements on hospital network coordination and waiting times. *Management Science* **65**(5) 1969–1994.

Dubé, Jean-Pierre, Günter J Hitsch, Pranav Jindal. 2014. The joint identification of utility and discount functions from stated choice data: An application to durable goods adoption. *Quantitative Marketing and Economics* **12**(4) 331–377.

Edbrooke, David L, Cosetta Minelli, Gary H Mills, Gaetano Iapichino, Angelo Pezzi, Davide Corbella, Philip Jacobs, Anne Lippert, Joergen Wiis, Antonio Pesenti, et al. 2011. Implications of icu triage decisions on patient mortality: a cost-effectiveness analysis. *Critical Care* **15**(1) R56.

Emadi, Seyed Morteza, Bradley R Staats. 2019. A structural estimation approach to agent attrition. *Management Science, to appear* .

Escobar, Gabriel J, Marla N Gardner, John D Greene, David Draper, Patricia Kipnis. 2013. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. *Medical care* 446–453.

Escobar, Gabriel J, Juan Carlos LaGuardia, Benjamin J Turk, Arona Ragins, Patricia Kipnis, David Draper. 2012. Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record. *Journal of hospital medicine* **7**(5) 388–395.

Fisher, Elliott S, David E Wennberg, Thérèse A Stukel, Daniel J Gottlieb. 2004. Variations in the longitudinal efficiency of academic medical centers: Increased intensity of care does not appear to be associated with higher quality or to result in better survival at amcs. *Health Affairs* **23**(Suppl2) VAR–19.

Franzini, Luisa, Kavita R Sail, Eric J Thomas, Laura Wueste. 2011. Costs and cost-effectiveness of a telemedicine intensive care unit program in 6 intensive care units in a large health care system. *Journal of critical care* **26**(3) 329–e1.

Frederick, Shane, George Loewenstein, Ted O'donoghue. 2002. Time discounting and time preference: A critical review. *Journal of economic literature* **40**(2) 351–401.

Gabler, Nicole B, Sarah J Ratcliffe, Jason Wagner, David A Asch, Gordon D Rubenfeld, Derek C Angus, Scott D Halpern. 2013. Mortality among patients admitted to strained intensive care units. *American journal of respiratory and critical care medicine* **188**(7) 800–806.

Gattinoni, Luciano, Danilo Radrizzani, Bruno Simini, Guido Bertolini, Luca Ferla, Giovanni Mistraletti, Francesca Porta, Dinis R Miranda, et al. 2004. Volume of activity and occupancy rate in intensive care units. association with mortality. *Intensive care medicine* **30**(2) 290–297.

Green, Linda V. 2002. How many hospital beds? *INQUIRY: The Journal of Health Care Organization, Provision, and Financing* **39**(4) 400–412.

Green, Linda V, Sergei Savin, Nicos Savva. 2013. "nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science* **59**(10) 2237–2256.

Halfin, Shlomo, Ward Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations research* **29**(3) 567–588.

Halpern, Neil A, Stephen M Pastores. 2010. Critical care medicine in the united states 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Critical care medicine* **38**(1) 65–71.

Halpern, Neil A, Stephen M Pastores. 2015. Critical care medicine beds, use, occupancy and costs in the united states: a methodological review. *Critical care medicine* **43**(11) 2452.

Hotz, V Joseph, Robert A Miller. 1993. Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies* **60**(3) 497–529.

Hu, Wenqi. 2018. Managing hospital care: Data-driven decisions and comparisons. Ph.D. thesis, Columbia University.

Hugonnet, Stéphane, Jean-Claude Chevrolet, Didier Pittet. 2007. The effect of workload on infection risk in critically ill patients. *Critical care medicine* **35**(1) 76–81.

Ibanez, Maria R, Jonathan R Clark, Robert S Huckman, Bradley R Staats. 2017. Discretionary task ordering: Queue management in radiological services. *Management Science* **64**(9) 4389–4407.

Kahn, Jeremy M, Gordon D Rubenfeld, Jeffery Rohrbach, Barry D Fuchs. 2008. Cost savings attributable to reductions in intensive care unit length of stay for mechanically ventilated patients. *Medical care* 1226–1233.

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

Kim, Song-Hee, Carri W Chan, Marcelo Olivares, Gabriel Escobar. 2015. Icu admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science* **61**(1) 19–38.

Kim, Song-Hee, Jordan Tong, Carol Peden. 2019. Admission control biases in hospital unit capacity management: How occupancy information hurdles and decision noise impact utilization. *Available at SSRN 3219451* .

Komarova, Tatiana, Fabio Sanches, Daniel Silva Junior, Sorawoot Srisuma. 2018. Joint analysis of the discount factor and payoff parameters in dynamic discrete choice models. *Quantitative Economics* **9**(3) 1153–1194.

Li, Jun, Nelson Granados, Serguei Netessine. 2014. Are consumers strategic? structural estimation from the air-travel industry. *Management Science* **60**(9) 2114–2137.

Lu, Yina, Andrés Musalem, Marcelo Olivares, Ariel Schilkrut. 2013. Measuring the effect of queues on customer purchases. *Management Science* **59**(8) 1743–1763.

Magnac, Thierry, David Thesmar. 2002. Identifying dynamic discrete decision processes. *Econometrica* **70**(2) 801–816.

Manski, Charles F. 1993. Identification of endogenous social effects: The reflection problem. *The review of economic studies* **60**(3) 531–542.

Mehta, Nitin, Jian Ni, Kannan Srinivasan, Baohong Sun. 2017. A dynamic model of health insurance choices and healthcare consumption decisions. *Marketing Science* **36**(3) 338–360.

Miessi Sanches, Fabio A, Daniel Junior Silva, Sorawoot Srisuma. 2016. Ordinary least squares estimation of a dynamic game model. *International Economic Review* **57**(2) 623–634.

Mullan, Fitzhugh. 2004. Wrestling with variation: An interview with jack wennberg: The creator of modern-day evaluative clinical sciences discusses what motivated him to define and pursue this area of study. *Health Affairs* **23**(Suppl2) VAR–73.

O'Connor, Annette M, Hilary A Llewellyn-Thomas, Ann Barry Flood. 2004. Modifying unwarranted variations in health care: Shared decision making using patient decision aids: A review of the evidence base for shared decision making. *Health Affairs* **23**(Suppl2) VAR–63.

Olivares, Marcelo, Christian Terwiesch, Lydia Cassorla. 2008. Structural estimation of the newsvendor model: an application to reserving operating room time. *Management Science* **54**(1) 41–55.

Pesendorfer, Martin, Philipp Schmidt-Dengler. 2008. Asymptotic least squares estimators for dynamic games. *The Review of Economic Studies* **75**(3) 901–928.

Rath, Sandeep, Kumar Rajaram. 2018. Staff planning for hospitals with cost estimation and optimization. *Kenan Institute of Private Enterprise Research Paper* (18-28).

Roberts, Rebecca R, Paul W Frutos, Ginevra G Ciavarella, Leon M Gussow, Edward K Mensah, Linda M Kampe, Helen E Straus, Gnanaraj Joseph, Robert J Rydman. 1999. Distribution of variable vs fixed costs of hospital care. *Jama* **281**(7) 644–649.

Rust, John. 1987. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society* 999–1033.

Rust, John. 1994. Structural estimation of markov decision processes. *Handbook of econometrics* **4** 3081–3143.

Shmueli, Amir, Charles L Sprung, Edward H Kaplan. 2003. Optimizing admissions to an intensive care unit. *Health Care Management Science* **6**(3) 131–136.

Singer, Adam J, Henry C Thode Jr, Peter Viccellio, Jesse M Pines. 2011. The association between length of emergency department boarding and mortality. *Academic Emergency Medicine* **18**(12) 1324–1329.

Song, Hummy, Anita Tucker, Ryan Graue, Sarah Moravick, Julius Yang. 2019. Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science, to appear* .

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.

Wang, Yang. 2014. Dynamic implications of subjective expectations: Evidence from adult smokers. *American Economic Journal: Applied Economics* **6**(1) 1–37.

Westert, Gert P, Stef Groenewoud, John E Wennberg, Catherine Gerard, Phil DaSilva, Femke Atsma, David C Goodman. 2018. Medical practice variation: public reporting a first necessary step to spark change. *International Journal for Quality in Health Care* **30**(9) 731–735.

Yu, Qiuping, Gad Allon, Achal Bassamboo. 2016. How do delay announcements shape customer behavior? an empirical study. *Management Science* **63**(1) 1–20.

Appendix A: Formulae and Supplementary Tables

A.1. Explicit expressions for state transition probability $g(s'|s)$

In this section, we provide explicit expressions for the function $g(s'|s)$ used in Proposition 1, which is the transition probability from the intermediate state s (after action is taken) to the state s' at the start of the next period. Denote $s = (\tilde{n}^E_{l,t}, \tilde{n}^E_{h,t}, \tilde{n}^I_t)$ and $s' = (n^E_{l,t+1}, n^E_{h,t+1}, n^I_{t+1})$, which are the state after action is taken in period t and the state at the beginning of period $t+1$ (before action) respectively. By (8), the transition from s to s' includes the ED arrivals from the two classes of patients, external arrivals to the ICU, and departures from the ICU. As the ED arrivals are independent of the ICU external arrivals and departures, we have

$$g(s'|s) = g_{Q,l}\left(n^E_{l,t+1}|\tilde{n}^E_{l,t}\right) g_{Q,h}\left(n^E_{h,t+1}|\tilde{n}^E_{h,t}\right) g_I\left(n^I_{t+1}|\tilde{n}^I_t\right), \tag{29}$$

where $g_{Q,l}$ and $g_{Q,h}$ denote the transition probabilities for the numbers of ED patients from low and high classes respectively, and g_I denotes the transition probability for the number of ICU patients.

For the ED transition probabilities $g_{Q,l}$ and $g_{Q,h}$, we only need to consider the new arrivals for class $i \in \{l, h\}$, which follow truncated Poisson distributions with rate $\lambda_{Q,i}$ and truncation by M_{A_i} from above. Additionally accounting for the ED capacity constraint, the number of ED arrivals is capped by $\max\{M_{A_i}, Q_i - \tilde{n}^E_{i,t}\}$, i.e., the bigger of the maximum arrival per period and the remaining ED capacity. Thus, the transition probability can be computed as

$$g_{Q,i}(m|n) = \begin{cases} (\lambda_{Q,i})^{m-n} \exp(-\lambda_{Q,i})/(m-n)! & \text{if } n \le m < n + \max\{M_{A_i}, Q_i - n\} \\ \sum_{j=\bar{A}_i}^{+\infty} (\lambda_{Q,i})^j \exp(-\lambda_{Q,i})/j! & \text{if } m = n + \max\{M_{A_i}, Q_i - n\} \end{cases} \tag{30}$$

and $g_{Q,i}(m|n) = 0$ elsewhere. The second line in (30) considers the case where the upper bound of the number of arrivals, $\max\{M_{A_i}, Q_i - \tilde{n}^E_{i,t}\}$, is reached.

For the ICU transition probability g_I, we need to consider both external arrivals and departures. The number of external arrivals E_t follows a Poisson distribution with rate λ_E, and is capped by the remaining ICU capacity $B - \tilde{n}^I_t$. With E_t external arrivals in the period, the ICU would have total $\tilde{n}^I_t + E_t$ patients. Then, the number of departures, D_t, follows a Binomial-$(\tilde{n}^I_t + E_t, \mu_I)$ distribution. We have following relationship:

$$n^I_{t+1} + D_t = \tilde{n}^I_t + E_t.$$

Thus, given the number of external arrivals, E_t, the number of departures follows by $D_t = \tilde{n}^I_t + E_t - n^I_{t+1}$. We note that E_t can be greater than $\max\{n^I_{t+1} - \tilde{n}^I_t, 0\}$. Summing up the probability of all possible choices of the Poisson-distributed E_t (and the Binomial-distributed D_t accordingly), we can derive transition probability $g_I(m|n)$ as

$$\begin{aligned} g_I(m|n) = & \sum_{j=\max\{m-n,0\}}^{B-n-1} \lambda_I^j \frac{\exp(-\lambda_I)}{j!} \frac{(n+j)!}{(n+j-m)!m!} \mu_I^{n+j-m} (1-\mu_I)^m \\ & + \left(\sum_{j=B-n}^{+\infty} \lambda_I^k \frac{\exp(-\lambda_I)}{j!} \right) \frac{B!}{(B-m)!m!} \mu_I^{B-m} (1-\mu_I)^m, \text{ for } 0 \le m \le B, \end{aligned} \tag{31}$$

where the second line considers the case that E_t is truncated by the remaining capacity $B - n$. Under this case, the ICU reaches the full capacity and D_t is a Binomial-(B, μ_I) variable. Combining (30) and (31), we obtain the explicit expression for state transition probability $g(s'|s)$ by (29).

A.2. Model Assumptions for Identification

This section documents the assumptions made in Komarova et al. (2018), which are also satisfied by our model. In their paper, x and a denote the system state and action, respectively; ε represents the random perturbation in utility.

Assumption 2 *(i) (Additive Separability) For all a, x, ε, the per-period utility follows:*

$$u(a,x,\varepsilon) = \pi(a,x) + \varepsilon(a).$$

(ii) (Conditional Independence) The transition distribution of the states has the following factorization for all x', ε', x, ε, a:

$$P(x',\varepsilon'|x,\varepsilon,a) = Q(\varepsilon')G(x'|x,a),$$

where Q is the cumulative distribution function of ε and G denotes the transition law of x_{t+1} conditioning on x_t and a_t. Furthermore, ε_t has finite first moments, and a positive, continuous, and bounded density.

(iii) (Finite Observed State) $X = \{1,\dots,K\}$.

Assumption 3 (Linear-in-Parameter) *For all a, x:*

$$\pi(a,x;\theta) = \pi_0(a,x) + \theta^\top \pi_1(a,x),$$

where π_0 is a known real value function, π_1 is a known p-dimensional vector value function and θ is the p-dimensional unknown parameter.

In our setting, we have per-period cost given by (3) as

$$c(s_t,d_t) = c_{r,l}r_{l,t} + c_{w,l}\left(n^E_{l,t} - a_{l,t} - r_{l,t}\right) + c_{r,h}r_{h,t} + c_{w,h}\left(n^E_{h,t} - a_{h,t} - r_{h,t}\right).$$

The deterministic part of the per-period utility follows by $u(s_t,d_t) = -c(s_t,d_t)$. Thus, it is indeed linear in parameters $\theta = \{c_{r,l}, c_{w,l}, c_{r,h}, c_{w,h}\}$. In our model, the functions π_0 and π_1 specify to

$$\pi_0(d_t,s_t) \equiv 0,$$

and

$$\pi_1(d_t,s_t) = -\begin{bmatrix} r_{l,t} \\ n^E_{l,t} - a_{l,t} - r_{l,t} \\ r_{h,t} \\ n^E_{h,t} - a_{h,t} - r_{h,t} \end{bmatrix}.$$

where $s_t = \left(n^E_{l,t}, n^E_{h,t}, n^I_t\right)$ and $d_t = \{a_{l,t}, r_{l,t}, a_{h,t}, r_{h,t}\}$.

A.3. Formulae related to reduced form regressions in Section 3

The logit regression with only admission and rerouting decisions is established as follows:

$$\begin{aligned} \ln\left[\frac{\Pr(\text{Adm}_{it}|\mathbf{X}_i,\mathbf{S}_t)}{\Pr(\text{Rer}_{it}|\mathbf{X}_i,\mathbf{S}_t)}\right] = {} & \gamma_0 + \gamma_G \text{Gender}_i + \gamma_A \text{Age}_i + \gamma_L \text{LAPS2}_i + \gamma_{CP}\text{COPS2}_i + \gamma_{CH}\text{CHMR}_i \\ & + \gamma_H \text{Hosp}_i + \gamma_{ICU}\text{ICUOccu}_t + \gamma_{ED}\text{EDNum}_t + \gamma_{Dep}\text{DepPre}_t \\ & + \gamma_{AvgL}\text{AvgLAPS2}_t + \gamma_{WD}\text{WeekDay}_t + \gamma_{DT}\text{DayTime}_t + \epsilon_{it}, \end{aligned} \tag{32}$$

where $\Pr(\text{Adm}_{it}|\mathbf{X}_i, \mathbf{S}_t)$ and $\Pr(\text{Rer}_{it}|\mathbf{X}_i, \mathbf{S}_t)$ denote the probabilities of admitting and rerouting the patients, respectively. The keeping wait decisions are dropped in the logit regression, i.e., we only consider whether the patient is rerouted or admitted eventually. The estimated coefficients are reported in Tables 18 and 19.

For the multinomial logit model (1), its McFadden's pseudo R-squared is defined as:

$$\text{Pseudo } R^2 = 1 - \frac{\ln l^{model}}{\ln l^{null}},$$

where l^{model} is the likelihood from the estimated model, and l^{null} is the likelihood from the "null" model that only includes the intercept and categorical variable for each hospital, i.e.,

$$\ln\left[\frac{\Pr(d_{it}|\mathbf{X}_i, \mathbf{S}_t)}{\Pr(\text{Rer}_{it}|\mathbf{X}_i, \mathbf{S}_t)}\right] = \gamma_{0,d} + \gamma_{H,d}\text{Hosp}_i + \epsilon_{it},$$

where $d_{it} \in \{Wait_{it}, Adm_{it}\}$; Hosp_i is the hospital categorical variable.

A.4. Supplementary Tables

Table 14 Sample Size: Numbers of observed days and hospitalizations for each hospital

Hosp	Num. of days	Num. of hospitalizations
1	667	11,676
2	500	9,902
3	667	8,039
4	667	14,595
5	576	5,082
6	667	10,577
7	578	4,915
8	653	8,400
9	514	12,355
10	609	5,978
11	667	2,655
12	388	6,751
13	667	8,061
14	667	12,841
15	575	7,208
16	667	7,190
17	548	3,511
18	667	7,702
19	547	7,476
20	666	5,096
21	333	2,109
22	333	2,048

Table 15 Estimation results for other coefficients in multinomial logit regression in (1) for all hospitals

		Waiting	*Admission*			*Waiting*	*Admission*
	γ_0	-3.574^{***}	-1.614^{***}		γ_{Hosp16}	-1.238^{***}	-0.667^{***}
		(0.064)	(0.065)			(0.055)	(0.052)
	γ_{Hosp2}	0.412^{***}	0.126^{**}		γ_{Hosp17}	-0.411^{***}	-0.489^{***}
		(0.033)	(0.041)			(0.057)	(0.063)
	γ_{Hosp3}	-1.518^{***}	-0.891^{***}		γ_{Hosp18}	-0.787^{***}	-0.699^{***}
		(0.058)	(0.053)			(0.052)	(0.054)
	γ_{Hosp4}	-0.725^{***}	-0.816^{***}		γ_{Hosp19}	-1.564^{***}	-0.226^{***}
		(0.036)	(0.041)			(0.063)	(0.045)
	γ_{Hosp5}	-0.619^{***}	-0.021		γ_{Hosp20}	-0.506^{***}	-0.503^{***}
		(0.054)	(0.051)			(0.050)	(0.057)
	γ_{Hosp6}	-0.613^{***}	-0.419^{***}		γ_{Hosp21}	0.828^{***}	-0.212^{***}
		(0.040)	(0.041)			(0.048)	(0.071)
	γ_{Hosp7}	-0.864^{***}	-1.073^{***}		γ_{Hosp22}	0.237^{***}	0.138^{*}
		(0.056)	(0.067)			(0.058)	(0.065)
	γ_{Hosp8}	-0.359^{***}	0.092^{*}	Gender$_i$	γ_G	0.005	-0.163^{***}
		(0.041)	(0.042)			(0.015)	(0.017)
	γ_{Hosp9}	-0.172^{***}	-0.318^{***}	Age$_i$	γ_A	-0.007^{***}	-0.027^{***}
		(0.034)	(0.041)			(0.000)	(0.000)
	γ_{Hosp10}	-1.741^{***}	-0.463^{***}	COPS2$_i$	γ_{CP}	-0.001^{***}	-0.003^{***}
		(0.077)	(0.057)			(0.000)	(0.000)
	γ_{Hosp11}	0.362^{***}	0.078	CHMR$_i$	γ_{CH}	0.143	-0.007
		(0.051)	(0.063)			(0.131)	(0.111)
	γ_{Hosp12}	0.248^{***}	-0.580^{***}	DepPre$_t$	γ_{Dep}	-0.200^{***}	0.040^{***}
		(0.037)	(0.052)			(0.011)	(0.011)
	γ_{Hosp13}	-1.015^{***}	-0.571^{***}	AvgLAPS2$_t$	γ_{AvgL}	0.001	-0.000
		(0.049)	(0.048)			(0.000)	(0.000)
	γ_{Hosp14}	-0.507^{***}	-0.512^{***}	WeekDay$_t$	γ_{WD}	0.351^{***}	-0.025
		(0.037)	(0.041)			(0.019)	(0.018)
	γ_{Hosp15}	-2.234^{***}	-0.528^{***}	DayTime$_t$	γ_{DT}	0.689^{***}	-0.097^{***}
		(0.085)	(0.051)			(0.016)	(0.017)

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. The other coefficients, not reported in Table 2, (intercept, hospital fixed effects, gender, age, severity scores (COPS2 and CHMR), number of departures in the previous 2 hours, the average LAPS2 score in the ICU, weekday indicator, and day time indicator) from regression (1) for waiting and admission decisions, respectively.

Table 16 Estimation results for other coefficients in multinomial regression (1) by hospital: decision = *Waiting*

Hosp	Intercept γ_0	Gender γ_G	Age γ_A	COSP2 γ_{CP}	CHMR γ_{CH}	DepPre γ_{Dep}	AvgLAPS2 γ_{AvgL}	WeekDay γ_{WD}	DayTime γ_{DT}
1	-3.657^{***}	0.023	-0.005^{***}	-0.001	1.169^{**}	-0.502^{***}	0.004^{*}	0.480^{***}	0.795^{***}
	(0.201)	(0.048)	(0.001)	(0.001)	(0.420)	(0.034)	(0.002)	(0.060)	(0.050)
2	-2.402^{***}	0.128^{**}	-0.005^{***}	0.001	-0.872^{*}	-0.261^{***}	-0.003	0.201^{***}	0.819^{***}
	(0.216)	(0.044)	(0.001)	(0.001)	(0.402)	(0.029)	(0.002)	(0.053)	(0.045)
3	-5.367^{***}	-0.152	-0.005	0.001	0.668	0.071	0.002	0.597^{***}	0.879^{***}
	(0.381)	(0.105)	(0.003)	(0.001)	(0.916)	(0.095)	(0.002)	(0.132)	(0.111)
4	-6.604^{***}	0.001	-0.006^{***}	-0.001	-0.933	-0.137^{***}	0.006^{*}	0.771^{***}	1.125^{***}
	(0.296)	(0.056)	(0.002)	(0.001)	(0.487)	(0.028)	(0.003)	(0.090)	(0.061)
5	-4.489^{***}	0.058	-0.003	-0.003^{*}	0.553	-0.152	0.002	0.330^{**}	-0.005
	(0.353)	(0.097)	(0.003)	(0.001)	(0.805)	(0.102)	(0.002)	(0.117)	(0.099)
6	-3.731^{***}	0.044	-0.008^{***}	-0.001	-0.851	-0.102^{*}	-0.010^{***}	-0.044	0.617^{***}
	(0.276)	(0.064)	(0.002)	(0.001)	(0.594)	(0.044)	(0.002)	(0.071)	(0.066)
7	-4.036^{***}	-0.227^{*}	-0.010^{**}	-0.001	-0.555	0.096	0.003	0.460^{***}	0.704^{***}
	(0.364)	(0.101)	(0.003)	(0.001)	(0.874)	(0.090)	(0.003)	(0.122)	(0.103)
8	-4.687^{***}	0.046	-0.005^{*}	0.001	0.379	-0.071	-0.003	0.234^{**}	0.647^{***}
	(0.289)	(0.068)	(0.002)	(0.001)	(0.509)	(0.053)	(0.002)	(0.082)	(0.068)
9	-4.612^{***}	0.016	-0.006^{***}	-0.001	0.258	-0.308^{***}	0.008^{***}	0.620^{***}	0.344^{***}
	(0.204)	(0.047)	(0.001)	(0.001)	(0.433)	(0.031)	(0.002)	(0.063)	(0.049)
10	-6.086^{***}	0.113	-0.006	-0.003	0.787	-0.174	0.008^{*}	0.612^{**}	0.617^{***}
	(0.570)	(0.149)	(0.005)	(0.002)	(1.035)	(0.130)	(0.004)	(0.191)	(0.152)
11	-1.988^{***}	0.146	-0.005	-0.004^{***}	-0.153	-0.137	-0.004^{*}	0.221^{*}	0.171
	(0.287)	(0.089)	(0.003)	(0.001)	(1.020)	(0.126)	(0.002)	(0.101)	(0.088)
12	-2.977^{***}	0.035	-0.010^{***}	-0.000	0.228	-0.080^{*}	0.001	0.279^{***}	0.648^{***}
	(0.218)	(0.057)	(0.002)	(0.001)	(0.514)	(0.039)	(0.002)	(0.067)	(0.058)
13	-5.143^{***}	-0.035	-0.010^{***}	-0.001	0.202	-0.241^{**}	0.010^{***}	0.102	0.651^{***}
	(0.364)	(0.085)	(0.003)	(0.001)	(0.722)	(0.084)	(0.002)	(0.093)	(0.086)
14	-3.422^{***}	-0.112^{*}	-0.006^{***}	0.001	0.102	-0.134^{***}	-0.005	0.176^{**}	0.453^{***}
	(0.240)	(0.055)	(0.002)	(0.001)	(0.439)	(0.030)	(0.003)	(0.068)	(0.056)
15	-2.171^{***}	-0.162	-0.016^{***}	-0.003	2.449^{*}	-0.611^{**}	-0.015^{***}	-0.113	0.270
	(0.531)	(0.164)	(0.005)	(0.002)	(1.186)	(0.210)	(0.004)	(0.178)	(0.165)
16	-5.102^{***}	-0.171	-0.004	-0.002	-1.217	-0.678^{***}	-0.001	0.661^{***}	1.330^{***}
	(0.392)	(0.100)	(0.003)	(0.001)	(0.986)	(0.135)	(0.003)	(0.136)	(0.111)
17	-3.759^{***}	-0.227^{*}	-0.021^{***}	-0.000	0.378	-0.357^{*}	0.003	0.174	0.634^{***}
	(0.408)	(0.101)	(0.003)	(0.001)	(0.710)	(0.150)	(0.002)	(0.117)	(0.102)
18	-3.668^{***}	0.281^{**}	-0.006^{*}	-0.001	-0.093	-0.189^{***}	-0.004	0.455^{***}	1.233^{***}
	(0.349)	(0.090)	(0.003)	(0.001)	(0.895)	(0.043)	(0.004)	(0.118)	(0.097)
19	-5.313^{***}	-0.194	-0.015^{***}	-0.002	0.466	-0.176^{*}	0.006	0.058	1.179^{***}
	(0.508)	(0.118)	(0.003)	(0.001)	(0.995)	(0.085)	(0.004)	(0.135)	(0.132)
20	-4.486^{***}	-0.037	-0.001	-0.001	0.643	-0.326^{**}	0.006^{***}	0.019	0.913^{***}
	(0.308)	(0.087)	(0.003)	(0.001)	(0.835)	(0.123)	(0.002)	(0.095)	(0.090)
21	-2.019^{***}	-0.035	-0.002	-0.002	1.752	-0.018	-0.006^{*}	0.877^{***}	0.492^{***}
	(0.335)	(0.081)	(0.003)	(0.001)	(0.989)	(0.081)	(0.003)	(0.108)	(0.086)
22	-4.550^{***}	0.059	0.000	-0.000	-2.529^{*}	0.144	0.008^{*}	0.513^{***}	0.371^{**}
	(0.461)	(0.106)	(0.003)	(0.001)	(1.254)	(0.089)	(0.004)	(0.134)	(0.113)

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. The other coefficients, not reported in Table 3, (intercept, gender, age, severity scores (COPS2 and CHMR), number of departures in the previous 2 hours, the average LAPS2 score in the ICU, weekday indicator, and day time indicator) from regression (1) for the *waiting* decision, estimated for each individual hospital respectively.

Table 17 Estimation results for other coefficients in multinomial regression (1) by hospital: decision = *Admission*

Hosp	Intercept γ_0	Gender γ_G	Age γ_A	COSP2 γ_{CP}	CHMR γ_{CH}	DepPre γ_{Dep}	AvgLAPS2 γ_{AvgL}	WeekDay γ_{WD}	DayTime γ_{DT}
1	-1.375^{***}	-0.153^{**}	-0.027^{***}	-0.004^{***}	1.691^{***}	-0.037	0.001	-0.003	-0.107^{*}
	(0.212)	(0.053)	(0.002)	(0.001)	(0.368)	(0.030)	(0.002)	(0.060)	(0.055)
2	-0.792^{**}	-0.163^{**}	-0.027^{***}	0.000	-1.268^{**}	-0.005	-0.003	-0.129	-0.372^{***}
	(0.279)	(0.060)	(0.002)	(0.001)	(0.412)	(0.035)	(0.003)	(0.067)	(0.063)
3	-3.380^{***}	-0.271^{**}	-0.028^{***}	-0.002	-1.402^{*}	0.201^{*}	0.005^{**}	0.083	-0.061
	(0.307)	(0.090)	(0.003)	(0.001)	(0.640)	(0.084)	(0.002)	(0.100)	(0.092)
4	-3.181^{***}	-0.023	-0.029^{***}	-0.003^{***}	0.817^{*}	0.002	0.001	0.018	-0.008
	(0.310)	(0.063)	(0.002)	(0.001)	(0.380)	(0.032)	(0.003)	(0.078)	(0.069)
5	-2.000^{***}	-0.166	-0.023^{***}	-0.002	-1.906^{**}	0.179^{*}	-0.000	-0.107	-0.013
	(0.297)	(0.086)	(0.003)	(0.001)	(0.669)	(0.081)	(0.002)	(0.094)	(0.088)
6	-2.165^{***}	-0.266^{***}	-0.026^{***}	-0.004^{***}	0.373	0.017	-0.002	-0.077	-0.064
	(0.252)	(0.064)	(0.002)	(0.001)	(0.416)	(0.042)	(0.002)	(0.070)	(0.066)
7	-2.018^{***}	-0.004	-0.024^{***}	-0.005^{**}	1.614^{*}	0.192^{*}	-0.002	0.057	-0.011
	(0.385)	(0.122)	(0.004)	(0.002)	(0.707)	(0.098)	(0.003)	(0.134)	(0.124)
8	-1.559^{***}	-0.116	-0.023^{***}	-0.000	-0.808	0.142^{***}	-0.000	0.035	-0.129^{*}
	(0.233)	(0.060)	(0.002)	(0.001)	(0.451)	(0.043)	(0.002)	(0.068)	(0.062)
9	-1.873^{***}	-0.178^{**}	-0.029^{***}	-0.000	-0.499	0.088^{*}	-0.001	-0.178^{**}	-0.258^{***}
	(0.253)	(0.063)	(0.002)	(0.001)	(0.435)	(0.035)	(0.002)	(0.068)	(0.064)
10	-1.590^{***}	-0.207^{*}	-0.035^{***}	-0.004^{**}	-1.421^{*}	0.048	-0.002	-0.007	0.143
	(0.335)	(0.100)	(0.003)	(0.001)	(0.618)	(0.081)	(0.002)	(0.110)	(0.103)
11	-1.663^{***}	-0.068	-0.019^{***}	-0.005^{**}	0.470	0.369^{**}	0.000	0.010	-0.533^{***}
	(0.362)	(0.114)	(0.004)	(0.001)	(0.923)	(0.126)	(0.002)	(0.125)	(0.115)
12	-2.607^{***}	-0.059	-0.028^{***}	-0.002	-0.830	0.048	-0.002	0.051	0.312^{***}
	(0.348)	(0.091)	(0.003)	(0.001)	(0.607)	(0.059)	(0.003)	(0.103)	(0.093)
13	-2.423^{***}	-0.153	-0.030^{***}	-0.004^{***}	-0.082	-0.002	0.001	-0.127	-0.131
	(0.293)	(0.078)	(0.002)	(0.001)	(0.520)	(0.069)	(0.002)	(0.085)	(0.079)
14	-1.833^{***}	-0.242^{***}	-0.031^{***}	-0.003^{***}	1.713^{***}	0.018	0.002	-0.016	-0.029
	(0.240)	(0.060)	(0.002)	(0.001)	(0.360)	(0.030)	(0.003)	(0.068)	(0.061)
15	-2.517^{***}	-0.423^{***}	-0.029^{***}	-0.003^{**}	-0.940	-0.066	0.003	-0.179	0.144
	(0.290)	(0.087)	(0.003)	(0.001)	(0.622)	(0.080)	(0.002)	(0.094)	(0.087)
16	-2.054^{***}	-0.190^{*}	-0.031^{***}	-0.002^{*}	-1.635^{**}	0.028	-0.003	-0.003	0.003
	(0.305)	(0.087)	(0.003)	(0.001)	(0.583)	(0.084)	(0.002)	(0.098)	(0.091)
17	-1.308^{***}	-0.184	-0.033^{***}	-0.003^{*}	-0.065	0.491^{***}	-0.001	0.034	0.008
	(0.376)	(0.110)	(0.003)	(0.001)	(0.705)	(0.110)	(0.002)	(0.122)	(0.111)
18	-2.863^{***}	-0.149	-0.026^{***}	-0.004^{***}	-0.607	-0.026	-0.001	0.160	0.126
	(0.343)	(0.092)	(0.003)	(0.001)	(0.635)	(0.043)	(0.004)	(0.108)	(0.096)
19	-2.057^{***}	-0.159^{*}	-0.028^{***}	-0.004^{***}	-0.391	-0.001	0.002	-0.026	-0.176^{*}
	(0.281)	(0.073)	(0.002)	(0.001)	(0.506)	(0.052)	(0.002)	(0.080)	(0.075)
20	-1.739^{***}	-0.031	-0.033^{***}	-0.002^{*}	-0.846	-0.001	-0.002	-0.034	-0.162
	(0.314)	(0.099)	(0.003)	(0.001)	(0.735)	(0.126)	(0.002)	(0.108)	(0.101)
21	-3.389^{***}	-0.132	-0.017^{***}	-0.003	-0.384	0.360^{**}	0.008	0.059	-0.492^{***}
	(0.566)	(0.135)	(0.004)	(0.002)	(1.205)	(0.116)	(0.004)	(0.154)	(0.139)
22	-1.809^{***}	-0.060	-0.013^{***}	-0.003^{*}	0.272	0.119	-0.004	0.076	-0.398^{***}
	(0.488)	(0.120)	(0.003)	(0.001)	(0.959)	(0.098)	(0.004)	(0.139)	(0.120)

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. The other coefficients, not reported in Table 3, (intercept, gender, age, severity scores (COPS2 and CHMR), number of departures in the previous 2 hours, the average LAPS2 score in the ICU, weekday indicator, and day time indicator) from regression (1) for the *admission* decision, estimated for each individual hospital respectively.

Table 18 Select coefficients for logit regression (32) by hospital

Hosp	Size	LAPS2 γ_L	ICUOccu γ_{ICU}	EDNum γ_{ED}	R^2
1	11,538	0.022^{***}	-0.125	-0.034	0.13
		(0.001)	(0.095)	(0.019)	
2	9,844	0.026^{***}	-0.454^{***}	-0.081^{***}	0.12
		(0.001)	(0.105)	(0.018)	
3	8,000	0.031^{***}	-0.230	-0.022	0.15
		(0.002)	(0.163)	(0.044)	
4	14,512	0.032^{***}	-0.069	-0.020	0.21
		(0.001)	(0.116)	(0.021)	
5	4,990	0.031^{***}	-0.503^{***}	-0.106^{*}	0.16
		(0.002)	(0.152)	(0.048)	
6	10,517	0.029^{***}	0.188	-0.080^{**}	0.16
		(0.001)	(0.112)	(0.026)	
7	4,748	0.019^{***}	-1.085^{***}	0.064	0.12
		(0.002)	(0.205)	(0.065)	
8	8,236	0.024^{***}	-0.395^{***}	-0.053^{*}	0.11
		(0.001)	(0.106)	(0.026)	
9	12,266	0.029^{***}	-0.199	-0.053^{**}	0.16
		(0.001)	(0.113)	(0.019)	
10	5,834	0.031^{***}	-0.272	-0.113^{*}	0.16
		(0.002)	(0.174)	(0.051)	
11	2,611	0.030^{***}	-0.875^{***}	-0.038	0.18
		(0.002)	(0.202)	(0.090)	
12	6,627	0.035^{***}	-0.586^{***}	-0.103^{***}	0.20
		(0.002)	(0.167)	(0.030)	
13	7,977	0.031^{***}	-0.142	-0.065	0.18
		(0.002)	(0.140)	(0.038)	
14	12,633	0.022^{***}	-0.132	-0.016	0.16
		(0.001)	(0.106)	(0.021)	
15	6,918	0.030^{***}	-0.026	-0.011	0.15
		(0.002)	(0.152)	(0.042)	
16	7,063	0.031^{***}	-0.287	-0.090^{*}	0.16
		(0.002)	(0.154)	(0.045)	
17	3,396	0.029^{***}	-0.961^{***}	-0.068	0.19
		(0.002)	(0.190)	(0.080)	
18	6,979	0.029^{***}	0.066	-0.031	0.14
		(0.002)	(0.165)	(0.043)	
19	7,419	0.028^{***}	-0.006	-0.079^{*}	0.16
		(0.001)	(0.132)	(0.034)	
20	5,032	0.029^{***}	-0.366^{*}	-0.048	0.16
		(0.002)	(0.178)	(0.058)	
21	2,101	0.033^{***}	-0.769^{***}	0.040	0.21
		(0.003)	(0.233)	(0.076)	
22	2,030	0.027^{***}	-0.358	-0.167^{*}	0.15
		(0.003)	(0.208)	(0.078)	

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Select coefficients (LAPS2 score, ICU occupancy level, and number of ED patients) from (32) for admission (versus reroute) decision.

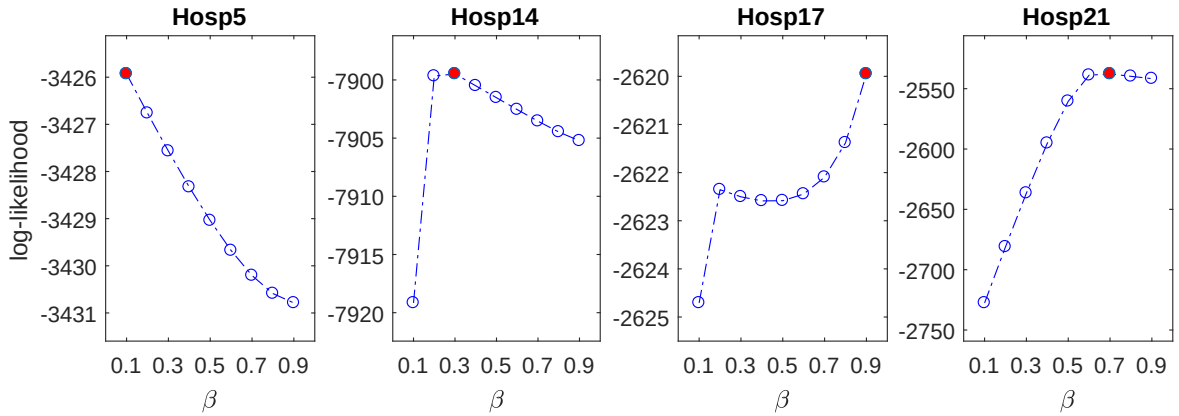
Table 19 Estimation results for other coefficients in logit regression (32) by hospital

Hosp	Intercept γ_0	Gender γ_G	Age γ_A	COSP2 γ_{CP}	CHMR γ_{CH}	DepPre γ_{Dep}	AvgLAPS2 γ_{AvgL}	WeekDay γ_{WD}	DayTime γ_{DT}
1	-1.411^{***}	-0.169^{**}	-0.028^{***}	-0.004^{***}	1.645^{***}	-0.039	0.001	-0.023	-0.124^{*}
	(0.216)	(0.054)	(0.002)	(0.001)	(0.374)	(0.031)	(0.002)	(0.060)	(0.055)
2	-0.785^{**}	-0.176^{**}	-0.027^{***}	-0.000	-1.170^{**}	-0.007	-0.003	-0.139^{*}	-0.400^{***}
	(0.282)	(0.060)	(0.002)	(0.001)	(0.421)	(0.036)	(0.003)	(0.068)	(0.064)
3	-3.380^{***}	-0.263^{**}	-0.028^{***}	-0.002	-1.334^{*}	0.208^{*}	0.005^{**}	0.072	-0.070
	(0.308)	(0.090)	(0.003)	(0.001)	(0.646)	(0.084)	(0.002)	(0.100)	(0.093)
4	-3.188^{***}	-0.040	-0.030^{***}	-0.003^{***}	0.755^{*}	0.012	0.001	0.010	-0.052
	(0.313)	(0.064)	(0.002)	(0.001)	(0.384)	(0.032)	(0.003)	(0.078)	(0.070)
5	-1.850^{***}	-0.172^{*}	-0.023^{***}	-0.002	-1.904^{**}	0.186^{*}	-0.001	-0.107	-0.045
	(0.298)	(0.087)	(0.003)	(0.001)	(0.672)	(0.083)	(0.002)	(0.095)	(0.089)
6	-2.212^{***}	-0.266^{***}	-0.026^{***}	-0.004^{***}	0.356	0.019	-0.002	-0.089	-0.081
	(0.254)	(0.064)	(0.002)	(0.001)	(0.418)	(0.042)	(0.002)	(0.071)	(0.066)
7	-1.977^{***}	0.001	-0.024^{***}	-0.005^{**}	1.648^{*}	0.184	-0.002	0.029	-0.008
	(0.385)	(0.123)	(0.004)	(0.002)	(0.714)	(0.099)	(0.003)	(0.135)	(0.125)
8	-1.605^{***}	-0.137^{*}	-0.023^{***}	-0.000	-0.903^{*}	0.139^{**}	-0.000	0.031	-0.128^{*}
	(0.237)	(0.061)	(0.002)	(0.001)	(0.458)	(0.043)	(0.002)	(0.069)	(0.063)
9	-1.824^{***}	-0.150^{*}	-0.030^{***}	-0.001	-0.285	0.087^{*}	-0.001	-0.185^{**}	-0.245^{***}
	(0.256)	(0.064)	(0.002)	(0.001)	(0.448)	(0.036)	(0.002)	(0.069)	(0.065)
10	-1.611^{***}	-0.198^{*}	-0.035^{***}	-0.004^{**}	-1.380^{*}	0.049	-0.002	-0.025	0.133
	(0.336)	(0.100)	(0.003)	(0.001)	(0.621)	(0.081)	(0.002)	(0.110)	(0.103)
11	-1.808^{***}	-0.034	-0.020^{***}	-0.005^{**}	0.060	0.377^{**}	0.000	-0.038	-0.544^{***}
	(0.366)	(0.117)	(0.004)	(0.001)	(0.926)	(0.131)	(0.002)	(0.128)	(0.118)
12	-2.607^{***}	-0.098	-0.028^{***}	-0.002	-0.916	0.056	-0.001	0.018	0.229^{*}
	(0.355)	(0.093)	(0.003)	(0.001)	(0.636)	(0.060)	(0.003)	(0.105)	(0.096)
13	-2.381^{***}	-0.152	-0.030^{***}	-0.004^{***}	-0.013	-0.005	0.000	-0.113	-0.144
	(0.294)	(0.079)	(0.002)	(0.001)	(0.523)	(0.070)	(0.002)	(0.085)	(0.080)
14	-1.872^{***}	-0.251^{***}	-0.031^{***}	-0.003^{***}	1.806^{***}	0.022	0.002	-0.012	-0.037
	(0.242)	(0.060)	(0.002)	(0.001)	(0.366)	(0.030)	(0.003)	(0.068)	(0.061)
15	-2.539^{***}	-0.407^{***}	-0.029^{***}	-0.003^{**}	-0.890	-0.064	0.003	-0.179	0.137
	(0.290)	(0.087)	(0.003)	(0.001)	(0.626)	(0.080)	(0.002)	(0.094)	(0.087)
16	-2.083^{***}	-0.163	-0.031^{***}	-0.002^{*}	-1.598^{**}	0.024	-0.003	0.004	-0.006
	(0.305)	(0.088)	(0.003)	(0.001)	(0.583)	(0.084)	(0.002)	(0.098)	(0.091)
17	-1.242^{**}	-0.175	-0.036^{***}	-0.003^{*}	-0.084	0.486^{***}	-0.001	-0.010	-0.018
	(0.381)	(0.112)	(0.003)	(0.001)	(0.699)	(0.111)	(0.002)	(0.124)	(0.113)
18	-2.780^{***}	-0.150	-0.026^{***}	-0.004^{***}	-0.560	-0.035	-0.001	0.179	0.112
	(0.343)	(0.093)	(0.003)	(0.001)	(0.649)	(0.044)	(0.004)	(0.109)	(0.096)
19	-2.077^{***}	-0.166^{*}	-0.028^{***}	-0.004^{***}	-0.422	-0.006	0.002	-0.027	-0.174^{*}
	(0.283)	(0.073)	(0.002)	(0.001)	(0.508)	(0.053)	(0.002)	(0.081)	(0.075)
20	-1.681^{***}	-0.030	-0.033^{***}	-0.003^{*}	-0.770	0.001	-0.002	-0.040	-0.172
	(0.317)	(0.100)	(0.003)	(0.001)	(0.739)	(0.126)	(0.002)	(0.108)	(0.101)
21	-3.510^{***}	-0.127	-0.019^{***}	-0.003	0.301	0.370^{**}	0.009^{*}	0.051	-0.536^{***}
	(0.577)	(0.140)	(0.004)	(0.002)	(1.240)	(0.120)	(0.004)	(0.158)	(0.143)
22	-1.758^{***}	-0.017	-0.014^{***}	-0.003^{*}	0.231	0.145	-0.003	0.068	-0.446^{***}
	(0.495)	(0.121)	(0.004)	(0.001)	(0.979)	(0.099)	(0.004)	(0.141)	(0.121)

Standard error is reported in parenthesis; $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

The other coefficients, not reported in Table 18, (intercept, gender, age, severity scores (COPS2 and CHMR), number of departures in the previous 2 hours, the average LAPS2 score in the ICU, weekday indicator, and day time indicator) from regression (1) for admission (versus reroute) decision by each individual hospital.

Figure 7 Examples of log-likelihood versus discount factor for a subset of hospital



Note: The estimated likelihood at $\beta = 0.1, 0.2, \ldots, 0.9$ for four different hospitals (5, 14, 17, and 21). The β with the best log-likelihood is highlighted in red.

Appendix B: Proof for Proposition 2

PROOF:

Preliminaries : From the definition of the value functions $V(s_t, \varepsilon_t)$ in (6) and $\tilde{V}(s)$ in (11), we have:

$$\begin{aligned}\tilde{V}(s) &= \sum_{s_t} \int_{\varepsilon_t} V(s_t, \varepsilon_t) g(s_t|s) q(\varepsilon_t|s_t) d\varepsilon_t \\ &= \sum_{s_t} \int_{\varepsilon_t} \sup_{d_t \in \Pi(s_t)} \mathsf{E}\left\{ \sum_{j=t}^{\infty} \beta^{j-t} U(s_j, d_j, \varepsilon_j) | s_t, \varepsilon_t \right\} g(s_t|s) q(\varepsilon_t|s_t) d\varepsilon_t. \end{aligned} \tag{33}$$

The expectation above is taken over the transition of (s_j, ε_j) starting from (s_t, ε_t).

Recall $g(s'|s)$ is the transition probability from the intermediate state s (after action is taken) to the state s' at the start of the next period, and $q(\varepsilon'|s')$ is the probability density of the random utility component in the next period. Thus, the function $\tilde{V}(s)$ represents the expected future utilities starting from intermediate state s and assuming the hospital always takes the optimal action.

We introduce following notations. We consider two systems s and s'. For every period t, we use s_t and s'_t to denote the system states at the start of period t, which are

$$s_t = \left(n^E_{l,t}, n^E_{h,t}, n^I_t\right) \text{ and } s'_t = \left(\left(n^E_{l,t}\right)', \left(n^E_{h,t}\right)', \left(n^I_t\right)' \right).$$

Additionally, we use

$$\tilde{s}_t = \left(\tilde{n}^E_{l,t}, \tilde{n}^E_{h,t}, \tilde{n}^I_t\right) \text{ and } \tilde{s}'_t = \left(\left(\tilde{n}^E_{l,t}\right)', \left(\tilde{n}^E_{h,t}\right)', \left(\tilde{n}^I_t\right)' \right)$$

to denote the intermediate states after actions d_t and d'_t are taken in period t, which are given by $\tilde{s}_t = \varphi(s_t, d_t)$ and $\tilde{s}'_t = \varphi(s'_t, d'_t)$ according to (7). Note that for notational compactness we suppress the dependence of the states on the action. Assume the two systems start from intermediate states $\tilde{s}_0$ and $\tilde{s}'_0$ with

$$\tilde{n}^E_{i,0} = \left(\tilde{n}^E_{i,0}\right)' \text{ for } i \in \{l, h\}, \text{ and } \tilde{n}^I_0 \leq \left(\tilde{n}^I_0\right)', \tag{34}$$

then Proposition 2 translates to

$$\tilde{V}(\tilde{s}_0) \geq \tilde{V}(\tilde{s}'_0). \tag{35}$$

Coupling: Our proof is based on a coupling argument and induction in time. We first introduce the coupling of the two systems s and s' as follows: First, the two systems witness identical arrivals to ED and external arrivals to ICU in every period, i.e., $A_{l,t} = A'_{l,t}$, $A_{h,t} = A'_{h,t}$, and $E_t = E'_t$ for every t. Next, we couple the ICU departures from the two systems as follows. Denote the numbers of ICU patients before departures by $\bar{n}^I_t$ and $(\bar{n}^I_t)'$ respectively and assume $\bar{n}^I_t \leq (\bar{n}^I_t)'$. Then, the departures D_t and D'_t are coupled as $D'_t = D_t + Z_t$, where D_t is a Binomial-$(\bar{n}^I_t, \mu_I)$ variable and Z_t is a Binomial-$((\bar{n}^I_t)' - \bar{n}^I_t, \mu_I)$ variable. That is, the number of departures in the s' system is always at least as many as the number in the s system. Finally, if an identical action d is taken in each system, the random utility components $\varepsilon_t(d)$ and $\varepsilon'_t(d)$ associated with that action d coincides for the two systems in every period t.

Under the coupling described above, we first prove following lemma that establishes the relationship between intermediate states and states at the start of next period.

Lemma 1 *Under the coupling, if the intermediate states in each system in period $t-1$ satisfy*

$$\tilde{n}^E_{i,t-1} = \left(\tilde{n}^E_{i,t-1}\right)' \text{ for } i \in \{l,h\}, \text{ and } \tilde{n}^I_{t-1} \leq \left(\tilde{n}^I_{t-1}\right)', \tag{36}$$

then the states at the start of period t satisfy

$$n^E_{i,t} = \left(n^E_{i,t}\right)' \text{ for } i \in \{l,h\} \text{ and } n^I_t \leq \left(n^I_t\right)'. \tag{37}$$

PROOF: See Appendix B.1. □

Mimicking Policy: We now define the policies used in each system. We assume the system s' always takes its optimal action which achieve the supremum in (33). For system s, we define a mimicking policy π which mimics the action taken in the s' system whenever possible; if it is not possible, it takes its own optimal action. We denote the value function associated with this policy by $V^\pi(s)$, which is defined by (33) with optimal action d_t replaced by the one under policy π. Such a policy is not necessarily optimal for system s and, by definition, we have

$$\tilde{V}(\tilde{s}_0) \geq V^\pi(\tilde{s}_0). \tag{38}$$

To prove the proposition, we will establish following two properties under our coupling and the policy π. First, two systems always have same number of patients in the ED, but system s has no more patients in the ICU:

$$n^E_{i,t} = \left(n^E_{i,t}\right)' \text{ for } i \in \{l,h\} \text{ and } n^I_t \leq \left(n^I_t\right)', \quad \forall t. \tag{39}$$

Second, the action taken in the s' system is always admissible for system s; thus system s always mimics the action of s' under π:

$$d_t = d'_t \in \Pi(s_t), \quad \forall t. \tag{40}$$

Note that (39) directly implies (40), as it follows from (2) that given the same number of patients in ED, the system with fewer ICU patients has a larger admissible action set, leading to $d'_t \in \Pi(s'_t) \subseteq \Pi(s_t)$.

Induction: We establish (39) for every t by induction.

Base Case: The base case follows directly from the relationship of the initial intermediate states $\tilde{s}_0$ and $\tilde{s}'_0$, which satisfy (34), and from Lemma 1.

$$n^E_{i,1} = \left(n^E_{i,1}\right)' \text{ for } i \in \{l,h\} \text{ and } n^I_1 \leq \left(n^I_1\right)'. \tag{41}$$

Inductive Step: We assume (39) holds for period j and show this implies it holds for period $j+1$.

In period j, under policy π, system s takes the same action of s' since by the inductive hypothesis the action is admissible, i.e., $d_j = d'_j$. Given the same action is taken in each system, the intermediate states after action, satisfy the following relationship:

$$\begin{aligned} \tilde{n}^E_{i,j} &= n^E_{i,j} - a_{i,j} - r_{i,j} = \left(n^E_{i,j}\right)' - a'_{i,j} - r'_{i,j} = \left(\tilde{n}^E_{i,j}\right)' \text{ for } i \in \{l,h\}, \\ \tilde{n}^I_j &= n^I_j + a_{l,j} + a_{h,j} \leq \left(n^I_j\right)' + a'_{l,j} + a'_{h,j} = \left(\tilde{n}^I_j\right)'. \end{aligned} \tag{42}$$

Finally, we can apply Lemma 1 to prove the relationship (39) holds for period $j+1$. This completes the inductive step.

Per-Period Utilities: We have shown that under our coupling and the policy π for the s system, at the start of each period, the two systems always have same numbers of patients in the ED, and system s always has fewer patients in

the ICU than that in system s'. Thus, the system s always mimics the action by s' under the policy π. We next prove the per-period utilities always coincide for the two systems, which follows by:

$$\begin{aligned} U(s_t, d_t, \varepsilon_t) &= -c(s_t, d_t) + \varepsilon_t(d_t) = -\sum_{i\in\{l,h\}} c_{r.i} r_{i,t} - \sum_{i\in\{l,h\}} c_{w,i}\left(n^E_{i,t} - a_{i,t} - r_{i,t}\right) + \varepsilon_t(d_t) \\ &= -\sum_{i\in\{l,h\}} c_{r.i} r'_{i,t} - \sum_{i\in\{l,h\}} c_{w,i}\left(\left(n^E_{i,t}\right)' - a'_{i,t} - r'_{i,t}\right) + \varepsilon'_t(d'_t) = U(s'_t, d'_t, \varepsilon'_t). \end{aligned}$$

This is because: (i) Both systems take the same action, thus they admit and reroute same numbers of patients, this leads to same rerouting costs; (ii) As both systems have same numbers of patients in the ED, the number of patients remaining in the ED after actions are also the same, leading to the same waiting costs; (iii) By our coupling, the random utility components coincide for the same action $d_t = d'_t$, i.e., $\varepsilon_t(d_t) = \varepsilon'_t(d'_t)$.

As the per-period utilities coincide for every period given system s takes policy π and system s' takes its own optimal policy, we have

$$V^\pi(\tilde{s}_0) = \tilde{V}(\tilde{s}'_0),$$

then it follows by (38)

$$\tilde{V}(\tilde{s}_0) \geq V^\pi(\tilde{s}_0) = \tilde{V}(\tilde{s}'_0).$$

This proves the proposition. □

B.1. Proof of Lemma 1

PROOF: The result follow directly from the coupled arrivals and departures in the two systems. Since we start from the intermediate state, the system evolution to period t is only dictated by the stochastic arrivals to the ED, external arrivals to the ICU, and departures from the ICU during period $t-1$.

It is trivial to see the relationship for the ED patients holds by our coupling of the ED arrival processes. Since, by our coupling $A_{i,t-1} = A'_{i,t-1}$, then $n^E_{i,t} = \left(n^E_{i,t}\right)'$ for $i \in \{l,h\}$ directly follows from our assumption that $\tilde{n}^E_{i,t-1} = \left(\tilde{n}^E_{i,t-1}\right)'$.

We now consider the ICU patients. By our coupling, the s and s' systems see the same number external arrivals, $E_{t-1} = E'_{t-1}$. Then, the total number of ICU patients in each system before departures satisfies the following relationship:

$$\bar{n}^I_t = \min\left\{\tilde{n}^I_{t-1} + E_{t-1}, B\right\} \leq \min\left\{(\tilde{n}^I_{t-1})' + E'_{t-1}, B\right\} = (\bar{n}^I_t)', \tag{43}$$

By our coupling, the number of departures from the ICU in the s system is related to the number of departures in the s' system as follows: $D_{t-1} + Z_{t-1} = D'_{t-1}$. Thus,

$$\begin{aligned} \left(n^I_t\right)' - n^I_t &= \left(\bar{n}^I_{t-1}\right)' - \bar{n}^I_{t-1} - \left(D'_{t-1} - D_{t-1}\right) \\ &= \left(\bar{n}^I_{t-1}\right)' - \bar{n}^I_{t-1} - Z_{t-1} \geq 0. \end{aligned}$$

The last inequality follows as $Z_{t-1} \leq \left(\bar{n}^I_{t-1}\right)' - \bar{n}^I_{t-1}$. This completes the proof for Lemma 1. □