Capacity Management in Networks: A Structural Estimation Approach for Hospital Inpatient Wards

Jing Dong
Columbia Business School, New York, NY 10027, jing.dong@gsb.columbia.edu

Pengyi Shi
Krannert School of Management, Purdue University, West Lafayette, IN 47907, shi178@purdue.edu

Fanyin Zheng
Columbia Business School, New York, NY 10027, fanyin.zheng@columbia.edu

Xin Jin
National University Health System, Singapore, Singapore 119228, xin_jin@nuhs.edu.sg

Problem Definition: Addressing capacity management within a multifaceted network of resources is a critical challenge in service operations management. As resources are often shared to serve multiple classes of customers in such systems, customer routing often depends on the congestion levels of these resources, which in turn are affected by customer routing, creating a feedback loop. Consequently, when evaluating the impact of substantial capacity changes in the network, one must account for the complexity resulting from resource sharing, endogenous routing policies, as well as the feedback loop between routing policies and congestion levels which has the potential to alter the equilibrium of the system. This complexity renders conventional approaches insufficient for an accurate assessment. **Methods/Results**: To tackle these challenges, we develop a structural estimation approach that relies on two key components. First, we estimate the routing policy via a choice model that allows the routing policy to depend on not only the focal resource's utilization but all connected resources' utilization. We adopt a control function approach with instrumental variables to estimate the loads' effect in routing without bias. Second, we incorporate the estimated routing policy into a queueing network model which captures the detailed system dynamics and evaluates the equilibrium performance of the entire network. We apply our approach to the specific empirical setting of the hospital inpatient ward network. We show that our proposed approach outperforms two alternative models and highlight the importance of accounting for network equilibrium effects when evaluating substantial capacity changes. **Managerial Insights**: We provide prescriptive capacity allocation recommendations to hospital managers. More generally, our findings underscore the importance of a comprehensive understanding of the interdependencies between customer routing decisions and the levels of congestion present at various resources, shedding light on broader strategies for improving the operational performance of service networks.

Key words: network capacity management; structural estimation; inpatient patient flow; empirical operations management.

1. Introduction

Capacity management plays a central role in the operation of service systems. In many industries, as demand and other financial and operational factors may change dramatically over time, managers often need to make significant capacity adjustments accordingly. For complex service systems such as hospitals, fulfillment centers, consulting companies, etc., evaluating the impact of large, instead of marginal, capacity changes on the performance of the overall system is particularly challenging for several reasons. First, as various degrees of resource sharing are common in those systems, changing the capacity of one type of resource might have a substantial impact on the performance associated with the other resources. Second, because the routing of customers or demand in such systems often depends on the congestion levels at various resources, changing capacity might lead to changes in routing decisions. Lastly, changes in congestion levels and changes in routing decisions might create a feedback loop that leads to a new system equilibrium. As a result, when evaluating the impact of substantial capacity changes in the network, it is imperative to consider the complexity arising from the network effect characterized by resource sharing, endogenous routing decisions, and the feedback loop which makes routing decisions and congestion levels at various resources interdependent. Traditional empirical methods focusing on marginal and static analysis are inadequate in capturing these intricacies and are thus insufficient for providing an accurate assessment of the impact of substantial capacity changes.

In this paper, we develop an integrated framework to evaluate the impact of substantial capacity changes in complex service systems. Specifically, we take into account the network nature of the system and the interdependence of customer routing on the heterogeneous workloads across different types of resources using a structural estimation approach based on a queueing network model. We study the hospital inpatient ward network as the specific empirical setting and provide prescriptive policy recommendations to hospital managers on the capacity allocation problem they face. More importantly, our findings highlight the importance of accounting for network effects when analyzing capacity management problems for complex service systems.

As demand for healthcare continues to increase, many hospitals across countries experienced substantial capacity expansions, especially in their inpatient wards, in recent years (Pallin et al. 2014). The ability to evaluate the impact of capacity change on the hospital's key performance metrics before implementing such large-scale expansion is, therefore, essential to hospital managers. However, the complex network nature of the inpatient wards makes the problem particularly challenging. Although the primary resource in the system–beds–are organized by medical specialties, resource sharing through off-service placement happens daily in most hospitals (Song et al. 2019). Once a patient arrives, the bed management team balances the workloads of all relevant units in the system and might route the patient to an off-service ward when their primary ward is heavily loaded. In other words, the patient routing policy depends on the real-time workload of various units in the system as well as the bed management team's preference when trading off off-service placement and congestion. As a result, evaluating the impact of capacity expansion in the inpatient ward network needs to account for the shared resources and endogenous routing decisions, and their resulting feedback loop which might change the equilibrium of the system.

We partner with a large teaching hospital in Southeast Asia and use their detailed patient-level data to tackle this problem. At the time, the hospital was planning to expand the inpatient ward capacity by building a new wing. Our study provides prescriptive policy recommendations to this specific hospital in evaluating the impact of such expansion on the system's key performance metrics and how such impact might vary across their preferred scenarios of capacity allocation across specialties. More importantly, we provide an empirical framework that combines structural estimation with a queueing network model to evaluate significant capacity changes to equilibrium system performance and highlight the importance of considering the complex network effects in providing unbiased managerial recommendations.

Our approach has two main components. First, we estimate the patient routing policy via a choice model. This allows the routing policy to depend on an extensive set of factors that might affect the bed management team's routing decision in practice. More importantly, it allows the routing policy to depend on not only the focal specialty's primary unit's occupancy, but all connected units' occupancy. To estimate the impact of occupancy on the routing decision without bias, we adopt a control function approach with instrumental variables. Our estimation approach takes into account that occupancy might be correlated with unobserved patient severity and other factors not captured in the data and the instruments can correct for such endogeneity. Our estimation results show that the occupancy of focal and connected units has a substantial impact on patient routing decisions. In other words, load balancing is of significant consideration when the bed management team decides which unit to place a patient in. The load-balancing behavior creates connectivity among different inpatient wards in the network.

Second, we plug the estimated patient routing policy into a queueing network model of patient flow (Dong et al. 2019), which allows us to capture the detailed system dynamics and evaluate the changes in capacity on the entire network in equilibrium. We show that our model output matches various key performance metrics in the data. More importantly, in the counterfactual analyses where we evaluate the capacity expansion scenarios of interest, we show that our model outperforms two existing alternative methods–one which only allows the routing policy to depend on the focal unit's occupancy (instead of all connected units' occupancy) and the other where the equilibrium network effect is not captured–in two important ways. First, our model predicts much more realistic results about the changes to the occupancy level and the number of high occupancy periods for both the focal unit and the connected units. The alternative models, by contrast, produce unrealistic results in both the direction and the magnitude of the changes. Second, we show that using alternative models might also lead to biased qualitative policy recommendations in terms of to which unit the additional capacity should the hospital allocate given their objective of reducing the overall congestion in the system.

Our study generates important insights about capacity management for hospitals and, more broadly, for complex service system operations. In addition to providing prescriptive capacity allocation recommendations to managers, a key takeaway from our study is the importance of a comprehensive understanding of the interdependencies between customer routing decisions and the congestion levels at various resources within the network. Our study offers a holistic examination of the underlying dynamic feedback loop, where changes in routing decisions impact congestion, and congestion, in turn, alters routing decisions. Accounting for this feedback loop is vital for providing an accurate evaluation of substantial capacity changes and making informed capacity allocation decisions. These insights extend beyond the specific context of hospital networks and shed light on broader strategies for improving operational performance in other service industries. Importantly, our study provides an integrated framework to capture the network effects and the dynamic relationships between routing decisions and congestion at various resources in the network.

1.1. Related Literature

Our work contributes to the literature that studies complex service and manufacturing networks. Wallace and Whitt (2005) study skill-based routing in call centers which reflects similar network structure and routing considerations to our setting. Shi et al. (2019) focus on the design of networks in manufacturing systems. DeValve et al. (2023) study the value of flexibility in a fulfillment network. While most of the related work in this literature takes a modeling perspective, our work complements the existing work by illustrating the importance of network effects using data, providing empirical evidence to motivate the use of complex network models, and developing an empirical framework to study such problems.

Within the application context of our study, our work is related to the growing body of literature on hospital capacity management. For long-term decisions, Green (2002), Gupta and Potthoff (2016), Pinker and Tezcan (2016), and Kim et al. (2023) study how to optimally allocate capacity and design wards. Best et al. (2015) compare pooled versus dedicated ward designs and point out several strengths and weaknesses in the two designs. On a shorter time scale, Freeman et al. (2017) and Sun et al. (2018) study dynamic admission control and patient routing. Our paper is most closely related to Bertsimas and Pauphilet (2023), Dai and Shi (2019), Helm and Oyen (2014), Samiedauluie et al. (2017), all of which study admission control or patient routing in inpatient ward networks. Aimed at deriving the optimal routing policy, these studies typically impose a specific objective function for the decision maker. In contrast, we acknowledge the complex objective of the decision maker in practice and take an empirical approach to estimate the decision rule. More importantly, we highlight the role of network effects and demonstrate the importance of correctly capturing the effects in providing accurate capacity management recommendations.

Our work is also closely related to the literature on empirical behavioral queueing and behavioral healthcare operations management. Song et al. (2015) empirically investigate the effect of queueing configuration on the productivity of ED physicians. Green et al. (2013) study how anticipated workload affects the nurses' absenteeism behavior. Kim et al. (2019) use behavioral models and controlled experiments to study admission control biases in ICU capacity management. Ding et al. (2019) estimate the effect of waiting on ED patient scheduling. Batt and Terwiesch (2017) and Soltani et al. (2019) study various load-adaptive behaviors in ED and their implications for patient outcomes. Other examples in the broader area of empirical behavioral operations management include Buell and Norton (2011), Mandelbaum and Zeltyn (2013), and Yu et al. (2017). Our work complements this literature by empirically investigating load-adaptive patient routing policy in managing a network of inpatient wards.

Our work also contributes to the literature on empirical operations management using structural estimation methods. To establish the causal effect of occupancy on patient routing, we use a control function approach to account for the endogeneity of occupancy. This method has been applied to solve endogeneity problems in estimating consumer choice models in economics and operations management (Arıkan et al. 2018, Guajardo et al. 2012, Petrin and Train 2010). More relevant to our particular setting, when studying retrospectively collected patient-flow data, it is important to account for the endogeneity caused by unobserved patient characteristics such as severity. In this context, occupancy is often used to construct instrumental variables (see, for example, KC and Terwiesch (2012), Kim et al. (2015), Song et al. (2019)). The rationale is that the determinants of the admission decision for an individual patient are unlikely to be correlated with the occupancy level. However, our work provides evidence that this assumption may be invalid in some settings and that empirical researchers should use occupancy-based IVs with caution in such settings.

2. Empirical Setting and Data

In this section, we describe the setting of our study and the data we use for the analyses. We start by providing an overview of the inpatient ward network in our partner hospital. Then, we discuss our research setting and provide some preliminary data analyses to motivate our model.

2.1. Overview of Inpatient Ward Operations

The inpatient ward network of our partner hospital has 13 wards serving patients from eight medical specialties: Cardiology (Cardio), Surgery (Surg), Orthopedics (Ortho), Respiratory disease (Respi), Gastroenterology and endoscopy (Gastro), Renal disease (Renal), General Medicine (GenMed), and Neurology (Neuro). Each ward contains around 40 beds. We index the wards from W1 to

W13. The hospital employs a focused-care model. Figure 1 shows the specialty-ward mapping, with circles representing wards and rectangles representing specialties. Note that each ward is designated to serve patients from one specialty (referred to as dedicated wards) or two specialties (referred to as shared wards). Each designated ward is a primary ward for the corresponding specialty or specialties. For shared wards, there is a nominal allocation of beds assigned between the two specialties, and are thus considered as the primary wards for the two specialties.

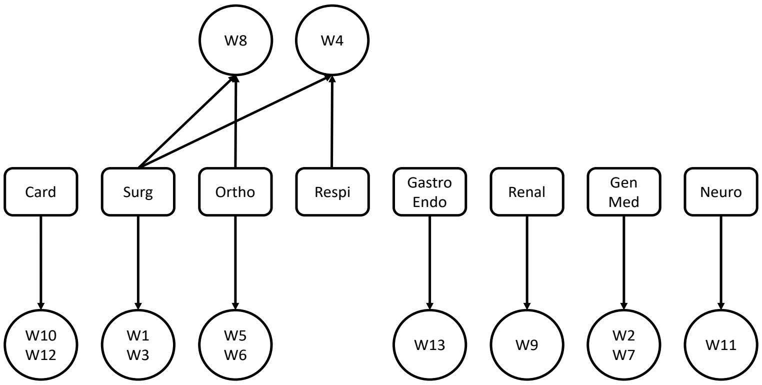


Figure 1 Specialty-ward assignment.

In our partner hospital, all bed assignments, including bed requests for both ED and non-ED patients, are managed by a central bed management team. When the team receives a bed request, one team member will start searching for an appropriate bed, from either a primary or a non-primary ward, and make a tentative bed allocation. After the bed allocation is confirmed, arrangements are made to admit the patient to the allocated bed.

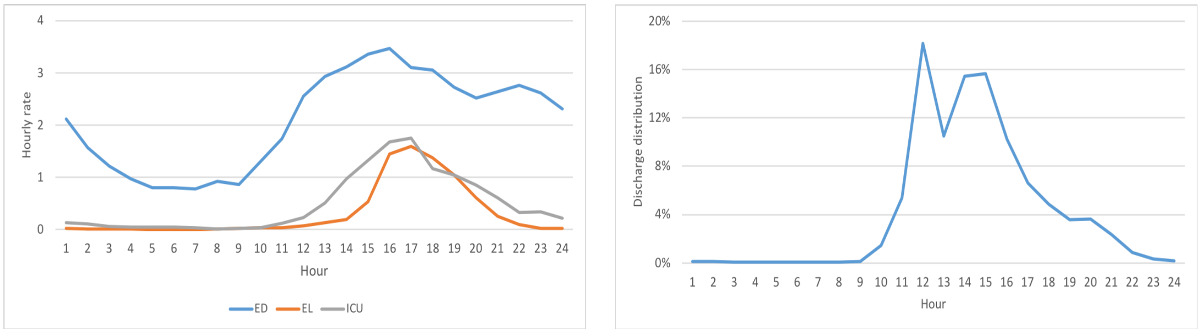
Table 1 summarizes the daily admission rate from each specialty (Λ_s), the average length of stay ($\mathbb{E}[LOS_s]$), and the off-service proportion of each specialty–i.e., the share of patients placed in a ward that is not a primary ward for the patient's specialty. We observe that different specialties receive different volumes of patients, as measured by $\Lambda_s\mathbb{E}[LOS_s]$. Cardio, GenMed, and Surg have the highest volume of patients. We also note that the hospital has a significant proportion of patients placed off service overall, 22%, which gives rise to the network connectivity among different inpatient wards. In addition, there are variations in the off-service proportions across different specialties. Since Cardio and GenMed have the highest volume of patients as well as a high off-service proportion, our partner hospital was considering allocating the new wing to either or both specialties. As we will discuss in Section 2.2, given Gastro serves as the main off-service ward to both Cardio and GenMed, Gastro is also considered for receiving the additional capacity in the new wing.

The patients of each specialty are admitted from three main sources: the emergency department (ED), elective admissions (EL), and transfers from the intensive care units (ICU). We refer to them

	Cardio	Ortho	Surg	Gastro	GenMed	Neuro	Renal	Resp	Overall
Λ_s	19.8	13.2	17.5	8.2	17.8	5.9	6.2	4.6	93.2
$\mathbb{E}[LOS_s]$	3.7	4.4	3.6	3.6	4.5	3.7	4.5	4.0	4.0
Off-service %	26%	6%	12%	33%	27%	54%	26%	14%	**22%**

Table 1 Summary of workload-related statistics.

as ED, EL, and ICU transfer patients for the rest of the paper. Figure 2a shows the number of patient arrivals (to the inpatient wards) in each hour from ED, EL, and ICU transfer. The arrival rates are estimated using the bed-request timestamp for ED patients; since we do not observe the bed-request timestamps for EL and ICU transfer patients, we use their admission timestamps instead. We can see that the arrival rates are time-varying, and the patterns for ED and non-ED patients are substantially different. In particular, EL admissions are pre-scheduled, so they arrive mainly in the afternoon and are admitted in clusters when beds become available–i.e., after the block discharges which peak between noon and 4pm, as shown in Figure 2b. Similar observations apply to ICU transfer patients. The arrivals of ED patients, on the other hand, are more spread out throughout the day. The exogeneity of the ED arrivals is the key to the identification of our routing model. We will return to this point and provide a more detailed discussion in Section 3.



(a) Hourly arrival rate

(b) Discharge time distribution

Figure 2 Hourly rates of arrivals by patient sources and hourly discharge rate. The arrival rates are estimated using the bed-request timestamps for ED patients and the admission timestamps for non-ED patients (no bed-request time for the latter). The numbers on the x-axis denote the hourly interval; e.g., 2 denotes the interval of 1-2am.

Since the hospital operates in a highly non-stationary environment, we divide each day into two periods: daytime, from 7:00 am to 9:00 pm; and nighttime, from 9:00 pm to 7:00 am the next day for the estimation of the patient routing policies. There are two reasons for this partition. First, the majority of patients are discharged by 9:00 pm, and there are almost no discharges between 9:00 pm and 7:00 am the next day (see the discharge time distribution in Figure 2b). Therefore, the bed management team's decision in the nighttime is not affected by patient discharges. Second, the majority of non-ED admissions happen before 9:00 pm, i.e., there are almost no EL or ICU transfer admissions between 9:00 pm and 7:00 am the next day (see the arrival rates plot in Figure 2a). This suggests that the bed management team's decision in the daytime and nighttime also differs in terms of the patient sources.

2.2. Connectivity of Inpatient Ward Network

The off-service placement creates connectivity among different inpatient wards which gives rise to the network structure of our system. In this section, we present some descriptive evidence to highlight the importance and complexity of the network. We focus on the three largest specialties: Cardio, GenMed, and Surg, and a highly connected specialty Gastro. As illustrated in Table 1, in addition to a large volume of patients, Cardio and GenMed also have high off-service placement proportions. We further restrict our study to ED admissions since EL and ICU transfer admissions are often scheduled in advance. The real-time decision making of the bed management team plays a much less important role in the routing of these patients.

Table 2 presents the share of ED admissions assigned to each ward for the four specialties. We observe that Cardio and GenMed both utilize W13, which is a primary ward for Gastro, for off-service placements. In addition, W8 and W9, which are two Ortho wards, are commonly used as the off-service wards for all specialties. These off-service placements balance the workload and create strong connectivity across different inpatient wards. As we will demonstrate later in the paper, the change in the capacity of one ward will lead to non-trivial changes in the occupancy level of the other connected wards.

Cardio		GenMed		Surg		Gastro	
W12	47%	**W2**	49%	**W3**	44%	**W13**	69%
W10	12%	**W7**	26%	**W1**	22%	W9	6%
W13	13%	W13	7%	**W4**	12%	W8	5%
W8	6%	W9	4%	W5	10%	W2	4%
W9	4%	W4	3%	W8	6%	W4	3%

Table 2 Proportions of ED admissions to top 5 wards for Cardio, GenMed, Surg, and Gastro. The wards in bold are the primary wards for the corresponding specialty.

Table 3 shows the occupancy at midnight, 6am, 11am, 3pm, and 8pm. We report the occupancy of six wards, which include both the primary wards and the major off-service wards for the four specialties. We observe that the primary wards for GenMed (W2), Cardio (W12), and Gastro (W13) have relatively high occupancy compared to the other wards. The occupancy for the Ortho wards (W8, W9) is particularly low, which partly explains why they are commonly used as the off-service wards by the bed management team. On the other hand, even though W13 has a high occupancy, it is also used heavily by Card and GeMed as the off-service ward. This suggests that the bed management team's objective is complex, and load balancing might be one of many considerations. In addition, we see that the wards tend to have much higher occupancy at 6am (near the beginning of daytime) and midnight than at 3pm and 8pm (near the beginning of nighttime). This is because most of the discharges are clustered between 11am and 5pm (Figure 2).

	W2	W3	W8	W9	W12	W13
midnight	92%	88%	75%	86%	96%	93%
6am	95%	91%	78%	89%	97%	95%
11am	90%	85%	76%	88%	93%	92%
3pm	84%	82%	70%	82%	88%	84%
8pm	88%	86%	72%	84%	94%	89%

Table 3 Ward occupancy at different times.

Next, we provide some descriptive evidence of the complex nature of the inpatient ward network. Table 4 shows the proportion of ED patients from Cardio and GenMed that are routed off service under three different primary ward occupancy levels: less than 90%, between 90% and 95%, and above 95%. These occupancy levels are measured at midnight, i.e., the beginning of each day, and the off-service proportion is calculated for the corresponding day. We observe that the off-service proportion increases as the primary ward occupancy increases, which highlights that the connectivity across wards in the network depends on the occupancy of the primary ward. It also suggests that load balancing is likely an important consideration in the bed management team's decision.

	Cardio	GenMed
< 0.9	19%	11%
$0.9 - 0.95$	24%	19%
> 0.95	40%	31%

Table 4 Average off-service proportion under three categories of primary ward occupancy levels: below 90%, between 90% and 95%, and above 95%.

In Table 5, we show how the proportion of Cardio and GenMed patients placed in different wards changes with the occupancy of W13, which is a commonly used off-service ward for Cardio and GenMed. We note that as the occupancy of W13 increases, the proportion of patients placed in W13 decreases, but we place a higher proportion of patients in the other off-service wards. These analyses suggest that patient routing decisions not only depend on the primary ward occupancy but also the occupancy of other connected wards, further highlighting the complex connectivity of the inpatient ward network.

Table 5 Overflow decomposition when primary unit occupancy is above 95%. The numbers reported in each parenthesis follow the format of "(primary, W13, other off-service wards)".

	Cardio	GenMed
W13 < 0.9	(63.4%, 14.4%, 22.1%)	(74.5%, 9.2%, 16.2%)
W13 $0.9 - 0.95$	(60.1%, 13.7%, 26.2%)	(70.2%, 7.5%, 22.2%)
W13 > 0.95	(59.0%, 12.2%, 28.8%)	(67.8%, 5.9%, 26.3%)

2.3. Road Map of Model and Empirical Analysis

As the descriptive evidence presented in Sections 2.1 and 2.2 illustrates, there are two important features of the inpatient ward system: the inpatient wards are connected as a network, and the routing of patients potentially depends on the occupancy of all connected units. As a result, we combine a structural estimation approach which allows us to estimate patient routing policy from data with a queueing network model that captures the equilibrium network effects. In what follows, we first describe the estimation of the patient routing policy in Section 3 and present the estimation results in Section 4. Then, in Section 5, we describe our stochastic network model and demonstrate its performance. Finally, we present the counterfactual analysis in Section 6.

3. Model and Estimation of Patient Routing

We adopt the structural estimation approach and model each ED patient's routing decision using a discrete choice model. We believe the model summarizes the key aspects of patient routing decisions and serves the objectives of our study for several reasons. First, the model allows us to identify and estimate the causal effect of the occupancy of different units in the network on the routing decision. Second, the model does not impose a specific objective function for the routing decisions but learns the main determinants of the decisions from data. Third, as the results in Section 6 demonstrate, our estimated model fits the data well in predicting key system performance metrics. Finally, it allows us to evaluate different capacity allocation policies in the counterfactual and provide prescriptive policy recommendations to hospital managers.

3.1. Structural Model

We describe the structural model in this section. Since the model is specialty-specific–i.e., all variables and parameters in the model are estimated for each specialty separately–we ignore the specialty subscript for brevity. We also model the daytime (7:00 am to 9:00 pm) and nighttime (9:00 pm to 7:00 am the next time) patient routing separately. Let U_{ijt} denote the decision maker's utility of routing patient i to ward j at time t. The decision maker in our setting is the bed management team. U_{ijt} takes the form

$$U_{ijt} = \beta_{0j} + \beta_1 O_{jt} + \beta_{2j} X_i + \beta_{3j} V_{S_j d_t} + \beta_{4j} Y_{d_t} + \epsilon^1_{ijt} + \epsilon^2_{ijt}, \tag{1}$$

where O_{jt} is the occupancy of ward j an hour before the start of the period t. We divide the day into four periods: 7:00 am to 12:00 pm; 12:00 pm to 4:00 pm; 4:00 pm to 9:00 pm; 9:00 pm to 1:00 am; and 1:00 am to 7:00 am. The ward occupancy is measured at 6:00 am, 11:00 am, 3:00 pm, 8:00 pm, and midnight on each day. For example, for a patient who is admitted between 7:00 am and noon, O_{jt}'s measure the occupancy of the wards at 6:00 am on that day. The time periods and occupancy observations are defined for the following reasons. First, the bed management team may not have the occupancy information nor the mental bandwidth to keep track of it in real time at a high frequency. Therefore, it is not reasonable to assume the patient routing decision is made based on real-time occupancy information. Based on our conversation with the bed management team and common in similar settings in other hospitals, the team pays more attention to occupancy levels at several routine times during the day: 6am is the time the inpatient wards experience the highest occupancy; 11am is the time when morning rounds complete, and some discharge decisions might have been made; 3pm is the time after the largest batch of discharges start during the day; 8pm is when the night shifts start, and discharges and non-ED admissions rarely happen after that; Midnight is when the daily hospital census is measured. Therefore, we assume that the bed management team keeps track of the occupancy levels of the inpatient wards at those five times of the day which divide each day into five time periods. The patient routing decisions during each time period are based on the occupancy at the beginning of that period. We acknowledge that we are making an assumption about the behavior of the bed management team, but we believe it is a reasonable assumption as it closely mimics the behavior of the bed management team in practice. In addition, we estimate the model with alternative time period definitions, and the estimation results do not change qualitatively. To better capture the bed management team's potential behavior of balancing workload across wards, we set O_{jt} to be 0.8 when the observed occupancy is at or below 80%, while O_{jt} takes on the value of the observed occupancy if it is above 80%, as the difference between 20% and 50% occupancy is unlikely to affect the routing decision. We also perform the estimation using alternative high occupancy definitions, and our main results do not change. X_i is a vector of patient-specific characteristics that include patient i's triage score, age, and gender. S_j is the set of specialties that could potentially route patients to ward j. $V_{S_j d_t}$ is a vector of non-ED patient admissions of specialties in S_j–i.e., the EL patients and the patients transferred from the ICU–of specialty $s \in S_j$ in the current day, and d_t is the day corresponds to period t. Since the EL and ICU transfer patients are typically scheduled in advance, it is reasonable to assume that $V_{S_j d_t}$ is known to the bed management team in advance. $V_{S_j d_t}$ is included in daytime estimation only. We exclude the non-ED admission variables from the nighttime estimation since these admissions rarely happen during the night. Y_{d_t} is a binary vector indicating the day of the week for the day d_t. ϵ^1_{ijt} denotes unobserved patient-, ward-, and time-specific characteristics that affect the routing decision, and, more importantly, are potentially correlated with O_{jt}. ϵ^2_{ijt} captures the idiosyncratic determinants of the routing decision and is i.i.d. extreme value distributed.

Parameter β_{0j} is the baseline utility of admitting the focal specialty patients to ward j. β_1 is the coefficient that governs the effect of occupancy on the patient routing decision–i.e., it captures the load-balancing behavior. β_{0j}, β_{2j}, β_3, and β_4 are other (vectors of) unknown coefficients to be estimated.

Let C_i denote the set of wards that can potentially accept patient i, which depends on the specialty of patient i. Based on the utility function, the probability of admitting patient i into ward j in period t is

$$P_{ijt} = \frac{\exp(\beta_{0j} + \beta_1 O_{jt} + \beta_{2j} X_i + \beta_{3j} V_{S_j d_t} + \beta_{4j} Y_{d_t} + \epsilon^1_{ijt})}{\sum_{k \in C_i} \exp(\beta_{0k} + \beta_1 O_{kt} + \beta_{2k} X_i + \beta_{3k} V_{S_k d_t} + \beta_{4k} Y_{d_t} + \epsilon^1_{ikt})}. \tag{2}$$

3.2. Endogeneity of Occupancy

Although the admission decision is made at the patient level and we allow for ward-specific effect through β_{0j}'s, estimating the model directly could still lead to biased results due to the endogeneity of occupancy. The occupancy is endogenous for two reasons. First, ϵ^1_{ijt} captures patient- and ward-specific unobserved characteristics that affect the routing decision. Most noticeably, the bed management team may have a preference for admitting more severe patients to the primary ward (Song et al. 2019). These patient-ward-specific unobserved characteristics or preferences are also correlated with ward occupancy. For example, since the primary ward of the specialty we are analyzing is more likely to have a higher occupancy than the non-primary wards, the preference for admitting more severe patients into the primary ward is positively correlated with the ward's occupancy. In other words, O_{jt} and ϵ^1_{ijt} are positively correlated, which introduces a positive bias if we estimate the model without taking the endogeneity problem into account. To be more specific, since the load-balancing effect represents a negative coefficient of O_{jt}, the positive bias of the unobserved patient-ward preference implies that the absolute value of the coefficient of O_{jt} is underestimated. The bias can also be interpreted as a classic selection bias in econometrics: More severe patients are selected for admission to the primary wards when the occupancy levels of those wards are higher. As a result, without properly dealing with the endogeneity problem, we may underestimate the size of the load-balancing behavior due to selection bias.

Second, ϵ^1_{ijt} also contains time-specific unobserved features that affect the routing decision and, thus, can bias the estimation result. For example, many specialties admit a substantial number of EL and ICU transfer patients. The bed management team has to decide how to allocate the limited number of beds over multiple types of patients. Some of the EL and ICU transfer patients may have priority over the ED patients for beds in the primary wards due to their higher severity or scheduled admission. Although we observe the number of EL and ICU admissions, their severity is not observed. Since these patients compete for the same set of beds as the ED patients, their unobserved severity is negatively correlated with ϵ^1_{ijt} and positively correlated with O_{jt}. In other words, this unobserved severity in EL and ICU transfer patients leads to a negative correlation between ϵ^1_{ijt} and O_{jt}. As a result, without properly taking into account the second type of unobservable, we may overestimate the magnitude of the dynamic load-balancing effect–i.e., the absolute value of the coefficient of O_{jt}.

To summarize, ϵ^1_{ijt} contains two sources of unobserved determinants of the routing decision. These unobservables introduce two types of biases of opposite directions in the estimation of β_1. It is a priori not clear which type of bias has a larger effect.

3.3. Instrumental Variables and Control Function

We propose a solution to the endogeneity problem using instrumental variables (IVs) and the control function approach. At a high level, we need the instrumental variables to provide exogenous variation in O_{jt}, which is not correlated with the unobserved factors ϵ^1_{ijt}. Since the model is nonlinear, instead of instrumenting O_{jt} directly, we use the control function to correct for the bias in the estimation. We describe the construction of the IVs, discuss their validity, and, finally, detail the use of the control function method.

We use the ED patient arrivals in the relevant specialties in the past 24 hours as instrumental variables. Specifically, let Z_{st} be the total number of ED admissions from specialty $s \in S_j$ during the last 24 hours before the time O_{jt} is measured. Recall that S_j is the set of specialties whose patients can potentially be admitted to ward j. We use Z_{st}, $\forall s \in S_j$, as the IVs for O_{jt}, and denote the vector $(Z_{st} : s \in S_j)$ as $Z_{S_j t}$. $Z_{S_j t}$ are valid IVs for the following reasons. First, the number of ED arrivals in the relevant specialties prior to period t is correlated with the occupancy of ward j at the beginning of period t as higher patient arrivals lead to more patient admissions and higher occupancy. This is the relevant condition, and we provide evidence in Section 4 to show that the relevance condition is satisfied. Second, the total ED arrivals in the 24-hour time window before the focal period, $Z_{S_j t}$, are unlikely to be correlated with any unobserved patient characteristics that might affect patient routing decisions in period t. This is because ED arrivals are exogenous, and the number of arrivals in the 24 hours prior to the current period is unlikely to be correlated with the focal patient's unobserved characteristics such as severity. Although we can not test that this exclusion restriction holds directly, we can provide suggestive evidence using observed patient severity levels. We run Pearson's product-moment correlation for the proportion of patients with triage score P1 (highest priority) on day t and the total number of ED-arrivals on day $t-1$, the p-value is 0.9296, suggesting the arrivals on the previous day are not correlated with the severity profile of patients on the current day. Finally, since there are eight specialties in our study, none with more than eight connected wards including the off-service wards, we have enough IVs for the occupancy of all the relevant wards. Given the network connectivity shown in Figure 1, we expect the set of relevant IVs to vary for different wards.

We apply the control function approach by following the procedure described in Petrin and Train (2010). Specifically, we model O_{jt} as follows:

$$O_{jt} = \alpha_0 + \alpha_1 Z_{S_j t} + \alpha_2 V_{S_j(d_t-1)} + \alpha_3 W_{jt} + \alpha_4 Y_t + \nu_{jt}, \tag{3}$$

where $V_{S_j(d_t-1)}$ and Y_{d_t} are defined as in Equation (1). W_{jt} are j-specific covariates that affect O_{jt}, which includes the occupancy of ward j 24 hours prior to period t and the number of patients discharged from ward j in the past 24 hours. Following the literature (Petrin and Train 2010), we assume that ν_{jt} and ϵ^1_{ijt} are jointly normally distributed. As a result, Equation (1) can be written as

$$U_{ijt} = \beta_{0j} + \beta_1 O_{jt} + \beta_{2j} X_i + \beta_{3j} V_{S_j d_t} + \beta_{4j} Y_{d_t} + \beta_{5j} V_{S_j(d_t-1)} + \beta_6 W_{jt} + \lambda \nu_{jt} + \sigma \eta_{ijt} + \epsilon^2_{ijt}, \tag{4}$$

where λ is the coefficient of the control function. η_{ijt} follows the standard normal distribution. σ is an unknown parameter to be estimated. Equation (2) then becomes

$$\begin{aligned} P_{ijt} = \int \Bigg(& \frac{\exp(\beta_{0j} + \beta_1 O_{jt} + \beta_{2j} X_i + \beta_{3j} V_{S_j d_t} + \beta_{4j} Y_t + \beta_{5j} V_{S_j(d_t-1)} + \beta_6 W_{jt} + \lambda \nu_{jt} + \sigma \eta_{ijt})}{\sum_{k \in C_i} \exp(\beta_{0k} + \beta_1 O_{kt} + \beta_{2k} X_i + \beta_{3k} V_{S_k d_t} + \beta_{4k} Y_t + \beta_{5k} V_{S_k(d_t-1)} + \beta_6 W_{kt} + \lambda \nu_{kt} + \sigma \eta_{ikt})} \Bigg) \\ & \times \phi(\eta_j) \prod_{k \neq j} \phi(\eta_k) d\eta_j d\eta_k, \end{aligned} \tag{5}$$

where ϕ is the probability density function of the standard normal distribution.

3.4. Estimation

We estimate (3) and (4) for the daytime and nighttime separately for each specialty. For specialty s, the set of possible wards for patient admission includes all wards into which patients in specialty s were ever admitted in the sample.[1] We normalize the coefficients of all variables for one of the primary wards of specialty s to be zero. Then, estimating (5) amounts to estimating a random coefficient discrete choice model with an endogenous covariate, which we deal with using the control function. We use maximum likelihood to construct the parameter estimates.

[1] We group the wards receiving fewer than 100 off-service patients from specialty s during the entire sample period into one choice of admission, defined as "other ward" for specialty s.

4. Estimation Results

In this section, we describe the estimation results. An example of the first-stage estimation is provided in Table 6, where we present the estimated coefficients that are statistically significant for a subset of the wards. The results of the rest of the wards and at other times are similar and are provided in the Appendix B. Our results show that ED arrivals from connected specialties in the previous 24 hours are positively correlated with the occupancy levels of different wards in the current period, which validates the relevance of our IVs. Note that because of off-service placement, ward occupancy can be affected by the arrival of patients from non-primary specialties. We also present in Table 6 the F-statistics which show that our IVs are strong.

	W1	W2	W3	W12	W13
ED Cardio				4.15×10^{-3} *** (0.62×10^{-3})	2.56×10^{-3} *** (0.79×10^{-3})
ED Surg	2.88×10^{-3} *** (0.78×10^{-3})		4.54×10^{-3} *** (1.09×10^{-3})		
ED Gastro	1.74×10^{-3} * (0.94×10^{-3})	2.11×10^{-3} * (1.10×10^{-3})		2.11×10^{-3} *** (0.78×10^{-3})	3.75×10^{-3} *** (1.02×10^{-3})
ED GenMed		3.89×10^{-3} ** (0.71×10^{-3})			
ED Neuro		2.67×10^{-3} *** (1.28×10^{-3})			
ED Renal					2.20×10^{-3} * (1.17×10^{-3})
ED Resp	4.93×10^{-3} *** (1.41×10^{-3})			2.53×10^{-3} ** (1.17×10^{-3})	2.62×10^{-3} * (1.52×10^{-3})
Num. of Obs.	365	365	365	365	365
F-stat	23.41	16.27	40.13	13.08	16.13

Table 6 Estimated coefficients and standard errors (in parentheses) of the IVs in the first stage for the midnight occupancy. Only statistically significant IVs are reported. $*$,$**$,$***$ indicate statistical significance at 10%,5%,1% level respectively. We also control for other covariates. See the full list of control variables in Appendix B.

4.1. Load-Balancing Behavior

The second-stage estimation results are summarized in Table 7. We present the results using a multinomial logit model without the IVs and the control function (MNL) and with the IVs and the control function for GenMed, Cardio, and Surg specialties, and for day and night separately.

We first note that the estimated coefficients for the occupancy of the wards are negative and statistically significant in all cases. This suggests that the bed management team takes the workload of the wards into account when routing patients. In other words, there is substantial load-balancing behavior in the patient routing decisions. Second, the magnitude of the load-balancing behavior is substantially higher in the nighttime than in the daytime. This finding suggests that the bed

	GenMed		Cardio		Surg	
	MNL	CF	MNL	CF	MNL	CF
Occupancy Day	-4.29***	-4.14***	-3.59 ***	-2.52 **	-1.62 **	-2.78***
	(0.57)	(0.86)	(0.66)	(1.07)	(0.59)	(0.79)
CF Day		-0.43		-1.42		1.43**
		(0.75)		(1.02)		(0.60)
Num. of Obs.	3200	3200	2005	2005	1847	1847
Log-Likelihood	-3449.8	-3446.8	-2337.9	-2336.8	-2386.2	-2382.4
Occupancy Night	-11.48***	-13.16***	-16.13 ***	-16.62 ***	-6.35 ***	-7.90***
	(0.75)	(0.96)	(1.16)	(1.62)	(0.79)	(0.95)
CF Night		1.63 *		0.74		2.63 ***
		(0.92)		(1.64)		(0.77)
Num. of Obs.	1421	1421	844	844	1043	1043
Log-Likelihood	-2193.1	-2185.7	-992.0	-991.8	-1465.4	-1458.8

Table 7 Estimated coefficients and standard errors (in parentheses) of key coefficients with and without control function by specialty ($*$,$**$,$* * *$ indicates statistical significance at 10%,5%,1% level.) We also control for other covariates in the model. See the full list of control variables in Appendix B

management team has a stronger preference for load-balancing when routing patients at night than during the day. One likely explanation for this finding is that, since discharges happen only during the day, the bed management team expects more beds to become available when routing patients during the day, and, thus, is less cautious about occupancy levels and balancing load during the day. At night, however, very few patients are discharged, and, thus, there are limited alternative resources available. Thus, the bed management team might be monitoring the primary-ward occupancy more closely at night. In addition, because the hospital is running at a lower level of resource availability at night, the bed management team may have a stronger tendency to reserve the primary ward beds for more severe patients.

To interpret the results, we calculate the average marginal effect of occupancy. If we increase one of GenMed primary wards' occupancy from 90% to 95%, the probability of admitting a patient to that primary ward decreases by 4.75% on average during the daytime and by 12.81% on average during the nighttime. Meanwhile, the probability of admitting a patient to the off-service ward W13 increases by 0.56% on average during the daytime and by 2.46% on average during the nighttime. If we increase one of Cardio primary wards' occupancy from 90% to 95%, the probability of admitting a patient to that primary ward decreases by 2.64% on average during the daytime and by 17.88% on average during the nighttime. Meanwhile, the probability of admitting a patient to the off-service ward W13 increases by 0.78% on average during the daytime and by 2.49% on average during the nighttime. If we increase one of Surg primary wards' occupancy from 90% to 95%, the probability of admitting a patient to that primary ward decreases by 3.12% on average during the daytime and 7.94% on average during the nighttime. If we increase the occupancy of W13 from

90% to 95% (W13 is a commonly used off-service ward for GenMed and Cardio), the probability of sending a GenMed patient to W64 decreases by 0.86% on average during the daytime and by 5.10% on average during the nighttime. The probability of admitting a GenMed patient to the GenMed primary ward increases by 0.49% on average during daytime and by 1.98% on average during nighttime. Meanwhile, the probability of sending a Cardio patient to W13 decreases by 1.4% on average during the daytime and by 5.24% on average during the nighttime. The probability of sending a Cardio patient to the Cardio primary ward increases by 0.76% on average during daytime and by 1.66% on average during nighttime.

Lastly, the coefficient of the control function is positive and significant for Surg throughout the day and for GenMed during the nighttime. This suggests that in these cases, the first type of bias due to the unobserved focal patient's severity is significantly bigger than the second type of bias due to unobserved factors in non-ED patients' needs. On the other hand, the coefficient of the control function is statistically insignificant for Cardio throughout the day and for GenMed during the daytime, which suggests that either the bias caused by the endogeneity of occupancy is small, or the two types of biases cancel out each other's impact on the estimated coefficient. Comparing the results of the MNL model and our model using the IVs and the control function, we find that the difference in the coefficients of occupancy can be as high as more than 40%. In other words, using the MNL model without instrumenting for occupancy can potentially lead to substantial bias.

4.2. Digression: Implication for Commonly Used Occupancy IVs

Our estimation results provide strong evidence for the load-balancing behavior in inpatient ward routing decisions. An interesting byproduct of the results is the finding that unobserved patient and system factors can play a role in such decisions, which is suggested by the significant coefficients on the control function observed in some estimations. This finding has important implications for the broader empirical healthcare research, in terms of the validity of one of the most commonly adopted identification strategies–i.e., using occupancy as the instrumental variable.

Many empirical healthcare studies focus on estimating the causal effect of certain operational decisions, such as admission delay or off-service placement, on patient outcomes. One common challenge is that unobserved patient features and system factors can potentially bias the estimation result. Many studies adopt the strategy of using occupancy to construct instrumental variables to solve the omitted variable bias (KC and Terwiesch 2012, Kim et al. 2015, Song et al. 2019). The main assumption for the validity of the occupancy-based IVs is that the system occupancy at the time of the patient's arrival is uncorrelated with any unobserved variables that could affect the operational decisions of the physician. Our estimation results, however, suggest that this assumption can be problematic in some settings.

The positive and significant coefficients of the control function found in some cases show that the occupancy of the ward is positively correlated with unobserved patient severity. The reason is, as discussed before, that when the occupancy level increases, the bed management team might select more severe patients to be admitted to the primary ward. The selection of patients based on unobserved severity invalidates the assumption under which occupancy is used for the construction of IVs. In particular, when we compare patient outcomes for those admitted to the primary ward and those who are not (driven by the change in the values of the IVs), we are studying significantly different patient subgroups with different unobserved severity levels. This introduces bias in the estimated causal effect of the admission (or waiting) decision of interest. Our empirical findings suggest that empirical researchers should use occupancy-based IVs with caution in such settings.

5. Capturing Equilibrium Behavior with Queueing Network Model

Our structural estimation reveals that ED patients' routing decision depends on the occupancy levels of all relevant units within the inpatient ward network. When capacity changes substantially, routing decisions and occupancy levels also change, and their changes interact with each other. To be more specific, as the capacity of the focal unit changes, the patient routing decisions change, thereby affecting the occupancy levels of both the focal unit and its connected units within the network. Subsequently, the updated occupancy of various units influences routing decisions. This interplay between routing decisions and occupancy continues until the system reaches an equilibrium. To evaluate the performance of the system under substantial capacity changes, one would require a network model of patient flow that captures these intricate interactions between patient routing and occupancy levels and how the effect of capacity changes on routing and occupancy propagates through the network.

We construct a queueing network model that adeptly reflects the interactions of the occupancy levels of connected units in the inpatient network and integrates our estimated routing model. The stochastic model is based on the one developed in Dong et al. (2019), which we adopt to integrate the estimated patient routing model. Our model is designed to capture the equilibrium occupancy levels of all units in the inpatient network, where the routing policies are endogenous–i.e., they depend on the occupancy levels of all relevant (connected) units in the network. We use this model to evaluate the performance of the system at equilibrium under various capacity allocation scenarios in Section 6. In what follows, we provide a brief overview of the network model and delegate details to Appendix C.

5.1. Queueing Network Model

The queueing network model contains multiple classes (different specialties and admission sources) of patients and multiple pools (different inpatient wards) of servers (beds). In particular, there are

$J = 13$ inpatient wards (server pools), where the j-th pool has N_j beds. Patients are classified into 8 medical specialties. Each specialty further has three different admission sources: ED, EL, and ICU transfer. Thus, there are $I = 8 \times 3 = 24$ patient classes in total. We model patient arrivals as time-varying Poisson processes with class-dependent rates. The LOS (service time) depends on both the patient class and the server pool (wards), which captures the potential off-service slowdown. In addition, the LOS contains an active treatment time and a discharge delay. We next elaborate on how we calibrate each of the modeling components.

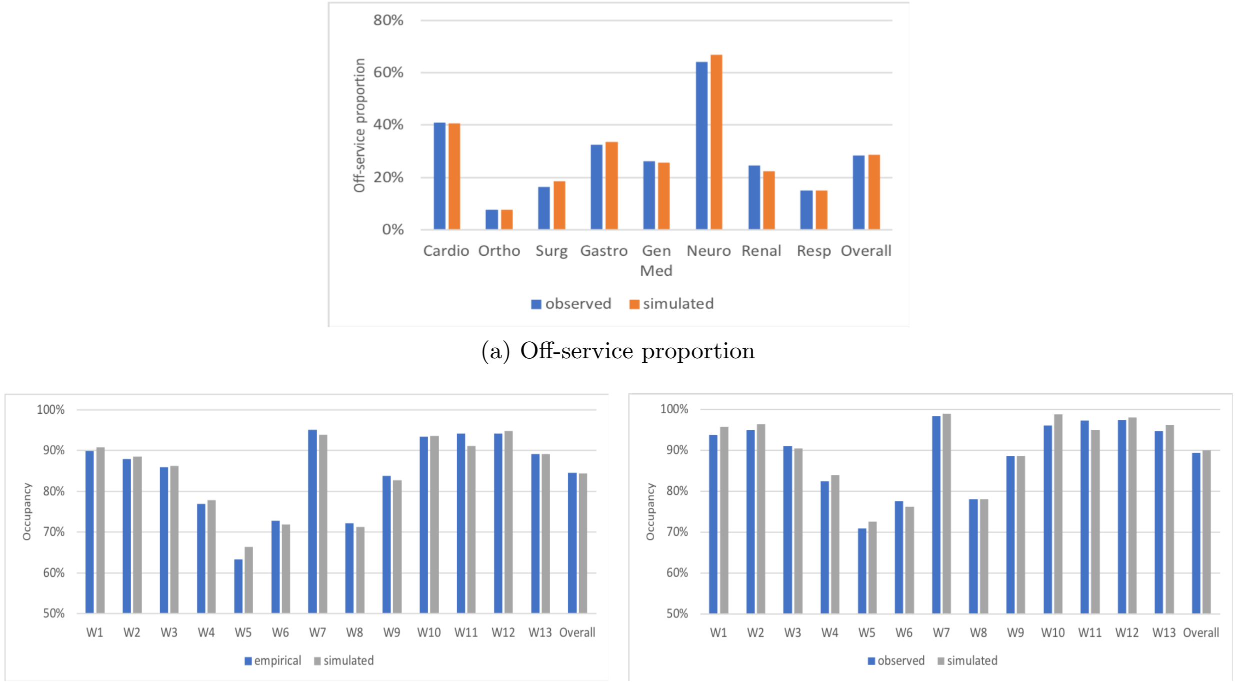
Arrivals. The arrival rates of patients are time-varying. For example, Figure 2a plots the hourly admission rates for patients from three different admission sources in our partner hospital. We model the arrival process for each patient class (specialty and admission source) as a time-varying Poisson process with a periodic piece-wise constant rate function where one piece is one hour and a period is one day. The arrival rate can be estimated directly from data.

Service Times. We model a patient's LOS in two time scales: an integer number of days, d_{los}, corresponding to the medically necessary LOS; and a real number of hours, h_{dis}, corresponding to when the patient release the bed on the day of discharge. The second part captures the need for physician inspection and discharge delays due to paperwork, transportation arrangements, etc (see Figure 2b for the hourly discharge rates). We assume there is a 17% increase in the medically necessary LOS for patients who are placed off-service based on the estimation in Dong et al. (2019) using data from the same hospital. The distribution of d_{los} and h_{los} can also be estimated directly from data.

Patient Routing. The routing decisions for patients from different admission sources are different. In general, the hospital has more control over EL and ICU transfer patients than ED arrivals. For EL and ICU transfer patients, we use a randomized routing policy where the routing probabilities are estimated using the corresponding sample proportions. For ED arrivals, we use the discrete choice model estimated in Section 3.

5.2. Calibration Results

Our model is carefully calibrated using the patient-flow data. Figure 3a shows the proportion of patients admitted to off-service wards for each specialty observed in the data and predicted by our model through simulation. Figures 3b and 3c compare the observed and simulated average occupancy for each ward at 8pm and 6am, respectively. The result shows that our model provides accurate predictions for the off-service proportions of each specialty and the occupancy of each ward. We also calculate the proportion of assignments to each ward for daytime and nighttime of different specialties; see Table 17 in Appendix C. Above all, our model enjoys high fidelity and is able to match key performance metrics well.



(a) Off-service proportion

(b) Ward occupancy at 8pm

(c) Ward occupancy at 6am

Figure 3 **Simulation calibration results.**

6. Counterfactuals

In this section, we use our estimated model to calculate key system performance metrics under different capacity expansion scenarios. The goal is to provide prescriptive policy recommendations to hospital managers. Compared with existing empirical work on patient routing in the literature, our approach treats inpatient wards as a network instead of individual units, which has two main advantages when evaluating system structural changes such as major capacity expansion. First, the patient routing decision in our model depends on the occupancy of all connected units in the network. This is in contrast to the simple regression-based models in the existing literature, where the routing decision depends on the occupancy of the focal unit only (see, e.g., Kim et al. (2019), Song et al. (2019)). Second, our model takes into account the equilibrium network effect of the capacity change in the focal unit on the entire network through the interactions between the occupancy levels and the routing decisions.

In the counterfactual analysis, we use our estimated model to quantify the impact of the capacity expansion on the overall system's performance in terms of its congestion level and off-service likelihood. More importantly, we compare our results to those using alternative models and highlight the importance of using the estimated network-dependent routing policy and calculating the equilibrium system performance. We show that using alternative models can lead to biased management decisions. This stresses the key role of the network effect in the capacity management of inpatient wards and highlights the advantage of our proposed empirical framework.

6.1. Benchmark methods

We consider two benchmark methods to highlight the two main advantages of our approach. The first method takes into account the off-service practice but not the network equilibrium performance. Let $\mathcal{S}$ denote the collection of all 8 specialties. Denote $\mathcal{P}^s_{jt}(\rho)$ as the probability of routing a patient of specialty s, $s \in \mathcal{S}$, to ward j in period t when the network occupancy is ρ, where $\rho = (\rho(1), \ldots, \rho(13))$ is a vector with 13 components, denoting the occupancy of the 13 wards. Let ρ_0 denote the current average occupancy levels of 13 wards, and $N' = (N'_1, \ldots, N'_{13})$ denote the updated capacity levels. We define

$$\rho'_0(j) = \sum_{s \in \mathcal{S}} \frac{\lambda_s \mathcal{P}^s_{jt}(\rho_0)}{N'_j \mu_{sj}},$$

which denotes the average occupancy of ward j immediately after the capacity increase. The updated occupancy will affect the patient routing probabilities, which in turn affect the occupancy levels. The updated occupancy for ward j, $j = 1, \ldots, 13$, takes the form

$$\rho_1(j) := \sum_{s \in \mathcal{S}} \frac{\lambda_s \mathcal{P}^s_{jt}(\rho'_0)}{N'_j \mu_{sj}}.$$

If we keep updating the routing probabilities and occupancy levels iteratively and take the detailed time-varying dynamics into account, we will reach an equilibrium average occupancy. However, for this method, we use the one-step update, ρ_1, as an estimate of the occupancy level after the capacity change. We refer to this method as the one-step method.

In the second benchmark method, we assume the routing probability only depends on the primary ward's occupancy– i.e., the routing policy does not depend on the occupancy of the other connected units in the network. We then use simulation to evaluate the equilibrium system performance. In particular, for specialty $s \in \mathcal{S}$, we set $\tilde{\rho}^s_t(j) = O_{jt}$ if ward j is a primary ward for specialty s, and $\tilde{\rho}^s_t(j) = \rho_0(j)$ if ward j is a non-primary ward. Recall that ρ_0 is the empirical average occupancy of the network. $\tilde{\rho}^s_t(j)$ tracks the real-time occupancy of the primary ward only. We then route patients according to $\mathcal{P}^s_{jt}(\tilde{\rho}^s_t)$. In this case, if the primary ward has high occupancy, regardless of the occupancy levels of the other connected units, patients are more likely to be placed off service. We refer to this method as the semi-static method.

The comparisons of our model with the two benchmark methods demonstrate the two main advantages of our model separately. By comparing the counterfactual results using our integrated model to the first benchmark method, we highlight the importance of taking into account the equilibrium network effect in the complex inpatient ward network: Although both the one-step model and our model allow the patient routing policy to depend on the primary ward's occupancy as well as the occupancy of the linked units in the network, our approach dynamically adjusts the occupancy with the routing policy until the system reaches an equilibrium. By comparing the counterfactual estimation using our estimation strategy to the second benchmark method, we highlight the importance of taking the occupancy of all the connected units into account in patient routing: Although both the second method and our proposed model take into account the equilibrium network effect, the routing policy in the second benchmark method only depends on the occupancy of the primary unit, while the estimated routing policy in our model depends on the occupancy of all linked units in the network.

To better demonstrate the key role of the network effects and the advantages of our method, we next briefly discuss a simple example. Consider two specialties, Cardio and Gastro, and the network depicted in Figure 4. The network connections represented by dotted lines in Figure 4 reflect the following observations in our data. First, the Gastro ward is a commonly used off-service ward for Cardio patients and also takes many off-service patients from other specialties, e.g., GenMed, Neuro, etc. Second, a significant share of Gastro patients is placed off-service (33%) [2].

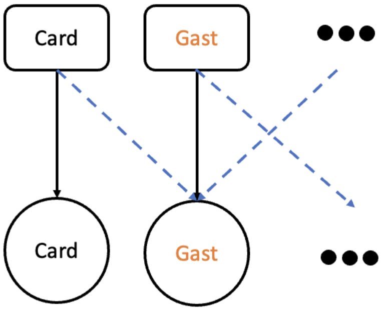


Figure 4 **A partial view of the inpatient ward network**

We consider two capacity expansion scenarios: One is to add more beds to the Cardio ward (Scenario A); the other is to add more beds to the Gastro ward (Scenario B). We first discuss the importance of capturing the network equilibrium behavior. When more beds are added to the Cardio ward, the Cardio ward occupancy will decrease, which increases the number of Cardio patients being routed to the Cardio ward. Therefore, fewer Cardio patients will be placed off service. However, as fewer Cardio patients are placed off-service and the Cardio ward occupancy continues to increase, the occupancy of the off-service ward, i.e., the Gastro ward, may decrease. This will in turn trigger more Cardio patients being placed off-service. The above two mechanisms may interact for a few iterations before the system reaches a new equilibrium. Since the one-step analysis fails

[2] Note that the two observations coexist as the patient arrivals across specialties and occupancy levels across wards fluctuate substantially over time.

to take such interactions into account, it may over-predict the occupancy of the Cardio ward since a lot more Cardio patients are routed to the Cardio wards now.

We next discuss the importance of fully capturing the network connectivity in the routing policy. Under Scenario A, as we place fewer Cardio patients in the Gastro ward, its occupancy decreases. However, the decrease in the Gastro ward's occupancy leads to fewer Gastro patients being placed off-service. Meanwhile, other specialties may also route more patients to the Gastro ward. Thus, not taking the network connectivity into account may over-predict the spill-over effect of adding beds to the Cardio ward on reducing the occupancy of the Gastro ward. For Scenario B, if we increase the capacity of the Gastro ward, it will not only result in fewer Gastro patients being placed off service, but also more Cardio patients being placed off service in the Gastro ward. In this case, not taking the network connectivity into account may lead to an underestimation of the spill-over effect of adding beds to the Gastro ward on reducing the occupancy of the Cardio ward. We next show that the findings of our counterfactual analyses on the full inpatient ward network are indeed consistent with the intuition we provide in this simple example.

6.2. Impact of capacity expansion on unit occupancy

In this section, we use our estimated model to compute the impact of adding a new wing with ten beds on the performance of the system. Our partner hospital was considering four possible ways of allocating the ten beds. Since Cardio, GenMed, and Gastro are the three busiest specialties (with high primary ward occupancy), the hospital was considering allocating all ten beds to one of those three specialties (scenarios I, II, and III) respectively, or allocating the ten beds to the three specialties proportionally to their current capacity: three beds to Cardio, three beds to Gen Med, and four beds to Gastro (scenario IV). We compare the predicted performance change under the four potential scenarios to provide hospital managers with prescriptive policy recommendations. In addition, we compare the results predicted by our model to the two benchmark models.

Table 8 presents the occupancy after the capacity expansion when all ten beds are allocated to the Cardio specialty (scenario I). We present the impact of the capacity expansion on the occupancy for Cardio as well as Gastro, since the Gastro ward is the main off-service ward for Cardio patients. Note that our counterfactual simulation provides occupancy for all units in the inpatient ward network, but we only present the results for the Gastro ward as the impact on other specialties is not significant. We focus on the occupancy at 12am, because it reflects the daily average occupancy well, and the results at the other times of the day follow very similar patterns. The results based on our model show that the occupancy of the Cardio primary ward decreases substantially after the capacity expansion. In addition, the decrease in occupancy also spillovers to the Gastro ward. Compared with our model, using the one-step method can lead to substantial bias in the occupancy predictions. First, the predicted occupancy in the Cardio ward after the capacity expansion increases, not decreases. The reason for this extremely large and counter-intuitive bias is that the one-step method can only capture the load balancing effect at the observed occupancy level right after the capacity expansion and does not take into account the equilibrium effect of the entire network: As the starting occupancy of the Cardio unit decreases after the capacity expansion, the simple model extrapolates linearly the number of patients routed to the Cardio ward using the observed load-balancing behavior in the data, which leads to an unrealistically large increase in the number patients being routed to the Cardio unit and the extremely high occupancy of that unit. In addition, due to the substantial influx of patients directed to the Cardio ward, the predicted occupancy of the Gastro ward is too low. In summary, the one-step method provides counterfactual results which are too unrealistic to be useful. Second, compared with our model, the semi-static method produces similar occupancy results for the Cardio unit but overestimates the spillover effect for the Gastro unit. The main reason for the overly optimistic estimates of the occupancy reduction in the Gastro ward is that the routing policy does not adjust with the occupancy of the Gastro ward. Instead, it only depends on the occupancy of the Cardio unit. In other words, as more patients are routed to the Cardio unit after the capacity expansion, the occupancy of the Gastro unit decreases. As the difference in the occupancy of Cardio and Gastro units narrows, the aggressive admission of patients to the Cardio unit might slow down, and more patients might be placed off service in the Gastro unit, as a result. However, the semi-static model does not take this factor into account, and, therefore, predicts too low an occupancy for the Gastro unit. We find similar results comparing our model and the two alternative models under scenario II (see Table 9), when all ten additional beds are allocated to the GenMed primary unit.

The results under scenario III are presented in the left panel of Table 10. When the additional beds are allocated to the Gastro ward, our results show that the occupancy of the Gastro ward reduces by about eight percentage points. As Gastro is the main off-service unit for Cardio and GenMed patients, we also find a significant spillover effect, of around one percentage point occupancy reduction, in the Cardio and GenMed wards. Compared with our model, the one-step method again produces unrealistic occupancy for the Gastro unit after the capacity expansion. In addition, the benefits to the Cardio and GenMed units are substantially overestimated as a result of overcrowding the Gastro unit. Meanwhile, the semi-static method finds a much larger occupancy reduction in the Gastro ward, but underestimates the spillover effect for the Cardio and GenMed wards, which is opposite in the direction of the bias we observed in scenarios I and II.

Similar to scenario III, under scenario IV, our model predicts that the occupancy levels of the Cardio, GenMed, and Gastro wards go down after the capacity expansion. The magnitude of the reductions is around three to five percentage points. The one-step model produces unrealistic occupancy predictions. Compared with our model, the semi-static method underestimates the benefits of the capacity expansion in reducing the occupancy of the Cardio and GenMed wards and overestimates that of the Gastro ward.

	before	after		
		our model	one-step	semi-static
Card	94.6%	81.3%	132.6%	81.6%
Gastro	92.6%	91.4%	82.7%	90.5%

Table 8 Scenario I: all ten beds to Card, occupancy at 12am

	before	after		
		our model	one-step	semi-static
GenMed	92.6%	83.2%	99.8%	84.0%
Gastro	92.6%	91.4%	87.3%	90.7%

Table 9 Scenario II: all ten beds to GenMed, occupancy at 12am

		Scenario III			Scenario IV		
	before	after			after		
		our model	one-step	semi-static	our model	one-step	semi-static
Card	94.6%	93.5%	84.8%	94.9%	89.6%	117.9%	90.8%
GenMed	92.6%	91.5%	83.0%	92.8%	89.1%	98.8%	90.1%
Gastro	92.6%	84.0%	136.5%	77.8%	88.0%	106.1%	85.2%

Table 10 Scenario III: all ten beds to Gastro, and Scenario IV: equal allocation, occupancy at 12am

6.3. Impact of capacity expansion on system congestion and off-service placement

To provide hospital managers with prescriptive policy recommendations in terms of choosing among the four capacity allocation scenarios, we compute the total impact of the expansion on the entire inpatient ward network. We examine two key performance measures, the number of high occupancy hours aggregated across all units in the system and the number of patients placed off service aggregated across all specialties. Although there are additional measures hospital managers might consider, we choose these two performance measures for the following reasons. First, the number of high occupancy hours provides a precise measure of congestion, which is one of the primary measures hospital managers consider when making capacity expansion decisions. Second, the number of off-service placements reflects an important aspect of the quality of care. In addition, as off-service placement is one of the most common ways to mitigate congestion in hospital inpatient ward networks, together with the number of high congestion hours, they provide hospital managers with a complete view of the system performance. Since the predictions of the one-step method prove to be unrealistic, we focus on the comparison between our model and the semi-static method.

Before the capacity expansion, the inpatient ward network is 95% occupied or above[3] for over 2.6 thousand hours per year, which accounts for 29.9% of the time. Table 11 presents the reduction of high occupancy hours per year for the four scenarios of interest. Based on the prediction of our model, scenarios I, II, III, and IV lead to reductions of 424, 356, 381, and 469 hours per year, respectively. In particular, scenario IV, the relatively equal allocation of the additional capacity to all three specialties, reduces high occupancy hours the most. Using the semi-static model, however, we find that scenarios I and IV lead to the same amount of reduction, 359 hours per year, and are the largest among all four scenarios. In addition, the semi-static model significantly underestimates the amount of reduction in all four scenarios. For scenario IV, compared to our model, the semi-static model underestimates the reduction by 23%.

We present in Table 11 the reduction in the total number of off-service placements per year under the four scenarios. If this is the performance metric the hospital prioritizes, both models predict that scenario II achieves the highest reduction. However, the semi-static model overestimates the reduction by close to 20% for scenario II, and the bias is even higher for scenarios III and IV.

In summary, due to its lack of ability to fully account for the network effects, the semi-static model may lead to biased managerial decisions in terms of both how to allocate the additional capacity and their estimated impact on the system.

		Scenario I	Scenario II	Scenario III	Scenario IV
Reduction of high occupancy hours	our model	424	356	381	469
	semi-static	359	324	179	359
Reduction of off-service placements	our model	223	335	-19	204
	semi-static	256	400	146	296

Table 11 Reduction of high occupancy hours and off-service placements per year

7. Conclusion

We develop a structural estimation approach to evaluate the impact of large capacity changes in a complex hospital inpatient ward network. This approach is underpinned by two main components. First, we estimate the patient routing policy using a choice model, which allows the routing policy to depend on an extensive set of factors, including the occupancy of interconnected wards in the inpatient ward network. Using a control function approach with instrumental variables, we accurately quantify how wards' occupancy affects the routing decision, which creates network

[3] We conduct robustness analysis using other high occupancy benchmarks, and the results are qualitatively similar.

connectivity. Second, we integrate the estimated patient routing policy into a carefully calibrated queueing network model, tailored to simulate the patient flow in the inpatient ward network. This allows us to evaluate the change in capacity on the equilibrium performance of the network. Based on the structural model, we conduct counterfactual analyses evaluating different capacity expansion scenarios of interest. We show that our model outperforms two alternative benchmark methods in producing accurate and reliable performance evaluations associated with the capacity changes. This demonstrates the importance of modeling and estimating the network effects to provide unbiased policy recommendations to hospital managers when making large capacity adjustments.

The setting of our study is inpatient wards in hospitals, but the modeling approach we propose can be applied to other service and manufacturing systems that involve complex networks of resources and various degrees of resource sharing. In particular, identifying the various factors that enter customer or demand routing decisions, quantifying the load-balancing behavior, and developing a tailored network model to evaluate equilibrium system performance under different capacity scenarios are vital to providing unbiased capacity management recommendations.

Our estimation results also have important implications for many empirical studies in healthcare operations management, where occupancy is commonly used as an instrumental variable. Our results suggest that there can be a selection in unobserved patient severity for those admitted patients when occupancy changes. The selection introduces bias in the estimated causal effect of the admission (or waiting) decision on patient outcomes. Thus, caution needs to be taken when using occupancy as an instrument in these settings.

References

Arıkan, Mazhar, Barış Ata, John J Friedewald, Rodney P Parker. 2018. Enhancing kidney supply through geographic sharing in the united states. *Production and Operations Management* **27**(12) 2103–2121.

Batt, Robert J, Christian Terwiesch. 2017. Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science* **63**(11) 3531–3551.

Bertsimas, Dimitris, Jean Pauphilet. 2023. Hospital-wide inpatient flow optimization. *Management Science* .

Best, Thomas J., Burhaneddin Sandıkçı, Donald D. Eisenstein, David O. Meltzer. 2015. Managing hospital inpatient bed capacity through partitioning care into focused wings. *Manufacturing & Service Operations Management* **17**(2) 157–176.

Buell, Ryan W, Michael I Norton. 2011. The labor illusion: How operational transparency increases perceived value. *Management Science* **57**(9) 1564–1579.

Dai, J.G., Pengyi Shi. 2019. Inpatient bed overflow: An approximate dynamic programming approach. *MSOM* Forthcoming.

DeValve, Levi, Yehua Wei, Di Wu, Rong Yuan. 2023. Understanding the value of fulfillment flexibility in an online retailing environment. *Manufacturing & service operations management* **25**(2) 391–408.

Ding, Yichuan, Eric Park, Mahesh Nagarajan, Eric Grafstein. 2019. Patient prioritization in emergency department triage systems: An empirical study of the canadian triage and acuity scale (ctas). *Manufacturing & Service Operations Management* **21**(4) 723–741.

Dong, Jing, Pengyi Shi, Fanyin Zheng, Xin Jin. 2019. Off-service placement in inpatient ward network: Resource pooling versus service slowdown. Working paper.

Freeman, M., N. Savva, S. Scholtes. 2017. Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science* **63**(10) 3147–3167.

Green, L. V. 2002. How many hospital beds? *Inquiry* **39**(4) 400–412.

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

Guajardo, Jose A, Morris A Cohen, Sang-Hyun Kim, Serguei Netessine. 2012. Impact of performance-based contracting on product reliability: An empirical analysis. *Management Science* **58**(5) 961–979.

Gupta, D., S.J. Potthoff. 2016. Matching supply and demand for hospital services. *Foundations and Trends in Technology, Information and Operations Management* **8**(3-4) 131–274.

Helm, Jonathan E., Mark P. Van Oyen. 2014. Design and optimization methods for elective hospital admissions. *Operations Research* **62**(6) 1265–1282.

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

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

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

Kim, Song-Hee, Fanyin Zheng, Joan Brown. 2023. Identifying the bottleneck unit: Impact of congestion spillover in hospital inpatient unit network. *Management Science* .

Mandelbaum, Avishai, Sergey Zeltyn. 2013. Data-stories about (im) patient customers in tele-queues. *Queueing Systems* **75**(2-4) 115–146.

Pallin, Daniel J, Janice A Espinola, Carlos A Camargo Jr. 2014. Us population aging and demand for inpatient services. *Journal of hospital medicine* **9**(3) 193–196.

Petrin, Amil, Kenneth Train. 2010. A control function approach to endogeneity in consumer choice models. *Journal of marketing research* **47**(1) 3–13.

Pinker, E., T. Tezcan. 2016. Determining the optimal configuration of hospital inpatient rooms in the presence of isolation patients. *Operations Research* **61**(6) 1259–1276.

Samiedauluie, S., B. Kucukyazici, V. Verter, D. Zhang. 2017. Managing patient admissions in a neurology ward. *Operations Research* **65**(3) 635–656.

Shi, Cong, Yehua Wei, Yuan Zhong. 2019. Process flexibility for multiperiod production systems. *Operations Research* **67**(5) 1300–1320.

Soltani, Mohamad, Robert Batt, Hessam Bavafa, Brian Patterson. 2019. Does what happens in the ed stay in the ed? the effects of emergency department physician workload on post-ed care use. *The Effects of Emergency Department Physician Workload on Post-ED Care Use (November 28, 2019)* .

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

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

Sun, Zhankun, Nilay Tanık Argon, Serhan Ziya. 2018. Patient triage and prioritization under austere conditions. *Management Science* **64**(10) 4471–4489.

Wallace, Rodney B, Ward Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manufacturing & Service Operations Management* **7**(4) 276–294.

Yu, Qiuping, Gad Allon, Achal Bassamboo. 2017. How do delay announcements shape customer behavior? an empirical study. *Management Science* **63**(1) 1–20.

Appendix A: Additional Descriptive Analysis

We provide in Table 12 the average number of admissions in daytime and nighttime, by admission sources, for the four specialties we focus on. First, we observe that, as shown in Figure 2a, most EL and ICU transfer patients are admitted during the daytime. Second, we find that the proportion of EL and ICU transfer patients varies for the two specialties. About half of the Cardio admissions are from EL or ICU transfers, while the majority of GenMed patients are admitted from ED.

	Cardio	GenMed	Surg	Gastro
	Daytime (7am-9pm)			
Total Adm Count	16.1	11.9	13.7	6.0
ED %	39%	80%	39%	58%
Elec/ICU %	61%	20%	61%	42%
Off-service %	34%	13%	13%	26%
	Nighttime (9pm-7am)			
Total Adm Count	3.8	6.1	3.9	2.4
ED %	84%	96%	88%	92%
Elec/ICU %	16%	4%	12%	8%
Off-service %	53%	45%	23%	45%

Table 12 Daily admission rates and off-service placement proportions by specialties and admission sources during the daytime and nighttime.

Table 12 also shows the proportion of patients placed off service from the four specialties during the daytime and nighttime. Interestingly, even though the primary wards' occupancy levels are lower at 8pm than at 6am, the off-service proportion is larger during the nighttime than during the daytime. This suggests that there might be differences in the bed management team's load-balancing considerations for daytime versus nighttime.

Appendix B: Additional Estimation Results

B.1. Control Variables

Our main treatment variable is the occupancy level of the wards one hour before the focal time interval. We also control for the following covariates in our estimation.

- **Patient-level information:** age, gender, triage score.
- **Unit-level information:** occupancy on the previous day, number of discharges on the current day
- **System-level information:** number of elective admissions and number of ICU transfer patients from relevant specialties (specialties that could potentially route patients to focal ward) on the current day.
- **Day of the week.**

B.2. First-Stage Estimation Results

Tables 13 – 16 summarize the first-stage estimation results at 6am and 12am for different wards, W1 – W13. Ward 11 is excluded since it is not used in our estimation. We only report the statistically significant IVs in the table. The estimation results at the other times, i.e., 11am, 3pm, 8pm, are similar.

	W1	W2	W3	W4	W5	W6
ED Cardio	2.01×10^{-3} *** (0.74×10^{-3})		1.63×10^{-3} * (0.93×10^{-3})	2.17×10^{-3} ** (1.03×10^{-3})		
ED Onco					6.78×10^{-3} * (3.68×10^{-3})	
ED Ortho	1.75×10^{-3} ** (0.81×10^{-3})				7.33×10^{-3} *** (1.24×10^{-3})	5.49×10^{-3} *** (1.30×10^{-3})
ED Surg	3.27×10^{-3} *** (0.77×10^{-3})		6.40×10^{-3} *** (0.97×10^{-3})	4.04×10^{-3} *** (1.06×10^{-3})	2.92×10^{-3} ** (1.18×10^{-3})	2.31×10^{-3} * (1.24×10^{-3})
ED Gastro						
ED GenMed	1.41×10^{-3} ** (0.61×10^{-3})	5.48×10^{-3} *** (0.59×10^{-3})		2.28×10^{-3} *** (0.84×10^{-3})		2.29×10^{-3} ** (0.98×10^{-3})
ED Neuro		2.96×10^{-3} *** 1.02×10^{-3}			5.62×10^{-3} *** (1.62×10^{-3})	4.50×10^{-3} *** (1.71×10^{-3})
ED Renal	1.99×10^{-3} * (1.15×10^{-3})	2.59×10^{-3} ** (1.10×10^{-3})				
ED Resp	3.76×10^{-3} *** (1.43×10^{-3})			11.88×10^{-3} *** (1.99×10^{-3})		
Num of Obs	365	365	365	365	365	365
F-stat	22.29	20.27	49.33	67.67	33.51	37.5

Table 13 Estimated coefficients and standard errors (in parentheses) of the IVs in the first stage for the 6am occupancy of wards W1 - W6. Only statistically significant IVs are reported. $*$,$**$,$***$ indicate statistical significance at 10%,5%,1% level respectively. We also control for other covariates.

Appendix C: A High-fidelity Queueing Network Model

In this section, we introduce the key components of the queueing network model introduced in Section **??**. We also provide some additional calibration results.

Key components. The key components of the queueing network model include a carefully designed network structure, time-varying arrivals and departures, routing policy based on our estimated choice model, and off-service slowdown–i.e., increased LOS due to off-service placement.

	W7	W8	W9	W10	W12	W13
ED Cardio		2.36×10^{-3} *** (0.86×10^{-3})		4.30×10^{-3} *** (0.80×10^{-3})	4.06×10^{-3} *** (0.52×10^{-3})	1.98×10^{-3} *** (0.70×10^{-3})
ED Onco						5.20×10^{-3} ** (2.28×10^{-3})
ED Ortho		5.81×10^{-3} *** (0.95×10^{-3})	1.70×10^{-3} * (0.90×10^{-3})			
ED Surg						
ED Gastro	1.19×10^{-3} ** (0.48×10^{-3})	2.27×10^{-3} ** (1.11×10^{-3})			1.23×10^{-3} * (0.66×10^{-3})	5.53×10^{-3} *** (0.90×10^{-3})
ED GenMed	2.31×10^{-3} *** (0.31×10^{-3})	2.41×10^{-3} *** (0.73×10^{-3})	1.84×10^{-3} *** (0.69×10^{-3})			
ED Neuro	1.80×10^{-3} *** (0.53×10^{-3})	2.78×10^{-3} ** (1.24×10^{-3})				3.27×10^{-3} *** (1.00×10^{-3})
ED Renal			8.82×10^{-3} *** (1.28×10^{-3})			2.83×10^{-3} *** (1.09×10^{-3})
ED Resp				4.19×10^{-3} *** (1.54×10^{-3})		2.92×10^{-3} ** (1.35×10^{-3})
Num of Obs	365	365	365	365	365	365
F-stat	6.54	43.46	74.07	7.68	13.88	18.36

Table 14 Estimated coefficients and standard errors (in parentheses) of the IVs in the first stage for the 6am occupancy of wards W7 - W13. Only statistically significant IVs are reported. $*$,$**$,$***$ indicate statistical significance at 10%,5%,1% level respectively. We also control for other covariates.

	W1	W2	W3	W4	W5	W6
ED Cardio					2.00×10^{-3} * (1.08×10^{-3})	
ED Onco					8.21×10^{-3} ** (3.40×10^{-3})	
ED Ortho					5.88×10^{-3} *** (1.22×10^{-3})	4.43×10^{-3} *** (1.35×10^{-3})
ED Surg	2.88×10^{-3} *** (0.79×10^{-3})		4.54×10^{-3} *** (1.09×10^{-3})	3.18×10^{-3} *** (1.22×10^{-3})	2.97×10^{-3} *** (1.14×10^{-3})	
ED Gastro	1.74×10^{-3} * (0.94×10^{-3})	2.11×10^{-3} * (1.10×10^{-3})				
ED GenMed		3.89×10^{-3} *** (0.71×10^{-3})		1.73×10^{-3} * (0.94×10^{-3})		
ED Neuro		2.67×10^{-3} ** (1.28×10^{-3})				
ED Renal						
ED Resp	4.93×10^{-3} *** (1.41×10^{-3})			5.12×10^{-3} ** (2.26×10^{-3})		
Num of Obs	365	365	365	365	365	365
F-stat	23.41	16.27	40.13	47.69	33.01	34.42

Table 15 Estimated coefficients and standard errors (in parentheses) of the IVs in the first stage for the 12am occupancy of wards W1 - W6. Only statistically significant IVs are reported. $*$,$**$,$***$ indicate statistical significance at 10%,5%,1% level respectively. We also control for other covariates.

- Network structure: There are $J = 13$ inpatient wards (server pools), where the j-th pool has N_j beds (servers). Patients are classified into $I = 8$ medical specialties (classes). Each specialty has three different subclasses representing different admission sources, including the ED, EL, and ICU transfer admissions.
- Time-varying arrivals and discharges: It is clear from Figure 2(a) that the arrivals have a time-varying pattern. We model the arrival process for each subclass as a nonhomogeneous Poisson process with its corresponding periodic arrival-rate function (a period is one day). In addition, the discharges are also highly time-varying, clustered in the afternoon as shown in Figure 2(b). To capture the block discharges,

	W7	W8	W9	W10	W12	W13
ED Cardio		2.90×10^{-3} ***	2.36×10^{-3} **	2.94×10^{-3} ***	4.15×10^{-3} ***	2.56×10^{-3} ***
		(0.94×10^{-3})	(0.92×10^{-3})	(0.88×10^{-3})	(0.62×10^{-3})	(0.79×10^{-4})
ED Onco						
ED Ortho		4.38×10^{-3} ***				
		(1.05×10^{-3})				
ED Surg		1.67×10^{-3} *				
		(0.98×10^{-3})				
ED Gastro				2.11×10^{-3} ***	3.75×10^{-3} ***	
				(0.78×10^{-3})	(1.02×10^{-3})	
ED GenMed	2.31×10^{-3} ***	2.00×10^{-3} ***	1.49×10^{-3} **			
	(0.36×10^{-3})	(0.77×10^{-3})	(0.76×10^{-3})			
ED Neuro	1.87×10^{-3} ***	3.70×10^{-3} ***				
	(0.66×10^{-3})	(1.39×10^{-3})				
ED Renal	1.26×10^{-3} *		5.66×10^{-3} ***			2.20×10^{-3} *
	(0.65×10^{-3})		(1.36×10^{-3})			(1.17×10^{-3})
ED Resp	1.79×10^{-3} **			2.53×10^{-3} **	2.62×10^{-3} *	
	(0.85×10^{-3})			(1.17×10^{-3})	(1.52×10^{-3})	
Num of Obs	365	365	365	365	365	365
F-stat	7.43	30.62	56.47	7.35	13.08	16.13

Table 16 Estimated coefficients and standard errors (in parentheses) of the IVs in the first stage for the 12am occupancy of wards W7 - W13. Only statistically significant IVs are reported. $*$,$**$,$***$ indicate statistical significance at 10%,5%,1% level respectively. We also control for other covariates.

we model a patient's service time as a two-time model: an integer number of days, d_{los}, estimated from the number of days each patient spent in the inpatient wards, using 10am as the start point of a day; and a real number of hours, h_{dis}, corresponding to the *discharge delay* between the morning rounds (10 am) on the day of discharge and the actual departure time of the patient.

- Routing decisions: We take a fully data-driven approach and fit randomized routing policies. More details are provided below.
- Off-service slowdown: Using data from the same partner hospital, Dong et al. (2019) estimate that there is a 17% increase in the LOS for patients who are placed off service. We incorporate this slowdown effect in the simulation model.

Routing policies. A major role of the simulation model is to provide a platform for us to conduct counterfactual analyses of how changes in capacity would change the system's performance. Note that the bed assignment decisions and occupancy levels are highly correlated. For example, higher occupancy leads to fewer bed assignments due to the load-balancing effect, which then reduces the occupancy and, consequently, affects future bed assignments. The network structure further complicates the evaluation of the system performance. This is why a simple back-of-the-envelope calculation is not sufficient to perform the counterfactual study. To calibrate the simulation model such that it has enough fidelity to capture the interplay between bed assignment decisions and occupancy levels, it is important to use a proper routing policy.

For all admissions, we use randomized routing policies, where a class i patient is assigned to a ward j at time t with some probability p_{ijt}. For EL and ICU transfer admissions, we estimate these probabilities, p_{ijt}'s, directly from the empirical routing probabilities, i.e., the empirical proportion of class i patients routed to ward j for the time period that t belongs to. We estimate the empirical proportions for three time periods: morning (7am-noon), afternoon (noon-9pm), and nighttime (9pm-7am) based on the admission patterns for EL and ICU transfer patients shown in Figure 2a.

For the ED admissions, we estimate p_{ijt}'s from the fitted choice models in Section 4. In particular, we randomly choose a ward j to admit the patient according to the probabilities estimated probabilities, which depend on the real-time occupancy of the wards relevant to the patient's specialty. We note that, very occasionally, a patient may be assigned to a full ward; this does not happen often due to the negative coefficient associated with high ward occupancy–i.e., the probabilities of assigning a patient to a full ward are very small. When this does happen, to ensure the simulated ward assignments are consistent with the estimated choice model, we use "surge capacities" in the assigned ward so that the patient will still be admitted–i.e., the capacity imposes only a soft constraint. The practice of using surge capacities, such as putting temporary beds in corridors, is used in our partner hospital.

Figure 3 in the main paper shows that our simulation can calibrate the occupancy levels of different wards at different time points and the off-service proportion for different specialties accurately. We emphasize here that we are able to calibrate not only the overall off-service proportion for each specialty, but also the assignment proportions to each individual ward for each specialty. Table 17 compares the observed and simulated assignment proportions to each ward for Cardio and GenMed during the daytime and nighttime.

	Daytime				Nighttime			
	Cardio		GenMed		Cardio		GenMed	
	observed	simulated	observed	simulated	observed	simulated	observed	simulated
W1	2%	2%	0%	0%	7%	7%	2%	2%
W2	2%	2%	56%	52%	4%	4%	36%	34%
W3	2%	2%	0%	0%	3%	3%	3%	3%
W4	3%	3%	3%	2%	5%	5%	6%	5%
W5	4%	4%	2%	1%	5%	4%	3%	3%
W6	1%	1%	1%	1%	4%	4%	3%	3%
W7	1%	1%	27%	33%	2%	2%	19%	21%
W8	5%	4%	2%	2%	9%	10%	7%	6%
W9	5%	5%	2%	3%	5%	5%	8%	9%
W10	13%	8%	0%	0%	11%	13%	1%	0%
W11	0%	0%	0%	0%	0%	0%	1%	1%
W12	49%	58%	0%	0%	36%	31%	1%	1%
W13	14%	10%	6%	4%	9%	11%	11%	12%

Table 17 Observed and simulated assignment proportions to each ward: Cardiology and General Medicine.