Structural Estimation of Intertemporal Externalities with Application in ICU Admissions

Yiwen Shen
Department of ISOM, HKUST, yiwenshen@ust.hk

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

Problem definition: In many service systems, the system manager needs to balance between addressing the needs of current customers and ensuring the system's ability to serve future customers. Such balancing behavior is particularly important in capacity-constrained systems with heterogeneous service levels, in which the manager needs to decide which level of service to provide to the current customer, taking into account the intertemporal externalities of their decisions.

Methodology/results: We develop a dynamic discrete choice model to describe the decision-making process in a gatekeeper system with multiple classes of servers and customers. The discount factor in the model captures how much the decision-maker internalizes the intertemporal externalities of their customer routing decisions. In contrast to most empirical studies in the literature which use a pre-specified discount factor, we establish joint identification of the discount factor and the utility parameters from data. We then apply the model to empirically study the Intensive Care Unit (ICU) admission decisions for Emergency Department (ED) patients. Using a large hospitalization data set, we find that there is large heterogeneity in the estimated discount factors across hospitals. Via counterfactual simulations, we show that correctly estimating the discount factor is crucial for hospitals to evaluate the ICU congestion levels and the impact of system changes.

Managerial implications: Our results suggest that it is important to understand how the decision-maker internalizes the intertemporal externalities from data. In addition, the balancing behavior regarding current customers and future available capacity provides a potential channel for improving system performance.

Key words: structural estimation, dynamic discrete choice model, empirical operations management, healthcare, intensive care unit

1. Introduction

In managing service systems, the decision-maker often has to carefully balance between providing immediate service to current customers versus the impact of such actions on the system's ability to serve future customers. Such trade-off is particularly relevant when the system is operating in a resource constrained environment. While providing immediate service may benefit current customers, it will utilize the limited resources in the service system; thus, such an action may hinder access to service for future, perhaps more

valuable, customers. This introduces *intertemporal externalities* in the decision-making, i.e., the impact of the current decision on the future system state matters.

These intertemporal externalities often arise in gatekeeper systems with heterogeneous service levels. In these systems, the decision-maker acts as a gatekeeper who decides which level of service to provide to the arriving customer. The higher-level service generally provides more benefits to the customers, but is usually more capacity-constrained than the lower-level service. When making decisions, the gatekeeper needs to consider not only the current customer's type, but also the system's state after the decision which determines its ability to serve future customers. Gatekeeper systems can be found in many real-world settings and application areas. For example, call centers and customer service departments may employ both general and specialist agents, with the latter focusing more on urgent and complex inquiries, and a gatekeeper deciding which type of agents should serve each customer (Shumsky and Pinker 2003, Hathaway et al. 2023). Border security-check systems usually have two stages of inspections. The inspectors in the first stage act as gatekeepers for the second stage, which is designed for more complicated cases (Zhang et al. 2011). Such systems are also common in healthcare settings, where different levels of care are often provided to the same type of patients. In maternity department, midwives act as gatekeepers to specialists, but may also try to serve patients themselves (Freeman et al. 2016). In emergency department, physicians may admit waiting patients to intensive care units or ordinary floor units (Kim et al. 2015).

We take a structural estimation approach to understand how the decision-maker balances the needs of current customers versus saving system capacity for future customers in such gatekeeper systems. We first develop a dynamic discrete choice model that describes the decision-making process. The model includes two service units: a first-tier service unit (FSU) which provides higher level of service but with limited capacity, and a second-tier service unit (SSU) that provides lower level of service but with ample capacity. Customers are categorized as the low-priority class and high-priority class, with the former more likely to benefit from the first-tier service. In each period, the system considers three actions for each customer arriving or waiting at the gatekeeper (GK): 1) send the customer to the FSU for service, 2) send to the SSU, or 3) keep the customer waiting at the GK.[1] We assume the system chooses the action that maximizes its discounted utility over an infinite horizon. Our model incorporates stochastic arrivals and departures, external arrivals to the FSU, and idiosyncratic shocks to the utility of each action.

While the FSU can provide the highest level of service, its limited capacity means that sending a customer to it may hinder the service for future high priority customers. This introduces intertemporal externalities in the decision-making process. That is, the system needs to balance the utility of current customer and its capacity to serve future customers. We quantify how the system balances such trade-off in practice by estimating the *intertemporal discount factor* in the model from observed data. The discount factor represents

[1] In some service systems, the gatekeepers can provide the lower-level service by themselves. However, the trade-off between current and future customers still applies. Our model can be extended to these settings as well.

the relative weight of the future periods' utility in the system's objective function. Thus, a larger discount factor implies that the system accounts more for the impact of its decisions on future system state; while a smaller discount factor means that the system is more focused on current customers.

In this work, we estimate the discount factor and the utility parameters *jointly* from data. This is in stark contrast to the majority of empirical studies with dynamic discrete choice models, which assume the discount factor to be known and only estimate the utility parameters (see, e.g., Rust 1987, Bajari et al. 2007, and Mehta et al. 2017). Moreover, the discount factor is generally set at a high level close to one (e.g., 0.975) without empirical support and formal justifications. However, the discount factors in dynamic discrete choice models can vary dramatically depending on the context of the problem, or as a consequence of the behavior of decision makers. We estimate the discount factor from data to understand how the decision-maker responds to intertemporal externalities in practice, which also facilitates accurate evaluations of how different shocks impact the system performance.

It is well known that the dynamic discrete choice model can not be identified from choice data without further conditions (see e.g., Manski 1993, Magnac and Thesmar 2002), which is a primary reason that most empirical studies need to assume a known discount factor. The non-identification stems from the existence of observationally equivalent structures: multiple combinations of discount factors and utility parameters can lead to the same choice probabilities for all states (Rust 1994). We circumvent this difficulty by leveraging Komarova et al. (2018) and extending the approach to our dynamic discrete choice model. Much like other general econometric methodologies, applying the theoretical results in Komarova et al. (2018) to our specific model is a challenging task that must be executed carefully for it to be valid. For this purpose, we develop the necessary components of the identification method tailored to our setting, including the modeling assumptions, estimation procedure, as well as discussion on the impact of the normalization choice in the estimation.

Next, we apply our discrete choice model and identification results of a gatekeeper system to a suitable and important empirical setting: hospital's Intensive Care Unit (ICU) admission decisions for patients admitted via the Emergency Department (ED). ICUs are specialized inpatient units which provide the highest level of care for the most critically ill patients. They are expensive medical resources and often operate at high occupancy levels (Coopersmith et al. 2012, Halpern and Pastores 2015). Consequently, the ICU is often identified as a critical process bottleneck in a hospital; 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). All these factors make ICU beds a key resource in a hospital which must be managed effectively. As the ICU provides the highest level of care, swift admission generally benefits the patients who need ICU care. However, given the constrained capacity and high occupancy of the ICU, this may restrict access to ICU care for future, perhaps more severe, patients. Such a trade-off introduces intertemporal externalities in the system gatekeeper's ICU admission decisions.

We use the dynamic discrete choice model to estimate the ICU admission decisions, with the discount factor reflecting how hospitals balance the tradeoff between current versus future patients. We divide ED patients into low- and high-severity classes based on their risk scores, capturing the heterogeneity in their needs for ICU care. For each ED patient, the hospitals can make three decisions: admit to the ICU, admit to non-ICU units (e.g., wards), or keep the patient waiting in the ED. In this setting, the ED plays the role of gatekeeper for the ICU, which represents the first-tier service unit (FSU) with the highest level of service with limited capacity. The non-ICU units provide lower-level of care with ample capacity, and are viewed as the second-tier service unit (SSU). Thus, the decision-making process for the ED patient fits into our gatekeeper's decision-making described above. We emphasize that the ICU admission is a complex process involving multiple stakeholders and resources. As such, the utility maximization problem in the structural model is used to estimate the hospitals' perceived behaviors consistent with the observed data, which is common in the healthcare literature (see, e.g., Olivares et al. 2008, Wang et al. 2019, Rath and Rajaram 2022).

We implement our structural model empirically on a large data set from a major US hospital network, including 21 hospitals with more than 300,000 hospitalizations. The discount factor and action utilities in the model are estimated jointly from the observed admission decisions. We find that the identified discount factor on average is much lower than the levels usually assumed in the literature. Moreover, there is large heterogeneity in the estimated discount factors across hospitals, revealing a novel aspect of the practice variation observed in medical literature (Corallo et al. 2014). In particular, hospitals with larger discount factors tend to be more responsive to ICU congestion by adjusting their admission probabilities proactively. Our estimation results suggest that it is important to identify discount factor from data in different empirical settings, instead of using a pre-assumed level. We show that our estimated model fits the data well in multiple aspects and support our main findings by additional robustness checks.

We use counterfactual analyses to quantify how the hospital's balancing behaviors regarding current patient and system-level considerations impact ICU congestion. First, we reveal the intertemporal trade-off captured by the discount factor. We show that with a larger discount factor in the model, i.e., assuming hospitals account more for the impact on future system state when making decisions, hospitals can reduce their high ICU congestion periods substantially. For some hospitals, the improvement is comparable to the costly intervention of adding an ICU bed and related staff. This reflects the benefit of smoothing workload across periods. However, increasing the discount factor leads to longer ED waiting time of the patients, with the negative effect being more significant for the low-severity class. Thus, the level of discount factor introduces a trade-off between ICU congestion and ED waiting time. Next, we show that misspecifying the discount factor in the model can lead to substantial bias in the estimates of ICU congestion. In particular, if we assume a large (resp. small) discount factor and only estimate the utility parameters, we would underestimate (resp. overestimate) the frequency of high ICU congestion periods. Such biases become more

significant under the scenarios where the ICU capacity becomes more constrained. Thus, correctly identifying the discount factor from data is important to accurately estimate the system performance and the impact of different shocks.

In summary, we make the following main contributions. First, we develop a general dynamic discrete choice model with heterogeneous service levels and customer types. The model captures how the decision-maker balances the needs of current customers and the system's ability to serve future customers. We then establish the joint identification of discount factor and utility parameters from data. Next, we apply our model empirically to the ICU admission problem using a large hospitalization dataset. We find that the identified discount factor is far from the levels usually assumed in the literature, and that there is large heterogeneity across hospitals. Finally, we use counterfactual simulations to show that correctly understanding the intertemporal externalities is crucial for managing ICU congestion.

The rest of the paper is organized as follows. We conclude this section with a brief literature review. Section 2 develops the main dynamic discrete choice model and establishes the identification of discount factor. Section 3 describes how we apply the model to the ICU admission decisions. Section 4 presents the main empirical results. Section 5 conducts counterfactual studies and reveals the intertemporal trade-off induced by the discount factor. We conclude the paper and discuss future research directions in Section 6. Other auxiliary results are collected in the electronic companion.

1.1. Literature Review

Our work is related to the following streams of literature: gatekeeper systems with heterogeneous servers; structural estimation in operations management; identification of dynamic choice models; and empirical healthcare operations management.

At the high level, our work is related to gatekeeper systems in operations management (Shmueli et al. 2003, Zhang et al. 2011). Various works have examined the behaviors and implications of gatekeepers, especially in healthcare settings. Batt and Terwiesch (2016) find that, when the ED is congested, nurses will initiate diagnostic tests at the triage stage to reduce waiting times. Freeman et al. (2016) show that midwives in delivery units ration resource-intensive service and increase the rate of specialist referrals when workload increases. Freeman et al. (2021) find that adding a second gatekeeping stage in the ED reduces the rates of unnecessary hospitalization and wrongful patient discharge. We consider a gatekeeper that routes customers to different levels of service. From the queueing aspect, our work is related to the literature on queueing models with heterogeneous servers (e.g., Mandelbaum et al. 2012). In contrast to solving the optimal referring/routing policy, we use a structural estimation approach to understand how the system behaves according to observed data.

The structural estimation approach has been increasingly used in operations management to understand the behaviors of customers and servers in different empirical settings. Bray et al. (2019) develop a dynamic

discrete choice model and show that ration gaming exists in a multiechelon supply chain. Using structural estimation from air-travel industry, Li et al. (2014) find that customers sometimes delay purchases strategically when they expect the price to fall. Akşin et al. (2013) and Yu et al. (2016) take the structural estimation approach to study caller's abandonment behaviors in call centers. Emadi and Staats (2020) find that attrition of agents at a management firm appears to be insensitive to salary. Structural models have also been used in healthcare operations management. 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 (OR) capacity for cardiac surgery. Rath and Rajaram (2022) use a choice model to estimate costs associated with OR scheduling of anesthesiologists.

In this study, we develop a dynamic discrete choice model to understand the discounting behavior in a gatekeeper system with heterogeneous service levels. The dynamic choice model we use is pioneered by Rust (1987). We estimate the discount factor and utility parameters jointly from data, and apply our model to empirically study the ICU admission decisions from a large hospital network. To our best knowledge, we are the first to estimate a dynamic structural model with unknown discount factor in the empirical operations literature. In most empirical studies with dynamic choice model, the discount factor is pre-assumed at a very high level such as 0.975 (e.g., Bray et al. 2019, Emadi and Staats 2020). This is because the identification of discount factor from data is generally a very difficult problem (Rust 1994, Magnac and Thesmar 2002). It has been done in only a few recent studies in very different settings, e.g., De Groote and Verboven (2019) and Ching and Osborne (2020). In both cases, identification is achieved by leveraging special features in the data based on their respective empirical settings. To overcome this challenge, we validate and extend Komarova et al. (2018), which established theoretically joint identification of the discount factor and payoff parameters for dynamic choice models with linear structure. We are not aware of other work in operations management that adapts this abstract methodology in an empirical setting.

Our empirical study of the ICU admission decisions contributes to the literature on empirical healthcare operations management. A number of works study patient admission decisions as we do in this work. For instance, Edbrooke et al. (2011) and Kim et al. (2015) investigate the impact of ICU admissions on multiple patient outcomes, including mortality, hospital length of stay (LOS), readmission rate, and transfers to higher levels of care. Chan et al. (2016) find that delays in ICU admission can increase ICU LOS, which, in turn, can create more congestion in an already busy unit. Patients who are not admitted to the unit of choice are typically rerouted to alternative units designated for a different service. Song et al. (2020) and Dong et al. (2018) study off-placement of patients when bed availability in the primary unit is limited. They show that such off-placement has important clinical and operational implications. There is substantial evidence that system state can influence physician admission decisions and patient flow dynamics. For instance, ICU congestion can impact who is admitted to the ICU (Kim et al. 2015) and when patients are discharged (Kc and Terwiesch 2012). In a setting similar to ours, Kim et al. (2020) study the ICU admission decision

from a behavioral perspective. They propose a behavioral model and use controlled experiments to identify a number of factors which can bias physician's admission decisions. In this study, we take a structural estimation approach to understand the balancing behaviors in ICU admissions decisions from data.

2. Model Set-up and Identification of Discount Factor

In this section, we introduce the structural model of the gatekeeper system, explain the identification of the discount factor, and describe the estimation procedure.

2.1. Dynamic Discrete Choice Model for Gatekeeper Systems

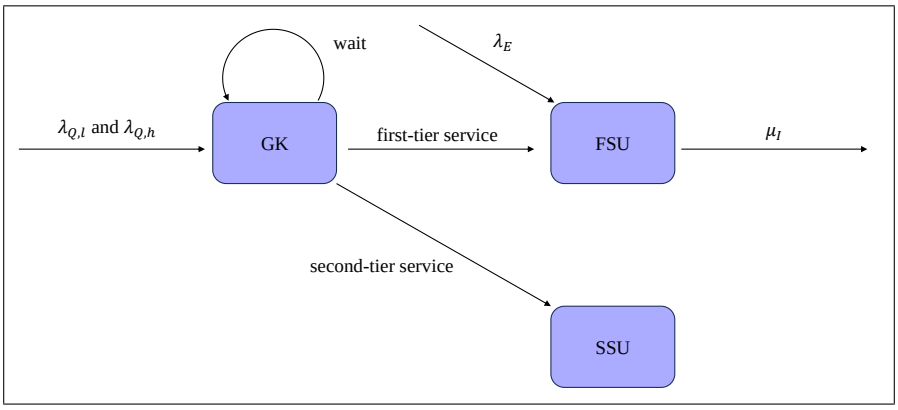
In our structural model, we consider the routing decision of a gatekeeper (GK) for each customer about the downstream unit which serves them. We assume that there are two units, i.e., groups of servers, that provide different levels of service: the first-tier service unit (FSU) and the second-tier service unit (SSU). The FSU provides a higher level of service that generally benefits the customers. However, its capacity is usually more constrained than the SSU. There are three key features of the model. First, in each period, the GK considers which service unit, if any, to allocate to each customer waiting in the system. Second, such routing decision depends on both the customer type and the current system status, i.e., the remaining capacity of the service units. Finally, the model allows for the system to potentially consider the future state and to make dynamic decisions. We provide detailed descriptions of the model below.

We divide the arriving customers into two classes: the low-priority class and the high-priority class, represented by subscripts l and h, respectively. The model set-up can be extended to include multiple customer classes as well. The customers in the high-priority class usually requires a higher level of service than those in the low-priority class. We assume that the GK has capacities Q_l and Q_h for the low-priority and high-priority classes, respectively. This can capture the limited waiting area at the GK. The two classes of customers arrive to the GK every period according to the following distribution. For $i \in \{l, h\}$, let $A_{i,t}$ be the number of class i customers arriving to the GK 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 a sufficiently high upper bound $M_{A,i}$. This reduces the state space of the model to keep the estimation computationally feasible without imposing additional assumptions.

In addition to customers admitted via the GK, there are other arriving customers who may also require the first-tier service. Let E_t denotes the number of customers arriving to the FSU via other channels in period t, referred to as the external arrivals. E_t is distributed according to a Poisson distribution with arrival rate λ_E. Every period, each current customer in the FSU completes their first-tier service and departs from the unit with probability μ_I. Thus, D_t, the number of customers departing from the FSU in each period, follows a binomial distribution. The total number of servers in the FSU is B, meaning that the first-tier service can be provided to at most B customers simultaneously. On the other hand, we assume the SSU has ample

capacity to focus on the intertemporal externalities related to the capacity-constrained FSU.[2] The system flow is depicted in Figure 1.

Figure 1 Overview of the system flow



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

$$s_t = (n^G_{l,t}, n^G_{h,t}, n^F_t),$$

where $n^G_{i,t}$ is the number of class i customers at the GK, $i \in \{l, h\}$, and n^F_t is the number of customers currently in the FSU. By the capacity constraints, $n^G_{i,t} \leq Q_i$ for $i \in \{l, h\}$ and $n^F_t \leq B$. For each customer currently at the GK, the decision-maker chooses one of the following three destinations: the FSU, the SSU, or keeping them waiting at the GK for one more period. Since the customers are treated as identical within each class, the system'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 customers sent to the FSU and SSU of class $i \in \{l, h\}$, respectively. Due to the capacity constraint in the FSU, the admissible action set for system state s_t is

$$\Pi(s_t) = \left\{(a_{l,t}, r_{l,t}, a_{h,t}, r_{h,t}) : a_{l,t} + r_{l,t} \leq n^G_{l,t},\ a_{h,t} + r_{h,t} \leq n^G_{h,t},\ a_{l,t} + a_{h,t} \leq B - n^F_t\right\}. \tag{1}$$

$\Pi(s_t)$ specifies the following set of constraints: The first two constraints state that the total number of customers sent to the FSU and SSU must be smaller than or equal to the total number of customers currently at the GK. The last constraint requires that the total number of customers sent to the FSU must be smaller than or equal to the current number of available servers in the FSU, i.e., $B - n^F_t$.

Let $u_{a,l}$ and $u_{a,h}$ be the system's expected utilities of sending a low-priority class and a high-priority class customer to the FSU, respectively. In addition, denote the expected utilities of sending each low and

[2] This is a reasonable assumption in many empirical settings, e.g., the ICU admission problem studied in Section 3. In addition, our model can be extended to include the capacity constraint of the SSU.

high-priority customer to the SSU by $u_{r,l}$ and $u_{r,h}$, and the expected utilities of keeping each low and high-priority customer waiting at the GK by $u_{w,l}$ and $u_{w,h}$, respectively. These utility parameters capture the expected utilities of each action for customers in each class. They incorporate all relevant factors that may affect the decision, such as customer demand, financial profit, and operational constraints. Such setting is standard in dynamic discrete choice models.

The system's expected utility $u(s_t, d_t)$ associated with state s_t and action d_t in period t is given by

$$u(s_t, d_t) = u_{a,l}a_{l,t} + u_{r,l}r_{l,t} + u_{w,l}\left(n^G_{l,t} - a_{l,t} - r_{l,t}\right) + u_{a,h}a_{h,t} + u_{r,h}r_{h,t} + u_{w,h}\left(n^G_{h,t} - a_{h,t} - r_{h,t}\right), \quad (2)$$

which is the sum of FSU routing, SSU routing, and waiting utilities for the two classes of customers. The system's total per period utility can be written as

$$U(s_t, d_t, \varepsilon_t) = u(s_t, d_t) + \varepsilon_t(d_t). \quad (3)$$

In above, $\varepsilon_t(d_t)$ is the idiosyncratic utility component associated with action d_t, which is observed by the system when making the decision, but not the researcher. For example, $\varepsilon_t(d_t)$ can capture the potential heterogeneity of customer conditions within a class. The additively separable form (3) follows Rust (1987) and the structural estimation of dynamic discrete choice literature.

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

$$\sup_{d_t \in \Pi(s_t)} \mathsf{E}\left\{\sum_{j=t}^{\infty} \beta^{j-t} U(s_j, d_j, \varepsilon_j) | s_t, \varepsilon_t\right\}. \quad (4)$$

The discount factor $\beta \in (0, 1)$ captures the trade-off between the system's perceived utility of current customers and future customers, which is the focus of our study. The expectation is taken over both the future random component ε_j and the transition of the system state–i.e., the arrivals, departures, and routing of customers in each period. In addition, note that the expectation in (4) is conditional on both s_t and ε_t, as the random component is observable to the system before making a decision in period t. As is common in the structural estimation literature, the optimization problem here models the system's decision making process under the observed flows of customers. Thus, the utility parameters and discount factor should be perceived as behavioral parameters that capture how the system balances complex and partially conflicting goals (e.g. mortality risk, reimbursement amounts, etc.) according to the data.

After the system chooses an action in period t, the number of GK customers of class i becomes $n^G_{i,t} - a_{i,t} - r_{i,t}$, and the number of customers in the FSU is $n^F_t + a_{l,t} + a_{h,t}$. To solve the model, we define an "intermediary" state, $\varphi(s_t, d_t)$, right after the action d_t is taken, i.e.,

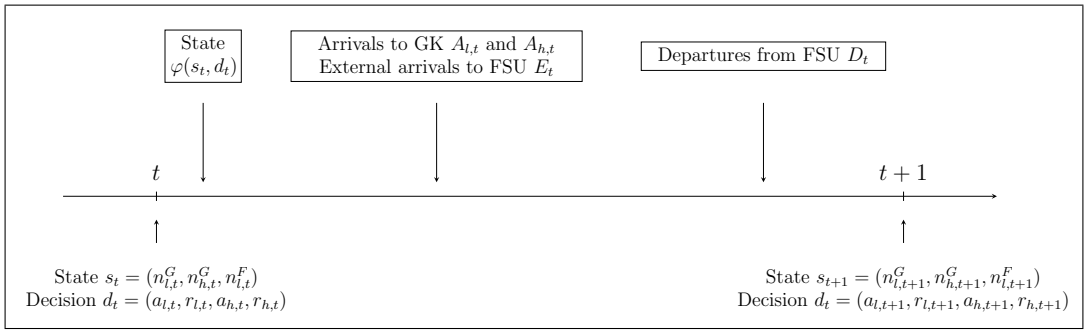
$$\varphi(s_t, d_t) = \left(n^G_{l,t} - a_{l,t} - r_{l,t}, n^G_{h,t} - a_{h,t} - r_{h,t}, n^F_t + a_{l,t} + a_{h,t}\right), \quad (5)$$

which describes the impact of action d_t on the system state. The system state then evolves according to the following two steps. First, $A_{i,t}$ new customers of class i arrive to the GK, and E_t customers arrive to the FSU through other channels–i.e., the external arrivals. If the GK or FSU is full, new arrivals can not be accepted. Thus, the total accepted GK and FSU arrivals are given by $A_{i,t}^{acc} = \min\left\{A_{i,t}, Q_i - (n_{i,t}^G - a_{i,t} - r_{i,t})\right\}$ for $i \in \{l,h\}$ and $E_t^{acc} = \min\left\{E_t, B - (n_t^F + a_{l,t} + a_{h,t})\right\}$, respectively. Second, D_t customers leave the FSU as they complete their first-tier service. This completes the system transition for period t. The system state at the beginning of period $t+1$ is thus given by $s_{t+1} = (n_{l,t+1}^G, n_{h,t+1}^G, n_{t+1}^F)$, with

$$n_{i,t+1}^G = n_{i,t}^G - a_{i,t} - r_{i,t} + A_{i,t}^{acc} \text{ and } n_{t+1}^F = n_t^F + a_{l,t} + a_{h,t} + E_t^{acc} - D_t. \tag{6}$$

It is clear that, by construction, 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 the system transition is summarized in Figure 2.

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



Both the current period's utility and how an action impacts the system state and future utilities may influence the system's decision in the current period. Define the value function as the objective in (4) given the optimal action sequence, i.e.,

$$V(s_t, \varepsilon_t) = \sup_{d_t \in \Pi(s_t)} \mathsf{E}\left\{\sum_{j=t}^{\infty} \beta^{j-t} U(s_j, d_j, \varepsilon_j) | s_t, \varepsilon_t\right\}. \tag{7}$$

By (4), the system chooses the action that maximizes

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

The last term on 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 equation

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

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

The above Bellman 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 (CI) assumption as in Rust (1987).

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

$$\Pr(s_{t+1}, \varepsilon_{t+1} | s_t, \varepsilon_t, d_t) = q(\varepsilon_{t+1}|s_{t+1}) g(s_{t+1}|\varphi(s_t, d_t)), \tag{8}$$

where the transition probability $g(s_{t+1}|\varphi(s_t, d_t))$ captures the random arrivals and departures shown by (6). 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 $\{s_t\}$. In addition, following the literature, we assume that ε_t is independent and identically distributed (i.i.d.) and follows a 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:*

$$f(d_t|s_t) = \frac{\exp\left(u(s_t, d_t) + \beta \tilde{V}(\varphi(s_t, d_t))\right)}{\sum_{d \in \Pi(s_t)} \exp\left(u(s_t, d) + \beta \tilde{V}(\varphi(s_t, d))\right)}, \tag{9}$$

where function $\varphi(s_t, d_t)$ *is given by (5). The function* $\tilde{V}(s)$ *is defined as*

$$\tilde{V}(s) = \sum_{s'} \int_{\varepsilon'} V(s', \varepsilon')\, g(s'|s) q(\varepsilon'|s') d\varepsilon'. \tag{10}$$

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 Section EC.1.1. The function* $\tilde{V}(s)$ *is the unique fixed point to the following functional equation*

$$\tilde{V}(s) = \sum_{s'} \ln \left\{ \sum_{d' \in \Pi(s')} \exp\left(u\left(s', d'\right) + \beta \tilde{V}(\varphi(s', d'))\right) \right\} g(s'|s). \tag{11}$$

As in most dynamic discrete choice models, the choice probability (9) has a closed-form representation, and the value function $\tilde{V}(s)$ solves the functional equation (11). By (10), the new value function $\tilde{V}(s)$ represents the expected future utility given s, the intermediary state after the action has been taken. In the proposition below, we show that it is monotonically non-increasing in the number of customers in the FSU.

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

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

That is, for any given number of customers at the GK, the function $\tilde{V}(s)$ *is monotonically non-increasing in the number of FSU customers.*

PROOF: See Section EC.2.1. □

This proposition shows that, as the system routes more customers to the capacity-constrained FSU, the future expected utility decreases. Thus, when routing customers, the system needs to balance the impact on current customers as well as its capacity to serve future customers.

2.2. Identification of Discount Factor

In this study, we aim to identify the utility parameters and discount factor in the model *jointly* from data, which is crucial for understanding how much the system reacts to the intertemporal externalities of its decisions. It is well-established in the literature that identifying the discount factor and utility parameters jointly is very difficult (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 (e.g., 0.975) and estimate the utility parameters only. However, the pre-specified discount factors usually lack empirical support and economic justifications. Indeed, the implied discount rate can vary substantially across different settings (Frederick et al. 2002).

The reason that the discount factor and utility parameters cannot be jointly identified in general is because multiple combinations of discount factor and utility parameters might lead to the same choice probabilities in (9) for all states and actions, i.e., they are observationally equivalent. Thus, an agent's observed actions in the data can be rationalized by different choices of discount factors and utility parameter values. Without further investigation of the model and the variation in the data, we cannot differentiate between such cases or identify the discount factor.

To identify the discount factor in our model, we leverage the recent theoretical development in Komarova et al. (2018). As highlighted in their paper, there is no prior identification results involving the discount factor for general parametric dynamic discrete choice models. The main steps for identification in Komarova et al. (2018) proceed as follows. Following the literature, they consider a dynamic discrete choice model where the utility function is linear in the unknown parameters. The choice set and transition probabilities are nonparametrically identified. For a given value of the discount factor β, they first construct estimates for the utility parameters following the standard two-step estimation procedure pioneered by Hotz and Miller (1993). By doing this, 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 estimated utility parameters, that minimizes the objective function, the model (discount factor and utility parameters) can be identified under some rank condition.

We extend and apply the identification results in Komarova et al. (2018) to our setting. By construction, 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. Besides, by (2) and (3), it is also clear that the deterministic part of the per-period utility $u(s_t, d_t)$ is linear in the utility parameters $\{u_{a,\iota}, u_{w,\iota}, u_{r,\iota}\}$ for $\iota \in \{l, h\}$. Thus, the linear-in-parameter assumption is satisfied in our setting. Details of the assumptions in Komarova et al. (2018) and how they apply to our setting are discussed in Section EC.1.2. An additional challenge in identifying discount factor in our structural model is due to the fact that our model has state-dependent action sets while the identification results in Komarova et al. (2018) are developed under the setting where all states have the same admissible action set. Consequently, the denominator for the choice probability in (9) is also state-dependent. However, note that for each state s_t, the admissible action set can

be fully determined by (1). Thus, the identification results in Komarova et al. (2018) can be extended to our model, given that we now plug in the admissible action set for each state according to (1).

We construct a one-dimensional criterion similar to the one in Komarova et al. (2018) based on the maximum likelihood estimator in Rust (1987). For each candidate β, we estimate the utility 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 utility parameter estimates) that maximizes the likelihood, and that ensures a rank condition is satisfied (see Theorem 1 of Komarova et al. 2018). We provide more details of the algorithmic approach in the next section. The intuition for the identification is that, if the variation in the data is rich enough, we can construct appropriate state and action pairs to separately identify the discount factor and the utility parameters of the model. For example, consider two states $s_1 = (1,0,0)$ and $s_2 = (0,1,0)$, i.e., with one low- and high-priority customer in GK respectively. Then, we can show by (9) that the difference in the SSU routing utilities can be identified as:

$$u_{r,l} - u_{r,h} = \ln\left(\frac{\Pr(r|s_1)}{\Pr(a|s_1)}\right) - \ln\left(\frac{\Pr(r|s_2)}{\Pr(a|s_2)}\right),$$

where $\Pr(a|s_1)$, $\Pr(r|s_1)$, $\Pr(a|s_2)$, and $\Pr(r|s_2)$ denote the probabilities of sending the customer to the FSU and SSU under the two states, respectively.

It is well known that to identify a dynamic discrete choice model, the utility of one action needs to be pre-specified in all states (Rust 1994, Magnac and Thesmar 2002). This is required even if the discount factor is assumed and predetermined in the model (Komarova et al. 2018). The common practice is to normalize the utility of one action to zero and interpret the estimated utilities of other actions as the differences in the utilities to the normalized choice. For general dynamic choice models, the normalization choice may or may not affect the estimation and counterfactual results, depending on the model set-up and the parametric assumptions (see, e.g., Norets and Tang 2014). In our model, we normalize the utility of admitting a high-priority customer to the FSU by pre-specifying a value for $u_{a,h}$. The following proposition shows how the normalization choice impacts the identification of our structural model.

Proposition 3 *Assuming the GK has the same number of total customers $n^G_{l,t} + n^G_{h,t}$ at the beginning of each period, then the model identification, i.e., values of discount factor and differences in utility parameters, does not depend on the normalization level of FSU routing utility $u_{a,h}$.*

PROOF: See Section EC.2.2. □

Proposition 3 shows that when the number of total customers in GK is fixed in each period, the identified discount factor and utility differences are unaffected by the normalization choice of $u_{a,h}$. This extends the results in Komarova et al. (2018), which considers a constant action space. In our setting, the action space $\Pi(s_t)$ can change with time even if the total number of customers at GK is fixed. First, the numbers of low-

and high-priority customers ($n^G_{l,t}$ and $n^G_{h,t}$) may vary even if their sum is fixed. Second, the FSU occupancy can change with time and there may not be enough servers to send all customers to the FSU.

In our general model set-up, the number of customers at GK can vary with time due to the routing decisions and random arrivals. Then, the model identification, including the identification of the discount factor, will generally hinge on the normalization choice. This is because the future states of GK, and thus future system utilities, will depend on the action taken in the current period. As such, the normalization level affects the underlying trade-off between the utilities of current versus future customers, which may impact the model identification. In these cases, the normalization choice should be justified based on the relevant objectives and constraints specific to the empirical setting. We discuss the normalization choice in our empirical study of ICU admission in Section 3.3.

2.3. Estimation Procedure

We now describe how our dynamic discrete choice model is estimated using the standard choice data, where we observe the system state and routing decisions of each customer. Suppose we have the state-action sequences (s_t, d_t) from the system for $t = 1, 2, ..., T$. First, the arrival and departure rates, as well as the GK and FSU capacities, are estimated directly from data – outside of the structural model. We estimate the customer 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 GK for the two classes in each period. We estimate the external arrival rate λ_E to FSU using the average number of customers admitted to the FSU in each period via other channels, The departure probability μ_I can be estimated as the ratio of the total number of departures to the total number of periods a group of customers spend in the FSU. The FSU capacity B is set as the maximum number of customers that can be served simultaneously. The GK capacities Q_l (resp. Q_h) can be estimated as the maximum number of low-priority (resp. high-priority) customers waiting at the GK.

The other parameters — the discount factor, the utility of routing the low-priority customer to the FSU, and the utilities of routing customers to the SSU and keeping them waiting in the GK for the two classes — are estimated within the structural model. Denote these parameters by $\theta = \{\beta, u_{a,l}, u_{r,l}, u_{r,h}, u_{w,l}, u_{w,h}\}$. Given the observed states and actions, the likelihood for a fixed set of parameters, θ, is given by

$$l^f(\mathbf{s}, \mathbf{d}|\theta) = \prod_{t=1}^{T} f(d_t|s_t, \theta) g(s_{t+1}|\varphi(s_t, d_t)) = \underbrace{\prod_{t=1}^{T} f(d_t|s_t, \theta)}_{l^d(\mathbf{s},\mathbf{d}|\theta)} \times \underbrace{\prod_{t=1}^{T} g(s_{t+1}|\varphi(s_t, d_t))}_{l^s(\mathbf{s},\mathbf{d})}, \tag{12}$$

where $(\mathbf{s}, \mathbf{d})$ denotes the observed state and action sequences, i.e., $\{s_t, d_t\}$ for $t = 1, 2, ..., T$; $f(d_t|s_t, \theta)$ is the choice probability in (9) given parameter θ. The state transition probability $g(s_{t+1}|\varphi(s_t, d_t))$ is explicitly given in the Section EC.1.1. The likelihood l^f can be decomposed to two parts: $l^d(\mathbf{s}, \mathbf{d}|\theta) := \prod_{t=1}^{T} f(d_t|s_t, \theta)$ is the part associated with choice probabilities, and $l^s(\mathbf{s}, \mathbf{d}) := \prod_{t=1}^{T} g(s_{t+1}|\varphi(s_t, d_t))$ is

the part from state transitions. We see that the structural parameter θ is only involved in the first part l^d, but not the second part l^s which only depends on the arrival and departure rates.

We employ the nested fixed-point algorithm in Rust (1987) to estimate the utility parameters (conditioning on discount factor) by maximizing the likelihood of observed choices. 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 (12). This is equivalent to finding the value of θ that maximizes $\ln l^d$ in (12) as it is the only part in l^f that involves θ. In the "inner" loop, the unknown function $\tilde{V}$ is computed by value iteration on the functional equation (11). In the "outer" loop, we use a gradient descent algorithm to find the optimal parameter θ. To summarize, for a candidate value of $\beta \in (0,1)$, we estimate the utility parameters $\{u_{a,l}, u_{r,l}, u_{r,h}, u_{w,l}, u_{w,h}\}$ that maximizes the log-likelihood $\ln l^d$. Then, we choose the discount factor and its associated utility parameters that leads to the largest likelihood among all candidate β. Finally, we can also verify the rank condition in Komarova et al. (2018) holds by simulation using the estimated parameters.

To examine the proportion of variation explained by the estimated model, we compute the McFadden's pseudo R^2 as

$$\text{Pseudo } R^2 = 1 - \frac{\ln l^d(\hat{\theta})}{\ln l^{null}}, \tag{13}$$

where l^{null} is the "null" likelihood from a multinomial logistic regression model with only intercept terms, i.e., the action probabilities do not depend on system states.

3. Empirical Study: ICU Admission Decisions

In this section, we apply our structural model and the identification method to a suitable and important empirical setting in healthcare operations management. Specifically, we study the ICU admission decisions for patients in the emergency department (ED) in a large hospital network. We first describe the data and the clinical setting. Then, we set up the structural model and provide the estimation results.

3.1. Data and Clinical Setting

We utilize a large data set from 21 hospitals in a large hospital network in California, US. The data contains 312,306 hospitalizations over a period of two years before the COVID-19 pandemic. All patients are covered by the same insurance program[3] and receive care at one of the hospitals. The hospitals cover a large geographic area (average distance between hospitals is more than 50 miles) and intra-hospital transfers are quite rare ($< 3\%$). As such, we will generally treat each hospital independently.

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 multiple severity scores. In addition to the patient level information, we also observe the admission and discharge time for

[3] The hospital network is vertically integrated.

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 or surgical ward, operating room (OR), or the postanesthesia care unit (PAR).

In this study, we focus on the routing decisions for the patients admitted to a medical service via the ED. This patient cohort comprises the largest proportion of admitted ICU patients. In addition, unlike surgical patients who usually have scheduled arrivals with fixed care protocols, there is much variation in the admission decisions for the ED medical patients (e.g., Chen et al. 2012). The ICU admission decision for an ED patient is generally made as follows. After a patient is stabilized in the ED, the ED physician provides an initial assessment about the condition and needs of the patient. If the patient may require ICU admission, an intensivist will be called to the ED for a consultation. The ultimate decision regarding the patient's disposition requires the communication and coordination of many people including the ED physicians, the ICU intensivist, and hospital administrators, as well as various system-level factors such as nurse staffing availability, diversion policies, or alternative interventions available. However, due to data limitations, we have no information on the ED physician and intensivist on duty. Therefore, we refer to this composite decision maker as the "hospital" and study the average behavior of the composite decision maker within each hospital. We emphasize that our estimation results and the counterfactual analyses are consistent with such interpretation of the decision maker and the complex decision making process.

The final study cohort for the ED medical patients consists of 164,166 hospitalizations for 22 hospitals over a horizon of two years.[4] The detailed data selection process is described in Section EC.3.1. In this patient cohort, 19,683 (12.0%) are admitted to the ICU, and the remaining are admitted to a non-ICU unit. Table EC.4 summarizes the patient characteristics of the final study cohort and the subset of which are admitted to the ICU. We find that the admitted cohort has higher average severity scores than the full cohort. The ED boarding time is similar for the two cohorts.

While we focus on the admission decisions for the ED medical patients, we utilize the data from all hospitalizations to compute the external arrivals to ICU and the real-time occupancy level of the ICU in each hospital. Among all ICU admissions, around 36.2% are directly from our study cohort described above, while external arrivals account for 63.8%. Thus, it is crucial to include the external arrivals when calculating ICU occupancy.

3.2. Intertemporal Externalities in ICU Admissions

We aim to study how the hospitals balance the utilities of current ED patients and the impact on future system state when making admission decisions. As the ICU provides the highest level of care, swift admission generally benefits the patients who need ICU care. However, given the limited capacity and high occupancy

[4] For Hospital 21, we find that its ICU capacity experienced a substantial change during the sample period. Thus, we split it into two parts, i.e., before and after the capacity change, and treat them as two hospitals in the estimation.

of ICU, this may restrict access to ICU care for future, perhaps more severe, patients. Thus, the admission decision introduces intertemporal externalities via the system's ability to treat future patients. This fits into our general framework in Section 2.

The hospital's balancing behavior can be reflected by how its ICU admission decisions vary with ICU occupancy levels. If a hospital primarily focuses on the current ED patients when making decisions, we would expect the ICU admission probability to be insensitive to ICU occupancy. On the other hand, if a hospital is relatively concerned with the system's ability to treat future patients, the ICU admission probability would drop when the ICU gets more congested, as the hospital is likely to reserve the limited ICU capacity for future patients (e.g., by stabilizing less severe patients in ED or wards). A priori, it is not clear how hospitals actually balance this trade-off when admitting patients.

Before setting up our structural model for ICU admission decisions, we first provide some descriptive evidence on hospitals' balancing behaviors between current patient and future system capacity. In particular, we develop a multinomial logistic model to estimate the ICU admission decisions. The model controls for a variety of factors that may affect the admission decisions, including patient characteristics, system states, as well as hospital and time fixed effects. The details of the multinomial logistic model and the estimated coefficients are provided in Section EC.4.

From the regression, we find that a higher ICU occupancy is associated with lower ICU admission probability for most hospitals. That is, the hospitals tend to slow their ICU admission when the ICU is more congested. We also find the effects of ICU congestion vary substantially across hospitals. This reflects that, at least to some extent, the hospitals take the impact on future system state into consideration when making admission decisions. In subsequent sections, we use the structural model developed in Section 2.1 to understand the discounting behaviors of hospitals. Importantly, the structural model allows us to conduct counterfactual analyses to quantify the impact of potential shocks on system performance metrics, which is not possible via regression analysis.

3.3. Model Set-up for ICU Admission Decisions

In this section, we apply the general structural model in Section 2 to the hospital's ICU admission process. For each patient currently in the ED, the hospital can make three decisions: admit the patient to the ICU, admit the patient to non-ICU units (e.g., mainly wards), or keep the patient waiting in the ED. In light of the admission process, we view the ED as the gatekeeper in our model, which sends patients to ICU or non-ICU units. The ICU is seen as the first-tier service unit (FSU), as it provides the highest level of care for the patients but has limited capacity. The non-ICU units provide a lower level of care and usually have ample capacity.[5] Thus, they are regarded as the second-tier service unit (SSU) in the model. We do not explicitly model the transfers from non-ICU units to ICU, but they are captured by external arrivals to ICU.

[5] This is a reasonable assumption as the proportion of periods where the ward occupancy exceeds 95% of its capacity is less than 0.8% in our sample.

We define each period to be a two-hour time interval, which provides a reasonable amount of time for transferring the patient from one unit to the next after an admission request is issued. 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. The LAPS2 score is assigned at hospital admission based on labs and vital signs taken in the last 72 hours (see Escobar et al. 2013). A score of 110 is considered to be critically ill. We define the low severity class as patients with LAPS2 score below the 85th percentile of the sample distribution (113), and the high severity class as those above the 85th percentile. That is, all medical patients admitted via the ED are included in our structural model as one of the two classes. The patients in the high severity class on average are more likely to require ICU service. Thus, the two severity classes correspond to the high-priority and low-priority customers in our general set-up. In Section EC.5.2, we perform a robustness check using three patient classes, and our main empirical findings remain the same.

We assume the hospital's utility maximization problem is formulated as (4), with the per period utility given by (3). Recall the utility parameters and discount factor in the model are perceived as behavioral parameters for understanding the hospital's actions from observed data. They incorporate all relevant factors that may affect the hospital's admission decisions, such as patients' short term and long term clinical outcomes, operational costs, financial revenue, system-level constraints, etc. In the structural model literature, it is common to define the utility function for modeling the decision-making process without explicitly delineate each of its determinants (see, e.g., Olivares et al. 2008, Wang et al. 2019, Rath and Rajaram 2022).

Our model includes six utility parameters for the low and high severity classes: the average utilities of admitting the patient to ICU ($u_{a,l}$ and $u_{a,h}$), admitting the patient to non-ICU units ($u_{r,l}$ and $u_{r,h}$), and keeping the patient waiting in ED ($u_{w,l}$ and $u_{w,h}$). As discussed in Section 2.3, we need to normalize the utility of one action in order to identify the model. In our setting, we choose to normalize the utility of ICU admission for a high severity patient to zero, i.e., $u_{a,h} = 0$, which is common in the literature. Then, in principle, the ICU admission utility of low severity class $u_{a,l}$ can be identified by the observed ICU admission probabilities when there are both low and high severity patients in ED. However, we find that such states are relatively infrequent in our sample. So the data does not allow us to accurately estimate the ICU admission utility for the low severity class separately from the high severity one. Thus, we additionally assume the ICU admission utility to be the same for the high and low severity classes in our main specification, i.e., $u_{a,l} = u_{a,h} = 0$. As such, all other utility parameters are assessed *relative* to ICU admission.

In our setting, assuming zero ICU admission utilities is equivalent to setting a reference point of utility across ED states: admitting all ED patients to ICU leads to the same average per-period utility (zero) regardless of the ED state. We expect this to be a reasonable assumption. First, admitting a patient to ICU is the best the hospital can do for the patient's clinical outcomes. Given improving patients' clinical outcomes is the hospital's primary goal,[6] admitting all ED patients to ICU should lead to the same per-period utility

[6] This is especially true for a vertically integrated hospital with capitated payments as in our study.

perceived by the hospital regardless of how many patients are in the ED. Second, ICUs have very high fixed and operating costs, thus the marginal utility of admitting one patient should be small. For example, ICUs have extremely high set-up cost for specialized equipment and tend to staff their beds based on fixed nurse-to-bed ratios. All these fixed set-ups do not affect the hospital's (marginal) admission decisions. We note that it is impossible to pin down the normalization level of ICU admission utilities from data, and some assumption has to be made. Moreover, setting a zero utility for normalized action is common in the literature. In Section EC.5.2, we show that our main findings are robust to the normalization choice.

For the other utility parameters, we assume that keeping patients waiting in the ED brings the hospital negative utilities for both classes, i.e., $u_{w,l}, u_{w,h} < 0$. This is a natural assumption since longer ED boarding time is generally undesirable, leading to a disutility for the hospital.[7] Longer ED waiting time is shown to be associated with negative outcomes, such as increased mortality risk and hospital LOS (Singer et al. 2011). Thus, reducing ED boarding time and ED crowding has been a common goal in the medical community. On the other hand, we do not restrict the sign of the non-ICU admission utilities $u_{r,l}$ and $u_{r,h}$. Patients who are not critically ill can receive sufficient care in the ward, so admitting them to the ward can bring the hospital positive utilities. For example, even for the high severity class, more than half of the patients (66%) are eventually admitted to the ward instead of ICU.

4. Main Empirical Results

This section presents our main empirical results on how hospitals balance the trade-off between the utilities of current patients and the impact on system state when making ICU admission decisions.

4.1. Estimation Steps for ICU Admission Decisions

We estimate the structural model for ICU admission decisions in two ways. First, we estimate the model for all hospitals combined. That is, we estimate one set of parameters that maximizes the sum of log-likelihood from all hospitals. Second, we estimate the model for each hospital individually. We choose to estimate the model in both ways for the following reasons. In the combined estimation, we can obtain more reliable estimates using a much larger sample. However, it does not provide information on how the discounting behavior varies across hospitals. For this purpose, we also estimate the model for each hospital individually to reveal potential heterogeneity across hospitals.

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 distinguish how the hospital balances the current patient versus future system capacity when making decisions. We conduct a robustness check with a finer grid for select hospitals, as will be discussed in the next section. We use the bootstrap method to get the standard errors of the estimates. For each hospital, we resample its data for 500

[7] Note that the decision to keep the patient waiting in the ED is mostly necessitated by system-level considerations; in the absence of capacity constraints, a patient would not be kept waiting for admission.

times by randomly selecting from its state-action pairs $\{s_t, d_t\}_{t=1}^{T}$ with replacement.[8] Each resampled set has the same number of periods as the original sample. We then estimate the model parameters for each of the 500 resampled sets to obtain the standard errors.

As described in Section 2.3, the system parameters (e.g., arrivals, capacities) are estimated outside the structural model. We provide the details of their estimation for our ICU admission problem in Section EC.3.2. The results are reported in Table EC.5, including the numbers of days and hospitalizations in our sample, ED and ICU capacities, average ICU occupancy, arrival and departure rates, and the proportion of patients that are eventually admitted to the ICU for the two classes for each hospital.

We find that the 22 hospitals in our study have very different sizes, work loads, and admission behaviors. For example, the ICU capacity varies from 7 to 36 beds across hospitals. The ICU admission probabilities also vary substantially across hospitals even for the same severity class. In addition, the ICUs in the hospitals are generally congested. The average ICU occupancy in most hospitals is higher than 50%, and, in some hospitals, higher than 70%. Finally, the ICU admission probability for the high severity class is above 30% for most hospitals, which is usually three or four times larger than that for the low severity class (around 8%). This implies that the admission decisions are very different for the two severity classes, which is captured in our structural model. However, with a much larger population, the low severity class contributes more (58%) to the total ICU admissions of ED medical patients than the high severity class (42%). Thus, it is important to include both severity classes in our model.

4.2. Estimated Discount Factor and Action Utilities

In what follows, we provide the main estimation results of the structural model, i.e., the estimated discount factor and the utility parameters. The estimated structural model reveals how the hospitals balance the utility of current patient and impact on system state in their observed admission decisions.

Table 1 reports the estimation results for all hospitals combined. The McFadden's pseudo R^2 of the structural model is 0.14, which is comparable to that from a comprehensive multinomial logistic regression, as will be discussed in the next section. The estimated discount factor is $\hat{\beta} = 0.3$, and is statistically significantly different from adjacent levels. In addition, we show in Figure EC.2 that the likelihood function decreases monotonically as we deviate away from the estimated $\hat{\beta} = 0.3$, which further validates the identification and estimation results.

At first glance, the estimated β is quite surprising. In most empirical studies with dynamic choice models, the discount factor is assumed to be relatively large, e.g., 0.95 or 0.99. However, we see that the estimated $\hat{\beta}$ is much smaller than these levels. In our model, the relatively small value of $\hat{\beta}$ implies that the hospitals are *not* very "forward-looking" when making ICU admission decisions. Given each period in our model is a two-hour interval, it suggests that the hospitals mainly focus on the current patients and barely consider the

[8] The log-likelihood depends on the model parameters only via the choice probabilities, which can be calculated explicitly for a given state-action pair (s_t, d_t) by (9).

Table 1 Estimation results of structural model: All hospitals combined (N = 154,140 hospital-periods)

Discount factor	Low Severity		High Severity		
$\hat{\beta}$	$\hat{u}_{w,l}$	$\hat{u}_{r,l}$	$\hat{u}_{w,h}$	$\hat{u}_{r,h}$	R^2
0.3	-0.071	1.950	-0.932	0.671	0.14
(0.005)	(0.001)	(0.010)	(0.018)	(0.013)	

Note: Standard error is reported in parenthesis.

impact of their decisions on the system beyond the next six hours (after three 2-hour periods, $0.3^3 \approx 0.03$). The above finding highlight the importance of identifying discount factor using real data instead of assuming a pre-specified value. In Section 5, we use counterfactual simulations to show that using a "wrong" discount factor in the structural model can lead to large bias when evaluating the impacts of different system changes.

Recall that the utility parameters represent the average perceived utilities measured relative to the ICU admission decision, which is normalized to have zero utilities. By Table 1, the non-ICU admission utilities $\hat{u}_{r,l}$ and $\hat{u}_{r,h}$ are positive and significant for both the low and high severity classes (1.950 and 0.671). This suggests that it is, on average, desirable to admit a patient to the ward units. This is because for the majority of patients, the ICU care is expensive and (likely) unnecessary, which can be seen by Table EC.5. On the other hand, the ED waiting utilities $\hat{u}_{r,l}$ and $\hat{u}_{r,h}$ are negative and significant for both classes (-0.071 and -0.932), which is inline with our assumption. Finally, both the waiting and non-ICU admission utilities are significantly lower for the high severity class than for the low severity class. Thus, it is on average more undesirable to keep the high severity patients temporarily waiting in the ED or to send them to the ward units, reflecting that the high severity patients are more likely to require ICU care. The large differences in the action utilities highlight the importance of differentiating the two severity classes in the model.

Next, we estimate the structural model for each hospital individually. To save space, we report the estimated discount factor and utility parameters, their standard errors estimated by bootstrapping, as well as the McFadden's R-squared for each hospital in Table EC.6 of the Electronic Companion. The results show substantial heterogeneity in the estimated discount factors across hospitals. In particular, 13 out of the 22 hospitals have relatively small estimated discount factors $\hat{\beta} \in \{0.1, 0.2, 0.3\}$, five have medium discount factors $\hat{\beta} \in \{0.4, 0.5, 0.6\}$, and the other four have relatively large discount factors $\hat{\beta} \in \{0.7, 0.8, 0.9\}$. The estimation results show that the hospitals behave very differently when making ICU admission decisions. Some of the hospitals have relatively small discount factors, suggesting that they focus primarily on the current ED patients. The others have relatively large discount factors, i.e., they account more for the impact of their admission decisions on the system's ability to treat future patients. Due to the smaller sample size, the standard errors of the estimates are larger than those in Table 1 for the combined estimation. That said, we find the estimated $\hat{\beta}$'s are generally accurate for individual hospitals (see the discussion in the notes of Table EC.6). Figure EC.3 shows how the likelihood varies with discount factor for select hospitals.

In addition, we find the patterns of the utility estimates for individual hospitals are very similar to those for all hospitals combined. As shown in Table EC.6, most of the waiting utilities are negative and significant. This holds for 17 out of the 22 hospitals for the low severity class ($\hat{u}_{w,l}$) and all hospitals for the high severity class ($\hat{u}_{w,h}$). The non-ICU admission utilities are all positive and statistically significant, except for the high-severity class ($\hat{u}_{r,h}$) in Hospital 11 and 22. In most cases, the non-ICU admission and waiting utilities for the high severity class are significantly lower than those for the low severity class.

We further investigate the heterogeneity in the estimated discount factor $\hat{\beta}$ by checking how it correlates with the observed system statistics of the hospitals. We find that hospitals with busier ICUs tend to have larger discount factors: across the 22 hospitals, the estimated $\hat{\beta}$ has a positive correlation with average ICU occupancy level (0.445), and a negative correlation with the ICU departure rate (-0.428) and the number of remaining beds at the 5% busiest time of ICU (-0.452). These correlations are statistically significant at the 5% level. This pattern can be explained as hospitals with busier ICUs have more motivation and need to account for the system-level impact when making admission decisions for ED patients.

Our empirical results show that both current patient and system considerations can influence the hospitals' admission decisions, and the overall effect varies substantially across hospitals. This reveals a novel aspect of the practice variation observed in medical literature, which has received much attention from medical professionals (see a review in Corallo et al. 2014). We note that the hospital-level results, as discussed above, tend to underestimate the actual variation in the system's behaviors, as it ignores the potential heterogeneity within the same hospital (e.g. night versus day, weekday versus weekend etc.). Thus, the significant heterogeneity across hospitals implies that there is likely even larger variation in practice.

4.3. Goodness of Fit and Robustness Checks

We briefly discuss the goodness of fit of the estimated structural model and several robustness checks for our main findings. The detailed results are provided in Section EC.5 of the Electronic Companion.

We show that our estimated structural model has good explanatory power for the hospitals' decisions and produces system statistics close to the data. First, the structural model leads to comparable McFadden's pseudo R^2 to the comprehensive multinomial logistic model described in Section EC.4 for all hospitals combined and for most individual hospitals. In addition, the heterogeneity in hospitals' discount factors is consistent with the estimated effects of ICU occupancy from the descriptive analysis. We also show, by simulation, that the key system statistics produced by the structural model are close to those observed in the data for each hospital, including overall ICU admission probabilities, average ICU occupancy, and average ED waiting time for the two severity classes. Finally, we estimate the structural model using the first half of our sample only and show it has good out-of-sample performance.

We conduct several robustness checks for our main findings. As discussed in Section 3.3, we normalize the ICU admission utilities of both severity class to zero in our main specification. We perform two robustness checks for the normalization choice. First, we estimate the model for all hospitals combined using the

sample with fixed ED states. Similar to the proof for Proposition 3, we can show that the identification is not affected by the normalization choice of $u_{a,l}$ and $u_{a,h}$ in a model with constant ED state $(n^G_{l,t}, n^G_{h,t})$. Second, we estimate the model for each hospital separately under small but non-zero ICU admission utilities. Our main results are robust in both settings. Next, we estimate the model with stratified samples based on flu vs. non-flu seasons as well as day and night periods. This addresses potential seasonality issues in our identification results. Finally, we estimate the model with three patient classes and a finer grid for discount factor for select hospitals. The results from these robustness checks are largely consistent with those in our main specification.

5. Counterfactual Studies

One of the main advantages of the structural estimation approach is its ability to conduct counterfactual studies. This allows us to quantify the impact of system changes on key system performance metrics, which cannot be obtained by descriptive or regression analysis. We focus on metrics for high ICU congestion in our counterfactual studies. For hospitals, managing ICU congestion is essential for fulfilling their clinical and managerial goals. A congested ICU is shown to be associated with worse clinical outcomes, such as higher mortality, longer hospital length-of-stay, and higher risk from complications (e.g., Hugonnet et al. 2007 and Kim et al. 2015).

We first consider the probability of high ICU congestion, which is defined as

$$\Pr(\text{HighCgstn}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left\{n_t^F \geq B-1\right\}; \tag{14}$$

i.e, the proportion of periods with only one or no empty bed(s) in the ICU. For all hospitals in our sample, operating above the 95% occupancy level implies that there is at most one empty bed. We can then estimate the number of ICU patients who spend their ICU stay under highly congested states by

$$\text{Pats HighCgstn} = \Pr(\text{HighCgstn}) \times 365 \times 12 \times \mu_I \times (B-1). \tag{15}$$

Here the product $\Pr(\text{HighCgstn}) \times 365 \times 12$ denotes the total number of high congestion periods in a year, given each period is two hours. We then divide this by the average LOS of each patient, $1/\mu_I$, to estimate the number of patients each bed can serve during the high congestion periods. Since there is at most one bed available during the high congestion periods, (15) provides a conservative estimate for the number of ICU patients who are exposed to the high congestion periods in a year.

In addition, we examine the probability that an external arrival patient will balk upon arrival because there is no ICU bed available. This is given by

$$\Pr(\text{Balk}|\text{External Arrival}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left\{n_t^F = B\right\}. \tag{16}$$

Using a similar argument as for (15), the number of external arrivals who balk in a year can be estimated by

$$\text{Pats Balk} = \Pr(\text{Balk}|\text{External Arrival}) \times 365 \times 12 \times \lambda_E, \tag{17}$$

where λ_E is the external arrival rate in each two-hour period.

In the counterfactual studies below, we use the parameters estimated for each hospital individually, which are reported in Table EC.6 of the Electronic Companion. The main findings are similar when we use the estimates from all hospitals combined in Table 1.

5.1. Intertemporal Trade-off Captured by Discount Factor

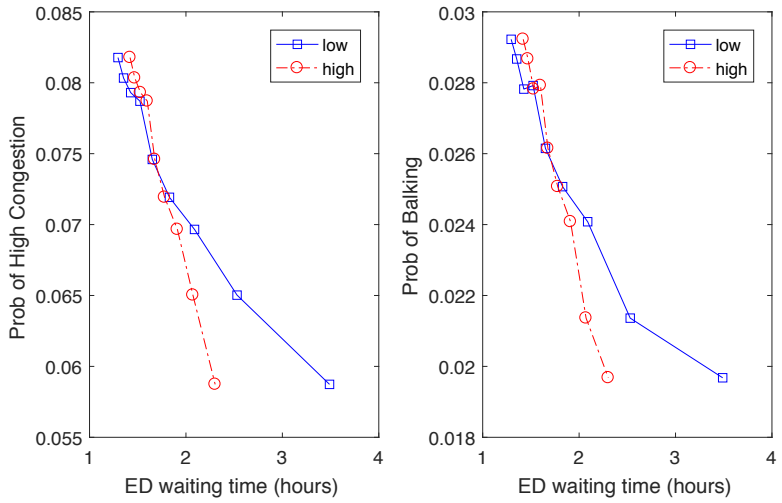
We first use counterfactual simulations to understand the intertemporal trade-off captured by the discount factor in our estimated structural model, which is the focus of this paper. For this purpose, we vary the discount factor in the model while keeping the utility parameters unchanged at their estimated levels. Then, we estimate the system performance metrics using the averages of 200 simulation trials: each trial has the same number of periods as in the data preceded by a three-month warm up period.

With a larger discount factor, the hospital accounts more for the impact of their admission decisions on the system state and the ability to care for future patients. As discussed in the previous section, hospitals with a larger discount factor tend to slow down their ICU admission when the ICU is already congested. Such adaptive behavior is expected to mitigate the occurrence of high ICU congestion states. However, it can also increase the ED waiting time of patients as they may be kept waiting in the ED for more periods. This introduces a potential trade-off between ED waiting time and high ICU congestion, which reflects the impact of discount factor in the model.

We use Hospital 2 as an example to illustrate the above trade-off. In Figure 3, we plot the probabilities of high ICU congestion and balking in (14) and (16) (y-axis) versus the average ED waiting time for the two patient classes (x-axis). The points on each line from left to right represent the estimates with $\beta = 0.1, 0.2, \ldots, 0.9$ respectively. We see that as we use a larger discount factor in the model, the probabilities of high ICU congestion and balking drop, while the average ED waiting time increases for both classes. This reveals the trade-off introduced by changing the discount factor in the model. However, we notice that the impact on ED waiting time is disproportionate for the two classes. As the discount factor becomes larger, the increase in ED waiting time is much larger for the low-severity class than for the high-severity one. When $\beta = 0.9$ (the rightmost points), the two classes are distinguished in terms of their ED waiting time. This is because the patients in the high-severity class are less likely to be delayed in the ED. The smaller impact on the high-severity class mitigates the undesired outcome from the longer ED waiting time.

We further quantify the impact of changing discount factor by the following counterfactual study. In particular, we increase the discount factor in the structural model to 0.9 for each hospital, while keeping

Figure 3 **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).

their utility parameters unchanged.[9] This study reveals how the system performance is affected if hospitals account more for the intertemporal externalities of their admission decisions. Table 2 compares the simulated system statistics under the structural models with the identified $\beta = \hat{\beta}$ and the new $\beta = 0.9$. The first three columns show the overall ICU admission probabilities for the two severity classes $\iota \in \{l, h\}$, average ICU occupancy, and ED waiting time for the two classes. These five statistics are averaged across the 22 hospitals for ease of interpretation. The last two columns report the total numbers of patients under high congestion and external arrivals who balk in a year, which are calculated by (15) and (17). To evaluate the overall impact, we sum up the number of patients from the two measures over all hospitals. The numbers in the parenthesis show the relative reduction from the base case with $\beta = \hat{\beta}$.

Table 2 **Simulated system statistics from structural models with $\beta = \hat{\beta}$ and $\beta = 0.9$**

Model	Average of hospitals			Sum of hospitals	
	$\Pr(\text{AdmICU}_l)$ / $\Pr(\text{AdmICU}_h)$	ICUOccu	EDWait_l / EDWait_h	Pats HighCgstn	Pats Balk
Structural: $\beta = \hat{\beta}$	8.55% / 32.66%	57.94%	1.33 / 1.39	2355	352
Structural: $\beta = 0.9$	8.06% / 32.51%	57.22%	2.63 / 1.76	2096 (-11.0%)	306 (-13.1%)

Note: This table reports the simulated statistics under the structural models with $\beta = \hat{\beta}$ and $\beta = 0.9$. Here we use the parameters estimated for each hospital individually.

We have the following observations for increasing the discount factor β. First, increasing β to 0.9 has little impact on the overall ICU admission probabilities and average ICU occupancy (first two columns).

[9] For the hospitals with identified $\hat{\beta} = 0.9$, this introduces no change in their models.

This can be explained as the utility parameters are kept unchanged in the intervention, thus the proportion of patients eventually admitted to ICU remains similar. However, increasing β leads to longer average ED waiting time for both classes (third column): the average ED waiting time for the low (resp. high) severity class increases from 1.33 to 2.63 (resp. 1.39 to 1.79) hours. We notice that the impact is smaller for the high-severity class, consistent with the pattern in Figure 3.

By the last two columns of Table 2, we see that increasing β to 0.9 substantially mitigates the high ICU congestion. Specifically, it reduces the patients under high ICU congestion for the 22 hospitals by 11% (from 2355 to 2096) in a year, and the external arrivals who balk by 13.1% (from 352 to 306). To assess the magnitude of the improvement, we compare the reductions in the two congestion measures from increasing β to 0.9 with those from adding one ICU bed and related staff. The latter represents a traditional operational intervention of increasing the ICU capacity B by one, which can be very costly to implement. Table 3 reports the comparison of the effects for select hospitals. It shows that for some hospitals, the reduction in high ICU congestion from increasing β to 0.9 is comparable to that from adding an ICU bed and related staff. Thus, understanding the balancing behavior between current patients versus system's capacity to serve future patients can be practically important for hospitals to manage their ICUs.

Table 3 Comparison of counterfactual effects from increasing β to 0.9 and adding one ICU bed (select hospitals)

Hosp	$\hat{\beta}$	ΔPats HighCgstn		ΔPats Balk	
		β to 0.9	$B+1$	β to 0.9	$B+1$
1	0.3	34.94	53.24	4.68	6.76
2	0.5	47.03	53.92	7.53	8.52
5	0.1	24.67	71.42	4.55	10.17
8	0.1	33.21	59.04	4.69	7.29
9	0.5	38.99	69.96	6.42	11.54
14	0.3	17.43	33.20	3.26	5.93

Note: Reductions in patients under high ICU congestion and balking from increasing β to 0.9 and adding one ICU bed.

The counterfactual study reveals the trade-off captured by the discount factor in our structural model, i.e., a larger discount reduces ICU congestion but increases ED waiting time. In light of this trade-off, the "proper" balancing behavior of current patients versus future system capacity would depend on the specific goals and constraints of hospitals, e.g., the patient cohort and available medical resources. In suitable circumstances, hospitals may reduce their ICU congestion by accounting more for the system-level impact of the admission decisions. Hospitals can take a set of approaches to facilitate such change, e.g., better educating their physicians about the benefit of being more responsive, improving the IT system to timely share the states of units, and providing more stabilising treatment in the ED for less urgent patients (e.g., Weingart

et al. 2013). This provides important managerial insights in ICU management by revealing the impact of smoothing workload over time and actively saving capacity during congestion periods. Such managerial insights may be applied to other real-world settings as well.

5.2. Impact of Misspecifying Discount Factor in Identification

The main contribution of this study is to identify and estimate the discount factor from data. In most related works in the literature, the discount factor is pre-specified, usually at a relatively high level (e.g., 0.99). In this section, we use counterfactual simulations to show how pre-specifying an "incorrect" discount factor leads to bias in evaluating the high ICU congestion periods.

To show the impact of misspecifying the discount factor in the structural model, we consider four choices of discount factors for each hospital: $\beta = 0$ for a pure myopic model, $\beta = 0.3$ as the average level from the combined estimation, $\beta = \hat{\beta}$ at the estimated level, and $\beta = 0.99$ as is commonly used in the literature. For a given discount factor, the utility parameters are estimated accordingly for each hospital. This mimics the common estimation procedure for dynamic discrete choice models: first pre-specifying a discount factor and then estimating the utility parameters from data. We then estimate the system performance metrics for each discount factor and its corresponding utility parameters.

We use counterfactual simulations to estimate the ICU congestion under three scenarios. The first is the base case with the estimated system set-up. We then consider two system changes. In the first one, we increase the ED arrival rates $\lambda_{Q,l}$ and $\lambda_{Q,h}$ by 50%.[10] Given the ED patients in our cohort consist of 36% of total ICU admissions, such an increase in ED arrivals translates to an 18% increase in ICU workload. In the second one, we keep the ED arrivals at the original level but reduce the ICU capacity by one bed, i.e., decreasing ICU capacity B by one. The two system changes reflect scenarios when ICU capacity gets constrained. For example, a pandemic outbreak may cause a surge in the ED arrivals; or hospitals may have to scale down their ICU due to financial constraints. Correctly evaluating the impact of such shocks is crucial to hospitals for managing their ICUs and related resources as well as meeting clinical and financial goals. We note that the two changes are large enough such that it is unreasonable to expect the local or marginal effects captured by regression analysis to make accurate projections on their impact.

Table 4 provides the estimates of total numbers of patients exposed to high ICU congestion and external arrivals who balk under different choices of discount factor (with corresponding utility estimates) and the three scenarios of system set-up. The results are obtained using the averages from 200 simulation trials described in previous section. The numbers in the table represent the sum of the 22 hospitals in our sample. The estimates for individual hospitals are qualitatively similar and available upon request. The top row "$\beta = \hat{\beta}$" shows the estimates when we jointly identify discount factor and utility parameters for each hospital.

[10] We indeed observe substantial variation in the ED arrival rates in our sample, which covers a two-year horizon before the COVID-19 pandemic. For example, the 90th percentile of daily (resp. weekly) ED arrival rates of all hospitals combined is 49% (resp. 22%) higher than the mean.

Table 4 Estimated ICU congestion metrics under difference choices of discount factors and corresponding utility estimates

Discount factor	Base case		Increasing ED arrival by 50%		Reducing one ICU bed	
	Pats HighCgstn	Pats Balk	Pats HighCgstn	Pats Balk	Pats HighCgstn	Pats Balk
$\beta = \hat{\beta}$	2,356	352	4,408	678	3308	546
$\beta = 0$	4,319 (83.4%)	718 (103.8%)	7,384 (67.5%)	1,282 (89.0%)	5,445 (64.6%)	995 (82.5%)
$\beta = 0.3$	2,961 (25.7%)	461 (30.9%)	5,335 (21.0%)	866 (27.6%)	3,993 (20.7%)	692 (26.8%)
$\beta = 0.99$	2,223 (-5.6%)	324 (-8.1%)	3,718 (-15.6%)	554 (-18.3%)	3,065 (-7.3%)	500 (-8.3%)

Thus, we use it as the benchmark for comparison. For other choices of β, their relative biases compared with $\beta = \hat{\beta}$ are reported in the parenthesis.

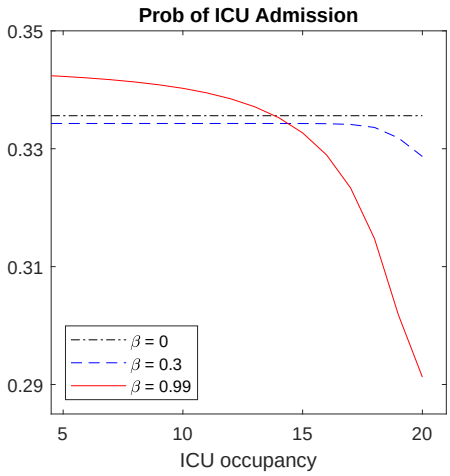
We have the following findings on the estimated ICU congestion under different discount factors. First, setting $\beta = 0$ or 0.3 in the structural model substantially overestimates the ICU congestion compared with using the identified $\hat{\beta}$. In the base case, setting $\beta = 0$ (resp. $\beta = 0.3$) overestimates the number of patients under high ICU congestion by 83.4% (resp. 25.7%) and the external arrivals who balk by 103.8% (resp. 30.9%). Similar bias holds under the system changes of increasing the ED arrivals and reducing one ICU bed. For example, using the identified $\hat{\beta}$ for each hospital, the 22 hospitals combined expect to see 678 external arrivals balking in a year after their ED arrivals are increased by 50%. However, this number increases to 1282 (resp. 866) when assuming $\beta = 0$ (resp. $\beta = 0.3$), causing a 89% (resp. 27.6%) upward bias. These observations suggest that using a full myopic model (setting $\beta = 0$) and ignoring the heterogeneity across hospitals (setting $\beta = 0.3$) tend to overestimate the high ICU congestion.

On the other hand, assuming $\beta = 0.99$, i.e., hospitals put substantial weight on the system impact of their decisions, would underestimate the ICU congestion. In the base case, it underestimates the patients under high ICU congestion (resp. externals who balk) by 5.6% (resp. 8.1%). Such bias becomes substantially larger when the ICU gets more constrained in adverse scenarios. After we increase ED arrivals by 50% (resp. reduce one ICU bed), the relative underestimation with $\beta = 0.99$ increases to 15.6% (resp. 7.3%) for high ICU congestion patients and 18.3% (resp. 8.3%) for external arrivals who balk. The absolute underestimation in the numbers of patients also becomes substantially larger than those in the base case.

The findings in Table 4 are in line with our interpretation of the discount factor in the model. As discussed, hospitals with larger discount factors respond to ICU state more proactively when making admission decisions, i.e., reducing ICU admission likelihood when ICU occupancy is high. This mitigates the high ICU congestion periods. Thus, setting $\beta = 0$ (resp. $\beta = 0.99$) in the structural model leads to overestimation (resp. underestimation) of high ICU congestion. As an illustration, we calculate the ICU admission probability for a single high-severity patient in the ED as the ICU occupancy increases. Figure 4 plots the results for Hospital 1 under three discount factors and the respective utility estimates: $\beta = 0$, $\beta = \hat{\beta} = 0.3$, and $\beta = 0.99$. We see that the choice of discount factor indeed affects how the ICU admission decisions

changes with the ICU state. With $\beta = 0$, the ICU admission probability is unaffected by the ICU occupancy. In contrast, if we set $\beta = 0.99$ in the model, the admission probability starts to decrease far before the ICU gets full, and the magnitude of drop is substantial.

Figure 4 Probability of ICU admission by ICU occupancy rates (Hospital 1)



We also compute other key system statistics under different discount factors, including overall ICU admission probabilities, ICU occupancy, and ED waiting times. We find that using $\beta = 0.99$ and $\beta = \hat{\beta}$ with their respective estimated utility parameters in the model produce close estimates for these system statistics. This reflects the potential "observational equivalence" issue that hinders the identification: multiple discount factor and utility parameters may generate similar system statistics. However, as shown in Table 4, the estimates of high ICU congestion, which are practically important, clearly hinge on the choice of discount factor. This is because the impact of the hospital's balancing behavior is more visible during the periods when the ICU is already congested, i.e., a future patient may be rejected due to lack of bed.

The counterfactual study highlights the importance of identifying the discount factor using observed data. Otherwise, misspecifying the discount factor in the model may lead to biased estimates of the ICU congestion, especially when the ICU demand is high and capacity is constrained. This is practically important for hospitals when they design their ICU capacities and evaluate the impact of different system shocks.

6. Conclusions

In many capacity-constrained service systems, it is important to balance the needs of current customers and the system's ability to serve future customers. We shed light on this aspect using a structural estimation approach. We develop a dynamic discrete choice model with multiple service levels and customer types, and measure the decision-maker's balancing behavior using the intertemporal discount factor. We show that the discount factor and utility parameters can be identified and estimated jointly from standard choice data, which contrasts with the common practice of pre-setting a discount factor in the literature.

We apply our model to an important problem in healthcare operations management: the ICU admission decisions of ED patients. Using a large US hospitalization data set, we find that the estimated discount factor is relatively small, far below the levels usually assumed in the literature. In addition, there is much heterogeneity in the balancing behaviors across hospitals. Thus, it is important to identify and estimate discount factor from data in different empirical settings. We then use counterfactual simulations to show that correctly identifying discount factor is important for estimating the ICU congestion level, especially in the scenarios when ICU becomes more capacity-constrained. Moreover, we find that changing the discount factor in the model introduces an intertemporal trade-off: increasing the discount factor reduces ICU congestion but leads to longer ED waiting time, with the negative impact being more significant for low severity class.

In this study, we consider a specific type of service system and use a dynamic discrete choice model to describe it. Future research may develop suitable structural models for other types of service systems and establish identification results of key behavioral parameters therein. In order to focus on the identification of discount factor, we have to make a number of simplifications in our model to ensure tractability. This inevitably ignores many realistic features of the system. For example, the customer arrival and departure rates may be heterogeneous and state-dependent. It would be interesting to investigate in future research whether these features can be incorporated into structural models. Moreover, there are other operational decisions that can be made in managing the system. For instance, decision-makers can interrupt the first-tier service for current customers to accommodate newly arrived customers that need the service more (e.g., the demand-driven discharge in ICU). As such, another potential direction is to study the joint impact of these decisions and the trade-offs they introduce.

Our empirical study on the ICU admission decisions has several limitations that may be explored on future research. First, the data we use have no direct information on the ICU admission process, as we can only observe the final admission units. Thus, 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 limited to hospitals within one healthcare system that uses a capitated payment model. Thus, it is possible that other payment models may also impact how the hospitals internalize intertemporal externalities. For example, a Fee-For-Service system may drive hospitals to be even more focused on the current patients.

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

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.

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, 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. 2020. Identification and estimation of forward-looking behavior: The case of consumer stockpiling. *Marketing Science* .

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.

De Groote, Olivier, Frank Verboven. 2019. Subsidies and time discounting in new technology adoption: Evidence from solar photovoltaic systems. *American Economic Review* **109**(6) 2137–72.

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.

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. 2020. A structural estimation approach to study agent attrition. *Management Science* **66**(9) 4071–4095.

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.

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

Freeman, Michael, Susan Robinson, Stefan Scholtes. 2021. Gatekeeping, fast and slow: An empirical study of referral errors in the emergency department. *Management Science* **67**(7) 4209–4232.

Freeman, Michael, Nicos Savva, Stefan Scholtes. 2016. Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science* **63**(10) 3147–3167.

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.

Hathaway, Brett A, Evgeny Kagan, Maqbool Dada. 2023. The gatekeeper's dilemma:"when should I transfer this customer?". *Operations Research* **71**(3) 843–859.

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

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.

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. 2020. Admission control biases in hospital unit capacity management: How occupancy information hurdles and decision noise impact utilization. *Management Science* **66**(11) 5151–5170.

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.

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

Mandelbaum, Avishai, Petar Momčilović, Yulia Tseytlin. 2012. On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science* **58**(7) 1273–1291.

Manski, Charles F. 1993. Identification of endogenous social effects: The reflection problem. *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.

Norets, Andriy, Xun Tang. 2014. Semiparametric inference in dynamic binary choice models. *Review of Economic Studies* **81**(3) 1229–1262.

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.

Rath, Sandeep, Kumar Rajaram. 2022. Staff planning for hospitals with implicit cost estimation and stochastic optimization. *Production and Operations Management* **31**(3) 1271–1289.

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.

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

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 L Tucker, Ryan Graue, Sarah Moravick, Julius J Yang. 2020. Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science* **66**(9) 3825–3842.

Wang, Guihua, Jun Li, Wallace J. Hopp, Franco L. Fazzalari, Steven F. Bolling. 2019. Using patient-specific quality information to unlock hidden healthcare capabilities. *Manufacturing & Service Operations Management* **21**(3) 582–601.

Weingart, Scott D, Robert L Sherwin, Lillian L Emlet, Isaac Tawil, Julie Mayglothling, Jon C Rittenberger. 2013. ED intensivists and ED intensive care units. *American Journal of Emergency Medicine* **31**(3) 617–620.

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

Zhang, Zhe George, Hsing Paul Luh, Chia-Hung Wang. 2011. Modeling security-check queues. *Management Science* **57**(11) 1979–1995.

Electronic Companion to "Structural Estimation of Intertemporal Externalities with Application in ICU Admissions"

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

EC.1. Formulae and Assumptions

EC.1.1. Explicit expressions for state transition probability $g(s'|s)$

We provide explicit expressions for the function $g(s'|s)$ used in Proposition 1, which is the transition probability from the intermediate state $s = (\tilde{n}^G_{l,t}, \tilde{n}^G_{h,t}, \tilde{n}^F_t)$ (after action is taken) to the state $s' = (n^G_{l,t+1}, n^G_{h,t+1}, n^F_{t+1})$ at the start of the next period (before action). As the GK arrivals are independent of the FSU external arrivals and departures, we have

$$g(s'|s) = g_{Q,l}\left(n^G_{l,t+1}|\tilde{n}^G_{l,t}\right) g_{Q,h}\left(n^G_{h,t+1}|\tilde{n}^G_{h,t}\right) g_F\left(n^F_{t+1}|\tilde{n}^F_t\right), \tag{EC.1}$$

where $g_{Q,l}$ and $g_{Q,h}$ denote the transition probabilities for the numbers of GK customers from low and high classes respectively, and g_F denotes the transition probability for the number of FSU customers.

For the GK 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 GK capacity constraint, the number of GK arrivals is capped by $\bar{A}_i = \min\{M_{A,i}, Q_i - \tilde{n}^G_{i,t}\}$. Thus, the transition probability can be computed as

$$g_{Q,i}(m|n) = \begin{cases} (\lambda_{Q,i})^{m-n} \exp(-\lambda_{Q,i})/(m-n)! & \text{if } n \le m < n + \min\{M_{A,i}, Q_i - n\} \\ \sum_{j=\bar{A}_i}^{+\infty} (\lambda_{Q,i})^j \exp(-\lambda_{Q,i})/j! & \text{if } m = n + \min\{M_{A,i}, Q_i - n\} \end{cases} \tag{EC.2}$$

and $g_{Q,i}(m|n) = 0$ elsewhere.

For the FSU transition probability g_F, 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 FSU capacity $B-\tilde{n}_t^F$. With E_t external arrivals in the period, the FSU would have total $\tilde{n}_t^F+E_t$ customers. Then, the number of departures, D_t, follows a Binomial-$(\tilde{n}_t^F+E_t,\mu_I)$ distribution. Given the number of external arrivals, E_t, the number of departures follows by $D_t=\tilde{n}_t^F+E_t-n_{t+1}^F$. We note that E_t can be greater than $\max\{n_{t+1}^F-\tilde{n}_t^F,0\}$. Summing up the probability of all possible choices of the Poisson-distributed E_t (and Binomial-distributed D_t accordingly), the transition probability $g_F(m|n)$ is given by

$$\begin{aligned} g_F(m|n) = & \sum_{j=\max\{m-n,0\}}^{B-n-1} \lambda_E^j \frac{\exp(-\lambda_E)}{j!} \frac{(n+j)!}{(n+j-m)!m!} \mu_I^{n+j-m}(1-\mu_I)^m \\ & + \left(\sum_{j=B-n}^{+\infty} \lambda_E^k \frac{\exp(-\lambda_E)}{j!} \right) \frac{B!}{(B-m)!m!} \mu_I^{B-m}(1-\mu_I)^m, \text{ for } 0 \le m \le B. \end{aligned} \tag{EC.3}$$

Combining (EC.2) and (EC.3), we obtain the explicit expression for state transition probability $g(s'|s)$ by (EC.1).

EC.1.2. Model Assumptions for Identification

This section documents the assumptions 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 is the random perturbation in utility.

Assumption 1 *(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 is 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) $X=\{1,\ldots,K\}$.

Assumption 2 (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 action utility given by (2) as

$$u(s_t, d_t) = u_{a,l} a_{l,t} + u_{r,l} r_{l,t} + u_{w,l} \left(n^G_{l,t} - a_{l,t} - r_{l,t}\right) + u_{a,h} a_{h,t} + u_{r,h} r_{h,t} + u_{w,h} \left(n^G_{h,t} - a_{h,t} - r_{h,t}\right).$$

This is the deterministic part of the per-period utility. Thus, it is indeed linear in parameters $\theta = \{u_{a,l}, u_{r,l}, u_{w,l}, u_{a,h}, u_{r,h}, u_{w,h}\}$. In our model, the functions π_0 and π_1 specify to $\pi_0(d_t, s_t) \equiv 0$ and

$$\pi_1(d_t, s_t) = [a_{l,t}, r_{l,t}, n^G_{l,t} - a_{l,t} - r_{l,t}, a_{h,t}, r_{h,t}, n^G_{h,t} - a_{h,t} - r_{h,t}]^\top.$$

EC.2. Proofs

EC.2.1. Proof for Proposition 2

PROOF: From the definition of the value functions $V(s_t, \varepsilon_t)$ in (7) and $\tilde{V}(s)$ in (10), we have:

$$\tilde{V}(s) = \sum_{s_t} \int_{\varepsilon_t} \sup_{d_t \in \Pi(s_t)} \mathsf{E}\left\{\sum_{j=t}^{\infty} \beta^{j-t} U(s_j, d_j, \varepsilon_j) | s_t, \varepsilon_t\right\} g(s_t|s) q(\varepsilon_t|s_t) d\varepsilon_t. \tag{EC.4}$$

The expectation above is taken over the transition of (s_j, ε_j) starting from (s_t, ε_t) based on the transition probability $g(s'|s)$ and probability density $q(\varepsilon'|s')$. Thus, the function $\tilde{V}(s)$ represents the expected future utilities starting from intermediate state s and assuming the system 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 $s_t = \left(n^G_{l,t}, n^G_{h,t}, n^F_t\right)$ and $s'_t = \left(\left(n^G_{l,t}\right)', \left(n^G_{h,t}\right)', \left(n^F_t\right)'\right)$. Additionally, we use $\tilde{s}_t = \left(\tilde{n}^G_{l,t}, \tilde{n}^G_{h,t}, \tilde{n}^F_t\right)$ and $\tilde{s}'_t = \left(\left(\tilde{n}^G_{l,t}\right)', \left(\tilde{n}^G_{h,t}\right)', \left(\tilde{n}^F_t\right)'\right)$ 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 (5). 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

$$\tilde{n}^G_{i,0} = \left(\tilde{n}^G_{i,0}\right)' \text{ for } i \in \{l, h\}, \text{ and } \tilde{n}^F_0 \le \left(\tilde{n}^F_0\right)', \tag{EC.5}$$

then Proposition 2 translates to $\tilde{V}(\tilde{s}_0) \ge \tilde{V}(\tilde{s}'_0)$.

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 GK and external arrivals to the FSU 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 FSU departures from the two systems as follows. Denote the numbers of FSU customers before departures by $\bar{n}^F_t$ and $(\bar{n}^F_t)'$ respectively and assume $\bar{n}^F_t \le (\bar{n}^F_t)'$. 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-$((\bar{n}^F_t)' - \bar{n}^F_t, \mu_I)$ 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.

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

$$\tilde{n}^G_{i,t-1} = \left(\tilde{n}^G_{i,t-1}\right)' \text{ for } i \in \{l,h\}, \text{ and } \tilde{n}^F_{t-1} \le \left(\tilde{n}^F_{t-1}\right)', \tag{EC.6}$$

then the states at the start of period t satisfy

$$n^G_{i,t} = \left(n^G_{i,t}\right)' \text{ for } i \in \{l,h\} \text{ and } n^F_t \le \left(n^F_t\right)'. \tag{EC.7}$$

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 GK, external arrivals to the FSU, and departures from the FSU during period $t-1$.

It is trivial to see the relationship for the customers in GK holds as we assume same GK arrival processes by coupling. For the FSU customers, we consider both external arrivals and departures. We have

$$\begin{aligned}\left(n^F_t\right)' - n^F_t &= \min\left\{(\tilde{n}^F_{t-1})' + E'_{t-1}, B\right\} - \min\left\{\tilde{n}^F_{t-1} + E_{t-1}, B\right\} - \left(D'_{t-1} - D_{t-1}\right)\\ &= \min\left\{(\tilde{n}^F_{t-1})' + E_{t-1}, B\right\} - \min\left\{\tilde{n}^F_{t-1} + E_{t-1}, B\right\} - Z_{t-1} \ge 0.\end{aligned}$$

The last equality follows from the coupling $E_t = E'_t$ for external arrivals and $D_{t-1} + Z_{t-1} = D'_{t-1}$ for departures. The last inequality holds as Z_{t-1} follows a Binomial-(L_{t-1}, μ_I) distribution with $L_{t-1} = \min\left\{(\tilde{n}^F_{t-1})' + E_{t-1}, B\right\} - \min\left\{\tilde{n}^F_{t-1} + E_{t-1}, B\right\}$. This completes the proof for Lemma 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 (EC.4). 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 (EC.4) 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

$$\tilde{V}(\tilde{s}_0) \ge V^\pi(\tilde{s}_0). \tag{EC.8}$$

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 customers in the GK, but system s has no more customers in the FSU:

$$n^G_{i,t} = \left(n^G_{i,t}\right)' \text{ for } i \in \{l,h\} \text{ and } n^F_t \le \left(n^F_t\right)', \quad \forall t. \tag{EC.9}$$

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:

$$d_t = d'_t \in \Pi(s_t), \quad \forall t. \tag{EC.10}$$

Note that (EC.9) directly implies (EC.10), as it follows from (1) that given the same number of customers in GK, the system with fewer FSU customers has a larger admissible action set, leading to $d'_t \in \Pi(s'_t) \subseteq \Pi(s_t)$.

Induction: We establish (EC.9) 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 (EC.5), and from Lemma 1. Thus we have $n^G_{i,1} = \left(n^G_{i,1}\right)'$ for $i \in \{l,h\}$ and $n^F_1 \le \left(n^F_1\right)'$.

Inductive Step: We assume (EC.9) 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, it is easy to verify the intermediate states after action also coincide, i.e., $\tilde{n}^G_{i,j} = \left(\tilde{n}^G_{i,j}\right)'$ for $i \in \{l,h\}$ and $\tilde{n}^F_j = \left(\tilde{n}^F_j\right)'$. Finally, we can apply Lemma 1 to prove the relationship (EC.9) 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 customers in the GK, and system s always has fewer customers in the FSU 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:

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

This is because: (i) Both systems take the same action, thus they send same numbers of customers to the FSU and SSU, leading to same routing utilities; (ii) As both systems have same numbers of customers in the GK, the number of customers remaining in the GK after actions are also the same, leading to the same waiting utilities; (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 (EC.8)

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

This proves the proposition. □

EC.2.2. Proof for Proposition 3

PROOF: With non-zero utility $u_{a,h}$ for sending the high-priority customer to the FSU, the per-period utility of action is given by

$$\begin{aligned} u'(s_t, d_t) &= u_{a,l} a_{l,t} + u_{r,l} r_{l,t} + u_{w,l}\left(n^G_{l,t} - a_{l,t} - r_{l,t}\right) + u_{a,h} a_{h,t} + u_{r,h} r_{h,t} + u_{w,h}\left(n^G_{h,t} - a_{h,t} - r_{h,t}\right) \\ &= \sum_{\iota\in\{l,h\}} \Delta u_{w,\iota}\left(n^G_{\iota,t} - a_{\iota,t} - r_{\iota,t}\right) + \sum_{\iota\in\{l,h\}} \Delta u_{r,\iota} r_{\iota,t} + \Delta u_{a,l} a_{l,t} + u_{a,h} n^G_t, \end{aligned} \tag{EC.11}$$

where $n^G_t = n^G_{l,t} + n^G_{h,t}$ denotes the total number of customers in GK in period t. In addition, we denote $\Delta u_{w,\iota} = u_{w,\iota} - u_{a,h}$, and $\Delta u_{r,\iota} = u_{r,\iota} - u_{a,h}$ for $\iota \in \{l,h\}$, as well as $\Delta u_{a,l} = u_{a,l} - u_{a,h}$. They represent the relative actions utilities compared with sending the high-priority customer to the FSU.

We see that, with non-zero utility for sending high-priority customer to the FSU, the last term $u_{a,h}n_t^G$ is involved in the per-period utility. Denote the per-period utility and value function with non-zero FSU routing utility $u_{a,h}$ by $U'(s_t, d_t, \varepsilon_t)$ and $V'(s_t, \varepsilon_t)$, respectively. By (7), we have

$$\begin{aligned} V'(s_t,\varepsilon_t) &= \sup_{d_t\in\Pi(s_t)} \mathsf{E}\left\{\sum_{j=t}^{\infty}\beta^{j-t}U'(s_j,d_j,\varepsilon_j)|s_t,\varepsilon_t\right\} \\ &= \sup_{d_t\in\Pi(s_t)} \mathsf{E}\left\{\sum_{j=t}^{\infty}\beta^{j-t}\left[U(s_j,d_j,\varepsilon_j)+u_{a,h}n_j^G\right]|s_t,\varepsilon_t\right\}, \end{aligned} \tag{EC.12}$$

where $U(s_j, d_j, \varepsilon_j)$ denotes the per-period utility with zero FSU routing utility $u_{a,h}$ and same relative utilities of other actions, i.e., $\Delta u_{w,\iota}$ and $\Delta u_{r,\iota}$ for $\iota \in \{l, h\}$ as well as $\Delta u_{a,l}$.

With same total number of customers in GK for each period, we can drop the subscript j in n_j^G and write it as n^G. Then the term $u_{a,h}n^G$ in (EC.12) is independent of action d_t and can be extracted out of the conditional expectation. We can write (EC.12) as

$$V'(s_t,\varepsilon_t) = \sup_{d_t\in\Pi(s_t)} \mathsf{E}\left\{\sum_{j=t}^{\infty}\beta^{j-t}U(s_j,d_j,\varepsilon_j)|s_t,\varepsilon_t\right\} + \frac{1}{1-\beta}u_{a,h}n^G, \tag{EC.13}$$

where the last term is a constant number and independent of the optimal actions. We note that by (7), the first term in (EC.13) is exactly the value function under the model with zero FSU routing utility for high-priority customers and same relative utilities for other actions. Thus, we can establish the following relationship between the value functions with different normalization levels of $u_{a,h}$:

$$V'(s_t,\varepsilon_t) = V(s_t,\varepsilon_t) + \frac{1}{1-\beta}u_{a,h}n^G.$$

Similarly, by (10), we can prove the function $\tilde{V}'(s)$ under non-zero utility $u_{a,h}$ satisfies

$$\tilde{V}'(s) = \tilde{V}(s) + \frac{1}{1-\beta}u_{a,h}n^G. \tag{EC.14}$$

By (EC.11) and (EC.14), we have

$$u'(s,d) + \beta\tilde{V}'(\varphi(s,d)) = u(s,d) + \beta\tilde{V}(\varphi(s,d)) + \frac{1}{1-\beta}u_{a,h}n^G.$$

Plugging this in to (11), it is easy to verify the new function $\tilde{V}'(s)$ satisfies the functional equation. Moreover, by (9), we can show that the new choice probability $f'(d_t|s_t)$ coincides with the original ones under zero FSU routing utility $u_{a,h}$. This shows our model identification is not affected by the normalization level of $u_{a,h}$ under constant number of customers in GK. □

EC.3. Data Selection and Estimation of System Parameters for ICU Admission

EC.3.1. Data Selection Process

We start from a total of 312,306 hospitalizations over two years. We restrict our study to the hospitalizations admitted to a medical service via the ED, which comprises the largest proportion of admitted patients. For the ED patients, 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 drop 12 hospitalizations with unknown gender and 9,128 (4.8%) hospitalizations for patients who experience hospital transfers or transports outside of the hospital network. 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 admit the patient to a non-ICU unit (e.g. the ward or TCU). 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).

We restrict our study cohort to the periods of each hospital with stable ICU capacity and occupancy. 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. The number of days and hospitalizations for each hospital in the final study cohort are summarized in the first two columns of Table EC.5 in Section EC.6. In total, we drop 11,268 (6.4%) hospitalizations that are outside the stable periods.

EC.3.2. Estimation of System Parameters

As described in Section 2.3, the arrival and departure rates, as well as the unit capacities, are estimated directly from data – outside of the structural model. We now describe how we do this for the ICU admission problem. Recall we define each period as a two hour interval in the 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 *all* patients (medical and surgical, emergency and elective) in the ICU observed from data. This is a reasonable assumption as ICU often operates under high congestion.

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. Thus, it is difficult to accurately determine the ED capacity in

our model. However, we note that the number of ED boarding patients is relatively low (e.g., average 1.21 patients). Thus, we set ED capacities Q_i using the following heuristic: $Q_i = M_{Q,i} + \lfloor \sqrt{M_{A,i}} \rfloor$, where $M_{Q,i}$ is the maximum number of ED patients observed in the data; $M_{A,i}$ is the maximum number of arrivals in each period; and $\lfloor \cdot \rfloor$ denotes the floor function.The square root term $\sqrt{M_{A,i}}$ is introduced as a heuristic "safety buffer" to ensure we have ample ED capacity to avoid balking upon arrival to the ED, which rarely happens in reality. We verify by simulation that the ED rarely reaches its full capacity Q_i in our structural model. In addition, the choice probabilities are very robust to alternative choices of the ED capacity. The estimated system statistics are reported in Table EC.5 of Section EC.6 for each hospital.

EC.4. Reduced-form Evidence for Discounting Behavior in ICU Admissions

In this section, we conduct reduced-form regressions to analyze the main determinants of the system's ICU admission decisions. This provides preliminary evidence for the discounting behaviors of hospitals. We apply a multinomial logistic model to estimate the ICU admission decisions. As with the structural model, we set a period to be two hours. In each period, the hospital chooses one of the three options for each patient: admit the patient into the ICU; admit the patient to non-ICU units; or make the patient wait in the ED. At the start of each period, we construct system "snapshots" which includes detailed information for each patient in the ED as well as the state of the ED and ICU. We include patient characteristics, system state variables, and seasonality effects as potential determinants of the admission decisions.

We use the non-ICU admission decision as the base case, and estimate the probabilities of the ICU admission ($d = \text{ICUAdm}$) and waiting ($d = \text{Wait}$) decisions relative to the non-ICU admission decision respectively. For patient i who is in the ED at the start of period t, we set up the multinomial logistic model:

$$\ln\left[\frac{\Pr(d_{it}|\mathbf{X}_i,\mathbf{S}_t)}{\Pr(\text{nonICU}_{it}|\mathbf{X}_i,\mathbf{S}_t)}\right] = \gamma_{0,d} + \gamma_{L,d}\text{LAPS2}_i + \gamma_{ICU,d}\text{ICUOccu}_t + \gamma_{ED,d}\text{EDNum}_t + \boldsymbol{\gamma}_{Z,d}^{\top}\mathbf{Z}_{i,t} + \epsilon_{it}, \tag{EC.15}$$

where $\Pr(d_{it})$ and $\Pr(\text{nonICU}_{it}|\mathbf{X}_i\mathbf{S}_t)$ denote the conditional probability of action d_{it} and admitting patient i to non-ICU units in period t, respectively. In (EC.15), LAPS2 is the patient's main severity score. 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. $\mathbf{Z}_{i,t}$ denotes other control covariates, given by

$$\mathbf{Z}_{i,t} = \{\text{Gender}_i, \text{Age}_i, \text{COPS2}_i, \text{CHMR}_i, \text{Hosp}_i, \text{DepPre}_t, \text{AvgLAPS2}_t, \text{DayOfWeek}_t, \text{HourOfDay}_t, \text{Month}_t\}.$$

It includes patient i's gender, age, identifier for hospital admitted, as well as other two severity scores COPS2 and CHMR. In addition, it includes several system state variables: the number of patients who left the ICU in the previous period (DepPre_t), the average severity level measured by the LAPS2 score of the current ICU patients in period t (AvgLAPS2_t), and several categorical variables to capture the potential

seasonality and time trend in the decisions. Here DayOfWeek_t and HourOfDay_t denote the day of week and hour of day respectively; Month_t is the dummy variable representing the month in the sample (total 23 months). To account for the heteroskedasticity, we cluster standard errors by hospitals in the regression.

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 Hosp_i in (EC.15). We use the McFadden's pseudo R-squared to measure the goodness of fit of the model. It is defined as $R^2 = 1 - \ln l^{mod}/\ln l^{null}$, where l^{mod} is the likelihood from the estimated model and l^{null} is the likelihood from the "null" model that only includes the intercept and a categorical variable for each hospital in (EC.15).

Table EC.1 Estimation results for Multinomial-Logistic Regression (EC.15), N = 183,691, R-squared = 0.16

	LAPS2_i	ICUOccu_t	EDNum_t
	γ_L	γ_{ICU}	γ_{ED}
Waiting	0.008	1.178	0.241
	(0.000)	(0.031)	(0.006)
ICUAdm	0.028	−0.396	−0.014
	(0.000)	(0.030)	(0.006)

Standard error is reported in parenthesis.

We report the estimated coefficients for three main variables in the multinomial logistic model: LAPS2_i, ICUOccu_t, and EDNum_t. Table EC.1 shows the estimation results for model (EC.15) with all hospitals combined.We find that all the coefficients are statistically significant and have the expected sign. In particular, higher LAPS2 score increases the probability of admission to ICU relative to other units, as these patients are more critically ill. In addition, 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 ED) decreases the probability of ICU admission and increases the probability of waiting. Such evidence suggests that the hospitals indeed internalizes the intertemporal externalities on the ICU admission decisions by adjusting their behaviors according to the current system state. We find similar results using the average marginal effects of the variables. The McFadden's pseudo R^2 for the combined estimation is 0.16. This is consistent with the magnitude seen for models of operational decisions in healthcare systems (see, e.g., Kim et al. 2015 and Song et al. 2020).

We also estimate model (EC.15) for each hospital separately. The results are qualitatively similar to that for all hospitals combined in Table EC.1. Most of the coefficients have the expected signs for individual hospitals, although some are not statistically significant due to a much smaller sample size. Specifically, higher ICU occupancy is still associated with decreased likelihood of ICU admission decisions for most hospitals. The McFaddean's R^2 for individual hospital regressions ranges from 0.14 to 0.24, with an average of 0.17. Full estimation results are available from authors upon request.

EC.5. Goodness of Fit and Robustness Checks

EC.5.1. Goodness of Fit

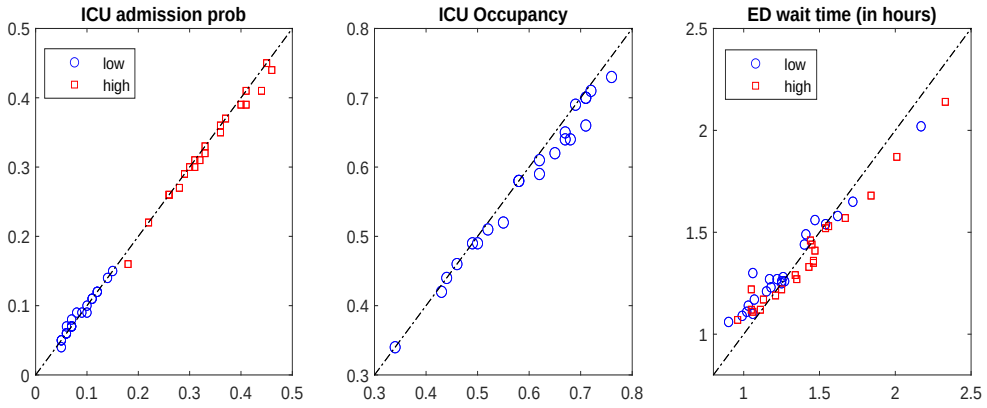
We show our estimated structural model provides good fits for both the hospitals' decisions and the overall system statistics. First, we compare the explanatory power for the hospitals' decisions, as measured by the McFadden's pseudo R^2 in (13), from our structural model with that from the reduced-form multinomial logistic regression model in Section EC.4. As shown in Tables 1 and EC.1, the pseudo R^2 for the combined estimation of all hospitals is similar from the structural model and logistic regression (0.14 versus 0.16). The average pseudo R^2 for individual hospital estimations is also comparable (0.13 vs 0.17). We note that the multinomial model contains a comprehensive set of variables that might influence admission decisions, including patient's characteristics, system states, and multiple dummies for seasonality fixed effects. Thus, the similar R^2 suggests that our structural model has reasonable explanatory power in capturing the hospitals' decisions. Of course, we acknowledge that there is still quite a bit of variation in the data that our model can not capture.

We find the structural model and reduced-form regression provide consistent evidences for the heterogeneity in hospitals' discounting behaviors. In the multinomial logistic model, we measure the impact of ICU occupancy on ICU admission likelihood by the coefficient γ_{ICU}. As discussed in previous section, γ_{ICU} is significantly negative for most hospitals, suggesting the ICU admission probability drops as ICU occupancy increases. We find that the estimated $\hat{\beta}$ and γ_{ICU} are negatively correlated: the correlation is -0.770 across the 22 hospitals, statistically significant at the 0.1% level. As such, hospitals with larger estimated discount factors indeed tend to be more responsive to ICU congestion when admitting patients.

Next, we use simulation to show that our structural model produces system statistics close to those observed in the data. As the arrival and departure rates are directly estimated from data, the average number of arrivals and departures in each period of our model are close to that observed in the data. Therefore, we examine other important system statistics, including the average ICU occupancy, overall proportion eventually admitted to the ICU for the two classes, as well as the ED waiting times for the two classes. The statistics estimated from our structural model are averaged from 200 simulation trials as in the counterfactual studies in Section 5.

Figure EC.1 compares the simulated and observed system statistics: ICU admission probabilities (left), ICU occupancy (middle), and ED waiting times (right). In the panels, 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 very close to the 45-degree line, implying that our estimated structural model produces system statistics that are very close to the observed data. Although our structural model is trained to fit the choice probabilities of the hospital's actions, it also produces a system that fits a number of observed key metrics very well. This further supports its effectiveness in modeling the admission process for ED patients.

Figure EC.1 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), including ICU admission probability (left), ICU occupancy (middle), and ED waiting times (right).

With a small number of parameters, over-fitting is unlikely for our structural model. To further address this concern, we divide the sample to first and second halves for each hospital. We estimate the structural model with all hospitals combined using the first half sample, and do out-of-sample prediction on the second half. The McFadden's pseudo R^2 from the out-of-sample prediction (0.13) is very similar to that from the in-sample estimation (0.14) and the level from the full sample estimation in Table 1 (0.14).

EC.5.2. Robustness Checks

In this section, we conduct several robustness checks for our empirical results in Section 4. Our main findings are largely robust in these settings.

Choice of Normalization Levels: As discussed in Section 2.2, the normalization level of the ICU admission utilities $u_{a,l}$ and $u_{a,h}$ can affect the identification results of our model. We perform two robustness checks for the normalization choice. First, similar to the proof for Proposition 3, we can show that the identification is not affected by our normalization choice of $u_{a,l}$ and $u_{a,h}$ in a model with constant ED state $(n^G_{l,t}, n^G_{h,t})$. So we now estimate our model using the periods with same or similar ED states, which is expected to mitigate the impact of the normalization choice. We select the state to be $(n^G_l, n^G_h) = (2, 1)$, i.e., with two low severity patients and one high severity patient. This state is selected for three reasons. First, there are both high and low severity patients in the ED. So the difference in the two patient classes can be captured by their admission decisions. Second, there are multiple patients in the ED, so the action space is large enough to reflect how the hospitals are balancing current versus future utilities. Third, the occurrence of the state is not too rare, so the model can be effectively estimated.

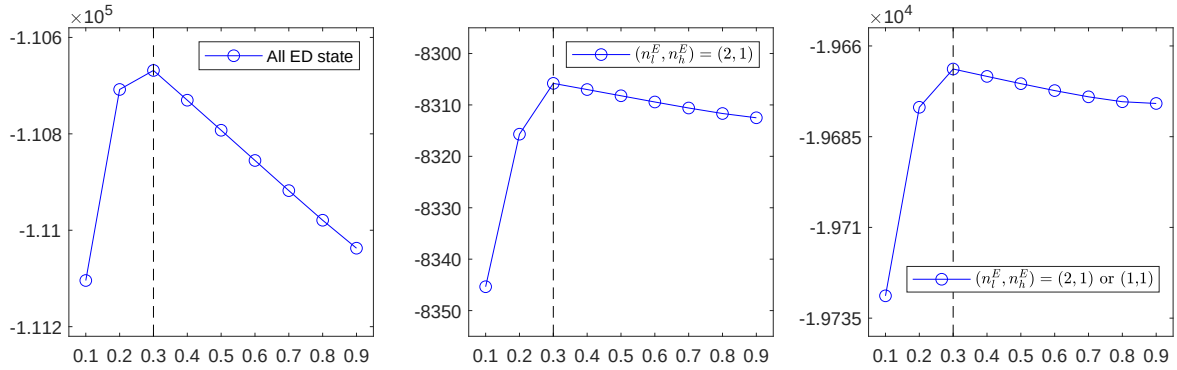
Due to sample size limit, we estimate the model by combining the sample from all hospitals by the procedure described in Section 4.1. The difference is we now calculate the log-likelihood l^f in (12) only using the actions associated with the ED state $(n^G_l, n^G_h) = (2, 1)$. The results are reported in Table EC.2.

The first row repeats Table 1 for our original model, in which all ED states are used. The second row reports the estimation results when we only use the ED state $(n_l^G, n_h^G) = (2, 1)$. As an additional check, we report in the third row the estimated parameters when we use the ED states $(n_l^G, n_h^G) = (2, 1)$ or $(1, 1)$. This further increases the sample size. The ICU admission utilities are assumed to be zero in the estimation, as in our main analysis. We see $\hat{\beta}$ equals to 0.3 in all the three cases, suggesting the identified discount factor is not affected when we limit to the select ED states. Figure EC.2 further shows how the log likelihood varies with discount factor in the three cases. We see the likelihood function decreases monotonically as we deviate from the optimal $\hat{\beta} = 0.3$ (vertical dashed line) in all three cases. In addition, the waiting and non-ICU admission utilities are also similar to the original model for both classes. These findings support the robustness of our identification results.

Table EC.2 Estimation results with different ED states (All hospitals combined)

(n_l^E, n_h^E)	Size	$\hat{\beta}$	$\hat{u}_{w,l}$	$\hat{u}_{r,l}$	$\hat{u}_{w,h}$	$\hat{u}_{r,h}$
All	154,140	0.3	-0.071	1.950	-0.932	0.671
		(0.005)	(0.001)	(0.010)	(0.018)	(0.013)
(2,1)	4,230	0.3	-0.013	2.001	-0.966	0.775
		(0.019)	(0.005)	(0.042)	(0.051)	(0.036)
(2,1) or (1,1)	11,706	0.3	-0.071	2.215	-1.059	0.727
		(0.023)	(0.044)	(0.031)	(0.040)	(0.022)

Figure EC.2 Log likelihood at different β for select ED states (all hospitals combined)



As discussed in Section 3.3, we expect the utilities of ICU admission to be small in our empirical setting. In the second robustness check, we further estimate our model for each hospital separately under small but non-zero ICU admission utilities. Here all ED states are used in the estimation. We consider multiple combinations of ICU admission utilities $u_{a,l}$ and $u_{a,h}$, including $(-0.05, -0.05)$, $(-0.05, -0.1)$, $(-0.1, -0.05)$, $(-0.1, -0.1)$, $(-0.1, -0.2)$, $(-0.2, -0.1)$, and $(-0.2, -0.2)$. We find the identified discount factors $\hat{\beta}$ are largely robust to these alternative choices of ICU admission utilities. Recall we estimate the discount factor

on the grid of $\{0, 1, 0.2, ..., 0.9\}$. In particular, the identified $\hat{\beta}$ remains the same or only changes to the adjacent level (i.e., with absolute change smaller than 0.1) for all hospitals in these alternative settings, with the only exception of Hospital 11 under $(u_{a,l}, u_{a,h}) = (-0.2, -0.1)$ and $(-0.2, -0.2)$. We also calculate the correlation (across hospitals) between $\hat{\beta}$ under the original model and each of the alternative settings. The correlation is higher than 0.89 in all cases. Thus, our results are robust to the choice of normalization levels.

Estimation with Stratified Sample: In our main specification, we assume the model parameters are constant for the entire sample. To address potential seasonality issues, we also estimate the structural model with stratified samples based on flu vs non-flu seasons, as well as day and night periods. The flu season ranges from November to March, and the non-flu season includes the rest. The day periods include the twelve hours from 7AM to 7PM. We re-estimate the ED arrival rates $\lambda_{Q,l}$ and $\lambda_{Q,h}$, external ICU arrival rate λ_E, and ICU departure rate μ_I for each stratified sample. The ED and ICU capacities Q_l, Q_h, and B are set as those for the full sample in Table EC.5.

We first estimate the structural model for all hospitals combined with the stratified samples. The results are shown in Table EC.3, with the first two rows repeating the estimates in Table 1 for the full sample. We can see the estimated parameters are close to those in our main specification. In particular, the estimated discount factor is 0.4 for the flu season, 0.2 for the non-flu season, and 0.4 for the day sample; all of which are close to the level $\hat{\beta} = 0.3$ for the full sample. The utility parameters and pseudo R^2 are also similar. The only difference is that the discount factor is estimated to be lower ($\hat{\beta} = 0.1$) for the night sample. Based on conversations with our clinical collaborators, this may be explained as follows. First, the demand for ICU beds is higher during the day due to more external arrivals from surgical patients as well as more medical patients from the ED.[1] Thus, the hospital is more likely to account for the system impact of its admission decisions during the day. Second, there may be more medical and operational constraints that hinder the hospital from responding to the system state during the night, such as shortage in nurse availability and delay in system state information.

We then estimate the structural model for each hospital separately with the stratified data. We find the heterogeneity in discount factors estimated from the stratified data is largely consistent with that from the full sample given in Table EC.6. In particular, the across-hospital correlations between the estimated $\hat{\beta}$ from stratified and full sample are significantly positive: 0.87 for the flu season, 0.67 for the non-flu season and day sample, and 0.46 for the night sample. Thus, our main findings from the structural model are robust to accounting for these potential temporal variations.

Finer Grid Search and Three Patient Class: For computational purposes, we conduct a grid search over the discount factor with a step of 0.1 in our main specification. This facilitates the numerical implementation of our estimation as the state space in our model is relatively large and the value function needs to be

[1] In our data, the average external arrival rate to ICU is 0.291 patients per period in the day versus only 0.149 in the night.

Table EC.3 Estimation of structural model with stratified data: All hospitals combined

	Size	Discount factor	Low Severity		High Severity		
	N	$\hat{\beta}$	$\hat{u}_{w,l}$	$\hat{u}_{r,l}$	$\hat{u}_{w,h}$	$\hat{u}_{r,h}$	R^2
All	154,140	0.3	-0.071	1.950	-0.932	0.671	0.14
		(0.005)	(0.001)	(0.010)	(0.018)	(0.013)	
Flu	60,031	0.4	-0.028	1.936	-0.755	0.674	0.14
		(0.004)	(0.010)	(0.012)	(0.021)	(0.019)	
Non-flu	108,737	0.2	-0.058	1.960	-1.086	0.670	0.13
		(0.005)	(0.010)	(0.009)	(0.023)	(0.016)	
Day	84,372	0.4	-0.001	2.006	-0.806	0.655	0.13
		(0.004)	(0.009)	(0.011)	(0.020)	(0.018)	
Night	84,396	0.1	-0.045	1.875	-0.971	0.684	0.14
		(0.005)	(0.011)	(0.009)	(0.024)	(0.016)	

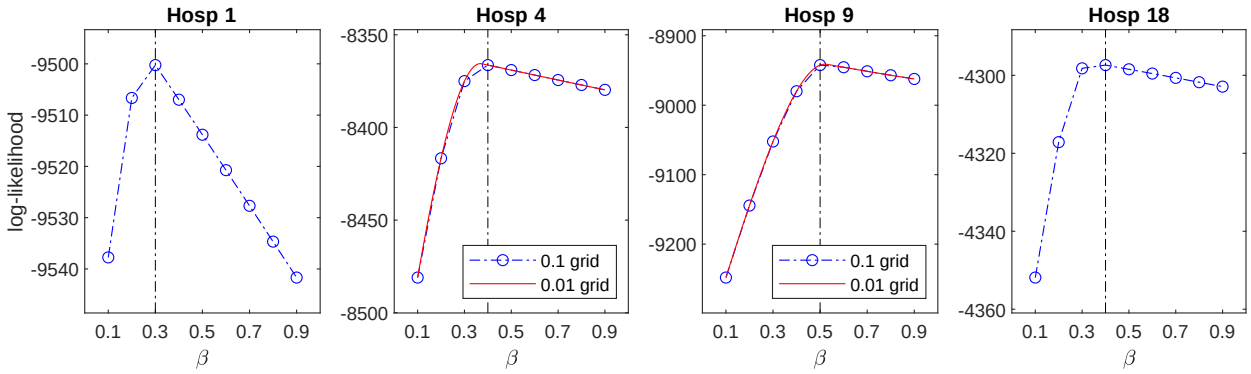
Standard error is reported in parenthesis.

solved at each candidate utility parameter set for a given discount factor. Specifically, the dimension of the state space is given by $(Q_l + 1) \times (Q_h + 1) \times (B + 1)$, which is around 810 on average for a hospital. Besides, the grid with a step of 0.1 is granular enough to measure the hospital's discounting behavior. As a robustness check, we have also done a finer grid search with 0.01 increment for select Hospitals 4 and 9. The likelihoods with respect to β for these two hospitals are shown in Figure EC.3 (red solid lines). We see that the identified β's from the finer grid are very close to those from the original grid.

Due to the high-dimensional nature of our dynamic model, we use two patient classes in our main specification. We run additional experiments with three patient classes for select hospitals to check the robustness of our findings. In the three-class model, we define the low severity class as those with LAPS2 score below the 60th percentile of the LAPS2 score distribution, the middle severity class as those between the 60th and 85th percentile, and the high severity class as those above the 85th percentile. That is, we use the same high severity patient class under the two-class and three-class models. We implement the three-class model for Hospital 5, 11, and 20, with identified discount factor of 0.1, 0.9, and 0.3 under the original two-class model respectively. The ICU admission utilities are assumed to be zero for all the three classes. Under the three-class model, the identified discount factor of Hospital 11 and 20 remains unchanged (0.9 and 0.3), and the discount factor of Hospital 5 changes slightly from 0.1 to 0.2. In addition, we find the waiting and non-ICU admission utility parameters for the high severity class are similar under the two-class and three-class models. These results show the robustness of our main findings.

EC.6. Supplementary Tables and Figures

Figure EC.3 Examples of log-likelihood versus discount factor for a subset of hospital



Note: The estimated log-likelihood at $\beta = \{0.1, 0.2, \ldots, 0.9\}$ for Hospitals 1, 4, 9, and 18. The β with the best likelihood is plotted by the black vertical line. The red solid line shows the likelihood with a 0.01 grid for Hospitals 4 and 9.

Table EC.4 Summary statistics of ED medical patient characteristics

Final study cohort: N=164,167

	Min	Max	Mean	Median	SD
Male			0.53		
Age (years)	18.00	113.00	67.27	70.00	17.59
LAPS2	0.00	294.00	74.11	70.00	37.47
EDWait (hours)	0.02	12.00	1.30	0.88	1.41

ICU admission cohort: N=19,683

	Min	Max	Mean	Median	SD
Male			0.48		
Age	18.00	111.00	64.52	67.00	17.48
LAPS2	0.00	294.00	105.03	102.00	45.98
EDWait	0.02	11.98	1.36	0.90	1.45

Note: LAPS2 is severity of illness score. EDWait corresponds to the ED boarding time

Table EC.5 System summary statistics by hospital

Hosp	Num. of days	Num. of hospitalizations	Q_l	Q_h	B	ICUOccu	$\lambda_{Q,l}$	$\lambda_{Q,h}$	λ_E	μ_I	$\Pr(a_l)$	$\Pr(a_h)$
1	667	11,676	11	5	21	0.67	1.230	0.221	0.252	0.035	0.12	0.41
2	500	9,902	14	7	26	0.76	1.429	0.227	0.268	0.026	0.11	0.36
3	667	8,039	8	4	12	0.49	0.859	0.146	0.101	0.030	0.05	0.22
4	667	14,595	13	7	31	0.71	1.529	0.327	0.519	0.031	0.05	0.29
5	576	5,082	7	3	11	0.65	0.633	0.098	0.108	0.030	0.11	0.41
6	667	10,577	9	5	21	0.58	1.123	0.208	0.278	0.036	0.08	0.36
7	578	4,915	7	4	11	0.71	0.583	0.111	0.188	0.030	0.05	0.18
8	653	8,400	10	6	16	0.67	0.868	0.185	0.158	0.033	0.14	0.40
9	514	12,355	15	5	22	0.71	1.807	0.264	0.354	0.036	0.07	0.31
10	609	5,978	8	4	12	0.52	0.714	0.092	0.228	0.048	0.06	0.28
11	667	2,655	6	4	7	0.55	0.284	0.045	0.056	0.029	0.12	0.46
12	388	6,751	12	6	24	0.69	1.256	0.173	0.299	0.026	0.06	0.33
13	667	8,061	9	4	16	0.50	0.817	0.182	0.110	0.028	0.07	0.30
14	667	12,841	11	6	36	0.72	1.332	0.257	0.493	0.026	0.07	0.33
15	575	7,208	8	4	16	0.43	0.897	0.111	0.114	0.031	0.07	0.31
16	667	7,190	8	4	13	0.44	0.752	0.133	0.108	0.033	0.06	0.26
17	548	3,511	7	4	9	0.62	0.418	0.099	0.066	0.024	0.09	0.32
18	667	7,702	9	6	32	0.58	0.796	0.122	0.557	0.034	0.06	0.26
19	547	7,476	8	4	25	0.46	0.966	0.170	0.196	0.031	0.10	0.37
20	666	5,096	8	4	11	0.34	0.545	0.087	0.065	0.035	0.07	0.29
21	333	2,109	8	4	13	0.68	0.451	0.076	0.139	0.025	0.10	0.44
22	333	2,048	8	4	16	0.62	0.422	0.090	0.182	0.028	0.15	0.45

Note: 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 ICU admission probability for the ED patients.

Table EC.6 Estimation results of structural model by individual hospital

Hosp	Num. of periods	Discount Factor	Low Severity		High Severity		R^2
		$\hat{\beta}$	$\hat{u}_{w,l}$	$\hat{u}_{r,l}$	$\hat{u}_{w,h}$	$\hat{u}_{r,h}$	
1	8016	0.3	−0.015	1.490	−0.749	0.301	0.19
		(0.025)	(0.016)	(0.031)	(0.060)	(0.041)	
2	6012	0.5	−0.001	1.465	−0.659	0.519	0.23
		(0.026)	(0.038)	(0.028)	(0.054)	(0.055)	
3	8016	0.1	−0.124	2.558	−1.560	1.185	0.08
		(0.052)	(0.157)	(0.052)	(0.155)	(0.067)	
4	8016	0.4	−0.065	2.307	−0.940	0.798	0.15
		(0.011)	(0.034)	(0.040)	(0.053)	(0.044)	
5	6924	0.1	−0.179	1.786	−1.147	0.336	0.10
		(0.036)	(0.096)	(0.053)	(0.117)	(0.076)	
6	8016	0.2	−0.077	1.904	−0.970	0.531	0.14
		(0.041)	(0.039)	(0.045)	(0.080)	(0.048)	
7	6948	0.9	−1.755	2.687	−1.876	1.453	0.10
		(0.177)	(0.515)	(0.083)	(0.328)	(0.096)	
8	7848	0.1	−0.249	1.446	−0.764	0.380	0.15
		(0.062)	(0.111)	(0.034)	(0.089)	(0.051)	
9	6180	0.5	0.000	1.745	−0.789	0.707	0.18
		(0.035)	(0.036)	(0.040)	(0.062)	(0.050)	
10	7320	0.1	−0.629	2.339	−1.630	0.929	0.07
		(0.325)	(0.800)	(0.053)	(0.466)	(0.086)	
11	8016	0.9	−1.127	1.774	−1.366	0.104	0.08
		(0.248)	(0.538)	(0.064)	(0.274)	(0.108)	
12	4668	0.6	−0.099	2.121	−0.558	0.637	0.22
		(0.014)	(0.033)	(0.051)	(0.058)	(0.071)	
13	8016	0.1	−0.044	2.229	−1.260	0.785	0.13
		(0.016)	(0.051)	(0.045)	(0.110)	(0.055)	
14	8016	0.3	−0.147	1.951	−0.956	0.662	0.19
		(0.050)	(0.063)	(0.036)	(0.082)	(0.044)	
15	6912	0.1	−1.082	2.155	−2.275	0.762	0.10
		(0.071)	(0.180)	(0.045)	(0.219)	(0.076)	
16	8016	0.1	−0.199	2.319	−1.233	1.005	0.11
		(0.000)	(0.072)	(0.050)	(0.123)	(0.066)	
17	6588	0.9	−1.669	2.065	−1.530	0.660	0.09
		(0.082)	(0.202)	(0.069)	(0.134)	(0.081)	
18	8016	0.4	−0.154	2.358	−1.391	1.005	0.04
		(0.046)	(0.065)	(0.060)	(0.114)	(0.069)	
19	6576	0.1	−0.886	1.809	−1.885	0.517	0.12
		(0.000)	(0.069)	(0.039)	(0.138)	(0.056)	
20	8004	0.3	−0.116	2.260	−1.335	0.862	0.09
		(0.043)	(0.054)	(0.063)	(0.131)	(0.084)	
21	4008	0.7	−0.135	2.002	−0.612	0.240	0.12
		(0.047)	(0.058)	(0.094)	(0.096)	(0.120)	
22	4008	0.3	0.000	1.513	−0.979	0.175	0.09
		(0.042)	(0.030)	(0.068)	(0.130)	(0.111)	

Note: The second column reports the number of two-hour periods in each hospital, and the last column provides the McFadden's pseudo R^2. Standard errors are reported in parenthesis. They are estimated by bootstrapping with 500 trials. We find that the standard errors of the estimated $\hat{\beta}$'s are smaller than 0.05 (resp. 0.1) for 15 (resp. 19) out of the 22 hospitals. Although not reported here, in the bootstrapping, we find that the estimated $\hat{\beta}$ remains the same or only changes to the adjacent levels (plus or minus 0.1) in more than 90% of the trials for 20 of the 22 hospitals (except Hospitals 14 and 15).

References

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.

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.

Song, Hummy, Anita L Tucker, Ryan Graue, Sarah Moravick, Julius J Yang. 2020. Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science* **66**(9) 3825–3842.