Off-service Placement in Inpatient Ward Network: Resource Pooling versus Service Slowdown

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

Inpatient ward beds play a central role in hospital operations. To better facilitate coordination of care, the beds are usually grouped into different specialized units, with each unit designated to serve patients in certain primary specialties. However, inpatient wards are often associated with high level of bed utilization and large variability in demand. When waiting time is excessively long before a bed in the primary ward becomes available, the patient may be assigned to a bed in a non-primary ward. This is referred to as off-service placement. In this paper, we take a data-driven approach to study off-service placement by taking into account three key aspects of the problem: the network structure of the wards; the complex bed assignment decisions; and the causal effect of off-service placement on patient outcome. Our analysis quantifies the trade-off between off-service placement and admission delay, and provides prescriptive solutions to improve system performance.

Key words: Inpatient Bed Management, Off-service Placement, Network Effect, Patient Outcome, Empirical Methods, Stochastic Models

1. Introduction

Inpatient ward beds are one of the most important resources in a hospital. The management of these beds has a direct impact on the majority of patients in the hospital. Moreover, it also affects the patients in the connected units, such as the Emergency Department (ED), the Intensive Care Unit (ICU), and the Operating Room (OR). The inpatient ward beds are typically grouped into specialized units, with each ward unit dedicated to a specific care type (specialty). This focused care model allows the hospital to better coordinate the care team, which consists of the specialized physicians, the nurses, and the technicians. It also allows the nurses to standardize and improve the process within the ward unit and better manage the specialized equipment and procedures (Best et al. 2015). As a result, the focused care model enables the hospital to provide better-quality of care to patients. However, there are also drawbacks associated with this focused care model.

Since inpatient ward units often experience high utilization of bed capacity and high variability in demand, when a designated ward unit is overloaded, patients may experience extensive delays in the admission process. To reduce this *admission delay*, a common strategy adopted by many hospitals, both in the U.S. and in other countries, is to place the patients in a non-designated ward unit, which is referred to as *off-service placement*. In other words, hospitals may often need to choose between excessive admission delay and off-service placement. As a result, it is necessary for hospital managers to understand the impact of both the admission delay and the off-service placement on patient- and system-level performance metrics and, more importantly, the tradeoffs they face when choosing between them.

While the impact of admission delay on an individual patient's medical outcome, as well as on system-level performance, has been well studied in the literature (Allon et al. 2013, Carr et al. 2010, Chan et al. 2016, Hoot and Aronsky 2008, Singer et al. 2011), we understand far less about the effect of off-service placement. Stylianou et al. (2017) and Song et al. (2018) show that off-service placement can lead to longer length-of-stay (LOS) for those patients who are placed in off-service units. We refer to this effect as the off-service slowdown in this paper. However, to the best of our knowledge, the literature has not studied how the off-service slowdown propagates through the complex inpatient ward network and impacts the overall *system* performance. On the one hand, off-service placement creates more resource pooling in the inpatient ward network, which may improve system performance by reducing admission delay. On the other hand, the longer LOS of the off-service patients may generate a greater workload for the system and block future admissions, triggering longer admission delay and even more off-service placements – or a 'snowball effect.' Thus, it is important to understand how the off-service slowdown affects admission delay in the inpatient ward system. In particular, it is important to analyze how the longer LOS propagates through the complex inpatient ward network and to quantify the overall delay in the system in the presence of off-service placement. This is the first goal of our paper.

Our second goal is to quantify the tradeoff between admission delay and off-service placement faced by hospital managers more generally. The perspective we take is similar to that of the "efficiency frontier" analysis, in which we trace out the full set of admission delay and off-service placement proportion combinations possible for given inpatient ward configurations. The resulting curve can provide hospital managers with a detailed quantification when deciding between the desired level of admission delay and off-service placement. More importantly, it provides a natural way to evaluate the effectiveness of off-service placement as an important control to reduce admission delay in inpatient ward units.

To achieve our two goals, we combine econometric tools with stochastic modeling and conduct a fully data-driven analysis using the detailed patient flow data from a large public hospital

in Singapore. Specifically, we first estimate the bed assignment policy using a multinomial logit model. Next, using an instrumental variable approach, we estimate the causal effect of off-service placement on the LOS of those misplaced patients. Then, we use these estimates to calibrate a high-fidelity stochastic model that captures the complex inpatient ward network structure and patient flow dynamics. Finally, we use the model to compute system performance measures and perform counterfactual analyses to answer our research questions.

We next summarize our main findings and highlight the advantage of our methodology.

Off-service placement and admission delay tradeoff. Our tradeoff curves illustrate that off-service placement is, in general, an effective way to reduce admission delay; that is, there is a negative relationship between the overall admission delay and the off-service placement proportion. However, the marginal reduction in admission delay diminishes quickly as the off-service placement proportion increases. In other words, the effectiveness of off-service placement as a control to manage delay varies substantially. As a result, it is important for hospital managers to have information about where the current operation lies on the tradeoff curve and the set of choices they face about trading off admission delay with off-service placement.

Capacity reallocation and network effect. To provide more general insights into the off-service and admission delay tradeoff, we deviate from the current operations of our partner hospital and analyze the tradeoff curves under more balanced bed allocation scenarios. First, we find that more balanced bed allocation can substantially improve the efficiency frontier. Second, we show that the diminishing marginal return to off-service placement also applies to the scenarios with balanced allocation. Third, by comparing the tradeoff curves under different capacity reallocation strategies, we find that the structure of the patient flow network plays a key role in determining the off-service and admission delay tradeoff. Allocating capacity to well-connected specialty wards (wards that are attractive for off-service patients from other specialties) improves the tradeoff curve substantially more than those wards that are less connected in the network.

Off-service slowdown. Finally, we analyze the impact of off-service slowdown on the tradeoff curve. We find that under the current load and patient composition of our partner hospital, the magnitude of the off-service slowdown factor does not have a big impact on the tradeoff curve. However, if the share of patients potentially subject to the off-service slowdown is larger or the system load is higher, the off-service slowdown can have a significant impact on the shape of the tradeoff curve. In those cases, when the off-service proportion is relatively high, further increasing the off-service proportion can lead to longer admission delays. In other words, the off-service slowdown effect cancels out the benefit of resource pooling created by off-service placement in the network. In addition, the off-service slowdown can have different impacts on different specialty

wards, depending on the ward network structure. In particular, well-connected wards suffer more from the higher workload generated by off-service slowdown.

Methodology. The stochastic network model we build incorporates several key features of inpatient flow dynamics. In particular, to model how patients are assigned to different wards in the network, we take a fully data-driven approach, as opposed to the stylized routing policies in the multi-class queueing literature. Using the data, we fit a discrete choice model to understand the importance of various determinants of patient routing policies in practice. We also demonstrate that even a highly sophisticated yet stylized index-based policy, constructed according to the insights from our choice model, is inadequate to capture the real system dynamics. More importantly, a comparison between the tradeoff curves constructed using our fitted bed assignment policy and the index-based policy shows that using the index-based policy can lead to substantial bias in the shape of the tradeoff curve and, therefore, biased managerial decisions. This highlights the importance of our method of combining econometric tools with stochastic models to provide accurate evaluations of system performance. We believe that our framework can be applied in other hospital management settings in which stylized policies may not be able to capture the complex nature of the managers' decision-making process.

The rest of the paper is organized as follows. We conclude this section with a brief review of the literature. In Section 2, we provide an overview of our dataset and the operation of our partner hospital. In Section 3, we introduce the stochastic network model, providing details about the key features of the model and how to calibrate the model. In Sections 4 and 5, we address two main estimation challenges: the routing policy; and the causal effects of admission delay and off-service placement on patients' outcome. In Section 6, we construct the tradeoff curve based on the stochastic network model. We also study how different factors, such as the ward network structure and the off-service slowdown, impact the tradeoff curves. We provide prescriptive policy recommendations and insights on inpatient flow management using the tradeoff curves. Section 7 concludes the paper.

1.1. Related Literature

Our work is related to several strands of the literature. First, our paper is closely related to the literature on hospital capacity management. Green (2002) is among the first to study capacity management in hospitals with the focused care model. She points out that different medical specialties have different service-level requirements and that hospital managers should carefully quantify the capacity needs for each specialty. Subsequent research also studies capacity planning and ward design for hospitals (Gupta and Potthoff 2016, Pinker and Tezcan 2016). Best et al. (2015) and Kuntz et al. (2019) analyze the design of medical wards and compare pooled versus separate ward

designs. Our work contributes to this line of research by using a data-driven approach to study the capacity management across wards and specialties in a hospital. In particular, we take into account three important features not explicitly captured in the literature: the inpatient ward network structure; the complex routing decisions; and the slowdown effect due to off-service placement.

Second, this paper also contributes to the literature on the common practice of off-service placement in hospitals. It has been acknowledged in the medical literature that there are potentially negative consequences associated with off-service placement (Goulding et al. 2012, Stylianou et al. 2017). In the operations management literature, a concurrent paper by Song et al. (2018) takes an empirical approach to rigorously quantify the magnitude of off-service placement's effect on various patient outcomes, including the LOS, readmission rates, and mortality risks for patients placed off-service. To correct for endogeneity in unobserved patient severity, they apply an instrumental variable (IV) approach using the occupancy in the primary wards and an indicator of hospital business as the instruments. We adopt a similar IV strategy and identify a similar magnitude of the increase in LOS among off-service patients, using data from a different hospital. Our outcome analysis in Section 5 confirms the substantial negative impact of off-service placement on patient outcome, suggesting that the effect is pervasive. However, quantifying this impact is just an important intermediate step in our analysis. The focus of our study is to understand how the longer LOS of the off-service-placed patients is propagated through the complex inpatient ward network and affects overall system performance. Moreover, we quantify the tradeoff between off-service placement and admission delay, and, thus, provide insights into and guidelines for inpatient-flow management.

Third, a growing body of literature studies admission control and scheduling policies in hospital settings (Freeman et al. 2017, Jacobson et al. 2012, Kim et al. 2015), and, more broadly, routing and scheduling policy in service operations (Ata and Van Mieghem 2009, Gurvich and Whitt 2009). The studies most relevant to our paper are Helm and Oyen (2014), Samiedauluie et al. (2017) and Dai and Shi (2019). In this line of work, the authors typically formulate the optimal routing policy as a solution to a stochastic optimization problem by imposing a specific objective function for the decision maker. For instance, the decision maker minimizes the sum of the holding cost and the off-service placement cost. In addition, this line of work usually makes stylized assumptions about the set of policies from which the decision maker is choosing, including imposing the first-come-first-served discipline or strict priority rules. By contrast, in this paper, we acknowledge that the objective of the decision maker is likely to be very complicated in practice. For example, there can be certain upper bounds on the amount of time that patients can stay in the ED, or different preferences for the proportion of off-service placement in certain wards. Thus, we take a fully

data-driven approach and estimate the decision rule. This allows us to provide a more reliable and realistic estimate of the impact of different capacity improvement strategies.

In terms of the methodologies used in the analysis, our paper adopts methods from two streams of the literature. First, to develop the stochastic model describing the patient-flow dynamics, we borrow insights from recent developments in patient-flow modeling (Armony et al. 2015, Dong and Perry 2018, Shi et al. 2016). In particular, our model incorporates features such as the time-varying patient arrival rate, physician rounding, and discharge delays. In addition, our model also captures the network structure. The sophistication of the model renders the exact analysis impossible. We rely, therefore, on extensive simulation experiments to conduct performance analysis. (Shi et al. 2016) and (Han et al. 2016) use similar strategies to evaluate patient discharge and overflow policies, while (Kim et al. 2015) also use such strategies to evaluate ICU admission policies.

Second, in model calibration, we use an IV approach to estimate the causal effect of off-service placement on patient outcomes. Similar empirical strategies have been adopted in many other empirical health care studies (Chan et al. 2016, KC and Terwiesch 2012, Kim et al. 2015). We also use a discrete choice model to estimate the bed assignment policy. This method has been used in estimating customer choice in service operations (Guajardo et al. 2015, Phillips et al. 2015). The advantage of the method is that it allows the decision maker to have a flexible and complex objective function. In other words, we use data to estimate how much weight the decision maker assigns to different factors, such as the system load, the waiting time, etc., in their decision making.

2. Overview of hospital operations

Our study is based on a collaboration with a large teaching hospital in Singapore. The data contain all patient admissions in 2010. To study the inpatient flow, we use a subset of 34,030 admissions out of the 92,081 total admissions. In particular, we exclude admissions to non-inpatient wards and those for certain (highly specialized) specialties that have little interaction with others. The selected subset contains patients admitted to eight specialties: Cardiology (Card), Surgery (Surg), Orthopedic (Ortho), Respiratory (Resp), Gastroenterology (Gastro Endo), General Medicine (Gen Med), Neurology (Neuro), and Renal. Each specialty admits patients from five different sources: emergency department admissions (ED), elective admissions (Elec), intensive care unit (ICU), transfer patients (Trans), and others. For all eight specialties, we also exclude patients admitted to the private wards which require private insurance and consists of a very small share of the population. We provide more details about the dataset in the online supplement; see Figure 1 for a brief summary.

We next provide an overview of the network of inpatient wards we are modeling. The network has 13 inpatient wards serving patients from eight specialties. Each ward contains 20 to 50 inpatient

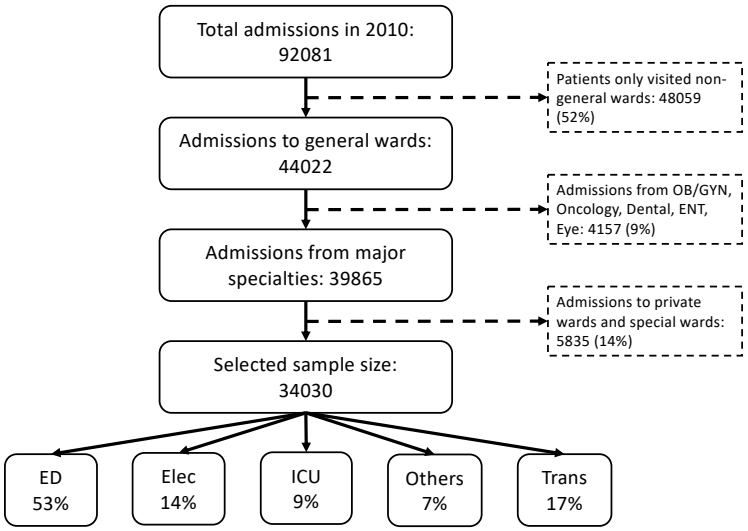


Figure 1 Selection of the patient sample.

ward beds. The hospital uses a focused care model, with each ward designated to serve patients from one specialty (referred to as dedicated wards) or two specialties (referred to as shared wards). For the shared wards, there is a nominal allocation of the beds between the two specialties. Analysis of the bed occupancy data suggests that bed assignments follow the nominal allocations. Figure 2 shows the specialty-ward mapping. The wards are listed in circles, and the specialties are listed in rectangles.

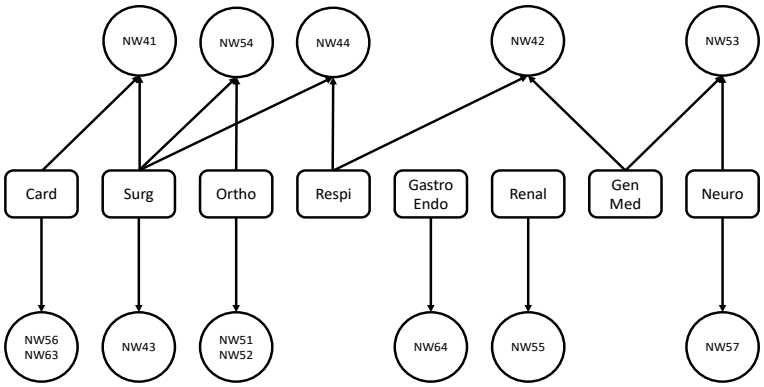


Figure 2 Specialty-ward assignment.

Due to the high operating costs, healthcare systems often operate under a very high load. To understand the workload in our partner hospital, we first calculate some basic statistics describing the size of the practice and the capacity utilization for each specialty. The (nominal) utilization for specialty i is defined as

$$\rho_i := \Lambda_i \mathbb{E}[LOS_i]/C_i,$$

where Λ_i is the daily arrival rate (average number of arrivals per day); $\mathbb{E}[LOS_i]$ is the average length of stay; and C_i is the nominal capacity allocated to that specialty. Specifically, C_i counts the number of beds in the dedicated ward and the nominally assigned number of beds in shared ward for specialty i. $\Lambda_i \mathbb{E}[LOS_i]$ is the average demand per day for specialty i. Thus, ρ_i measures the nominal occupancy rate (capacity utilization rate). As the hospital allows off-service placement, ρ_i can be larger than 100%, suggesting that the capacity allocated to specialty i is not enough to meet the corresponding demand. Table 1 summarizes these statistics. We observe that there is a mismatch between capacity and demand across the eight specialties. For example, Card, Gen Med, and Neuro have insufficient capacity allocated ($\rho_i > 100\%$), while Ortho has a very low occupancy rate (59%). This mismatch may be due to several reasons. For example, the demand for Card and Gen Med grows fast due to the aging population, and the capacity change can be very slow. Ortho is usually very profitable, so the hospital tends to allocate more capacity to it to ensure a good quality of service.

Table 1 also lists the off-service placement proportion for each specialty. We observe from the table that specialties that are overloaded ($\rho_i > 100\%$) have very high off-service percentages. This is because these specialties simply do not have enough capacity to handle their demand. However, even for specialties that are underloaded, the off-service percentage can still be significant (e.g., Gastro, Resp). The possible reason is twofold: (i) The demands are stochastic (randomness in the number of arrivals and the length of stay) and non-stationary. Off-service placement is employed to cope with these stochastic fluctuations. (ii) The primary wards for these specialties also take in a significant number of off-service patients. These off-service patients increase the occupancy of the ward and may, from time to time, prevent the ward from admitting its primary patients – who, in turn, need to be placed off-service in other wards. This suggests that to fully understand the tradeoff between off-service placement and admission delay, we need to take the complicated network structure and the interaction between different specialties into account. Indeed, we will introduce a high-fidelity stochastic network model in Section 3 to capture the underlying physics of the patient flow dynamics.

	Card	Ortho	Surg	Gastro	GenMed	Neuro	Renal	Resp	Overall
Λ_i	19.8	13.2	17.5	8.2	17.8	5.9	6.2	4.6	93.2
$\mathbb{E}[LOS_i]$	3.7	4.4	3.6	3.6	4.5	3.7	4.5	4.0	4.0
C_i	61	98	87	39	67	16	32	25	**425**
ρ_i	119%	59%	73%	76%	119%	134%	87%	74%	**87%**
Off-service %	26%	6%	12%	33%	27%	54%	26%	14%	**22%**

Table 1 Summary of workload related statistics.

3. A high-fidelity stochastic network model

To capture the inpatient flow dynamics, we build a special multi-class, multi-pool queue. In our setting, customers correspond to patients in need of inpatient care, and servers correspond to inpatient ward beds. This stochastic model provides the basis to quantify the tradeoff between off-service placement and admission delay. In what follows, we first introduce the key components of the model. These components are important for our application context, and differentiate our model from the stylized models used in the literature. We then discuss the calibration of the model. We highlight two major estimation challenges: estimating the routing policies and the causal effect of admission delay and off-service placement on patient LOS. Lastly, we compare our estimated routing policy with other highly sophisticated, yet stylized, index-based routing policies in the queueing literature. Our comparison indicates that the stylized routing policies are insufficient to capture the time-dependent dynamics of the system and do not match the empirical performances well.

3.1. Key modeling components

Our stochastic model incorporates unique characteristics associated with inpatient-ward operations. These are critical to take into account when studying inpatient flow, and are not particular to our partner hospital. Some features, such as time-varying arrival and block discharges, are studied in recent works on inpatient flow modeling (Armony et al. 2015, Dong and Perry 2018, Shi et al. 2016). Important new features studied in this paper include a detailed network structure, routing policies based on choice models that consider several key factors, and off-service slowdown–i.e., increased LOS due to off-service placement. We summarize the five key components of our model next; more details about the model can be found in the online supplement.

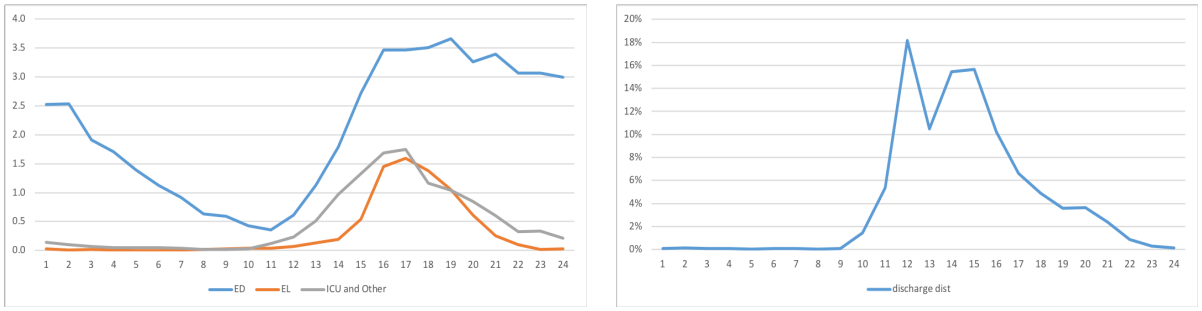
1. Network structure. There are $J = 13$ inpatient wards (server pools), where the j-th pool has N_j beds (servers). Patients are classified into $I = 8$ medical specialties (classes). Each specialty has seven different subclasses representing different admission sources, as shown in Figure 1. For the ED admissions, for example, we further divide them into three subclasses: i) short-stay observational patients, who stay for zero or one day; ii) focus-group patients, who stay for two to seven days (inclusive); and iii) long-stay patients, who stay for longer than seven days. We define the number of days as the number of midnights a patient spends in the hospital, following the literature. For the other four admission sources, we use one subclass to represent each. We model these subclasses separately because their arrival patterns and LOS distributions vary greatly, and the bed management team has different considerations when making routing (bed assignment) decisions for each of them. Figure 1 shows the proportion of patients from each admission source. For ED admissions: the focused group constitute 63% of ED admissions; short-stay observational patients constitute 26%; and long-stay patients constitute 14%. These correspond to 33%, 11%, and 5.8% of all inpatient admissions, respectively.

2. Nonstationarity. Like most service systems in practice, the arrival rates of patients are time-varying. Figure 3a plots the hourly admission rates for patients from three different admission sources in our partner hospital. We observe that the admission rate varies significantly for different hours of the day. Patients from different admission sources also have very different arrival-rate functions. Therefore, we model the arrival process for each subclass as a nonhomogeneous Poisson processe with its corresponding periodic arrival-rate function (the period is equal to one day).

3. Block discharge and LOS. Most discharge decisions are made during the morning rounds, which take place once a day at around 10:00 am. There are further delays between when the discharge is approved and the actual departure time of the patient (when the bed is released). These delays are due to reasons such as paperwork, need for transportation arrangements, coaching by professionals, etc. Figure 3b plots the hourly discharge rate of our partner hospital. We observe that there is almost no departure before 10 am, and the majority of patients are discharged between 11am and 8pm, with noon to 4pm being the peak discharge period. Thus, 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 the *discharge delay* between the morning rounds (10 am) on the day of discharge and the actual departure time of the patient.

4. Routing decision. Bed assignment decisions are complicated by many competing factors. For example, in addition to balancing the load among different wards, one must also take future bed availability into account. If a patient must be placed off-service, there may be different preferences over different non-primary wards. When analyzing the bed assignment decisions in our partner hospital, none of the stylized routing policies is able to capture this level of complexity. Even if we incorporate all the relevant factors, it is hard to decide how much weight to put on each of them. Therefore, we take a fully data-driven approach and fit randomized routing policies from the data. We are especially interested in the routing policies for ED admissions: due to the randomness in their arrival times and the negative consequences associated with admission delay, decisions involving the tradeoff between admission delay and off-service placement must be made in real time. For this class of patients, we fit a detailed discrete choice model, which incorporates key factors in bed-assignment decisions. These factors include preferences for different wards, wards' occupancy, admission delay, and future bed availability. Based on the fitted model, we use a randomized bed-assignment rule (routing policy).

5. Off-service slowdown. The literature suggests that both admission delay and off-service placement can lead to worse patient outcomes. From the operational perspective, we are especially interested in their effects on patients' medical length of stay, d_{los}, as this directly affects the workload of the system. Thus, it is important to account for this when analyzing system performance.



(a) Hourly admission rate (b) Discharge time distribution

Figure 3 Hourly rates of admissions and discharges by patient sources. The numbers on the x-axis denote the hourly interval, e.g., 2 denotes the interval of 1-2am.

Using data from our partner hospital, we estimate the causal effect of admission delay and off-service placement on the patient LOS. We find a longer LOS for patients who are placed off-service. Note that this estimated slowdown is only the immediate effect of off-service placement on the off-service-placed patients. We still need our stochastic model to quantify how this immediate effect on LOS is propagated through the complex patient flow network and affects overall system performance.

3.2. Model calibration

To populate the stochastic model, we need to estimate components (2)-(5) mentioned above for each class (subclass) of patients. That is, the arrival rate functions; the two-time scale service time distributions: d_{los} and h_{dis}; the routing policies; and the causal effects of admission delay and off-service placement on patient LOS. While the arrival rate and the service time distributions can be estimated from the data rather straightforwardly, estimating the routing policies and the causal effects are highly non-trivial.

Estimating routing decisions. The routing decisions for patients from different admission sources are very different in general. Hospitals have more control over the arrival time of elective patients and transfer patients, as their admissions are planned ahead of time. Thus, these patients usually arrive during the afternoon hours when most beds are becoming available due to batch discharge. As a result, we see a small proportion of patients from these admission sources been placed off-service. In contrast, hospitals have less control over ED admissions. In addition, when ED patients are delayed for admission, they occupy valuable resources in the ED. Therefore, we estimate a more detailed model for the routing decisions of ED patients than for the patients from other admission sources.

Specifically, for each sub-class within ED admissions, we estimate choice models that explicitly take the admission delay and the preferences over different wards into account. The fitted model provides insights into how much weight the bed management team places on different factors when

making routing decisions and serves as the basis for the randomized routing policy; we provide more details on the estimation for the focus-group patients in Section 4. It is important to note that we do not impose causal interpretations on these estimated weights–i.e., we do not intend to explain the routing decisions or impose a particular objective function for the decisions. Our goal is to provide a simple descriptive decision rule which we can then vary to generate different levels of off-service placement, and, after that, evaluate the tradeoff between admission delay and off-service placement.

For patients from other admission sources (other than ED), we also use randomized routing policies. Since these patients are irrelevant to the off-service placement and admission delay tradeoff, we estimate the routing probabilities using the sample proportion of admissions to each ward. We emphasize that even though, in the tradeoff analysis, we keep the routing probabilities of these patients fixed, it is important to model them because they affect the ward's occupancy level.

Outcome estimation. To quantify the effect of off-service placement on system performance, it is important to estimate the causal effects of admission delay and off-service placement on the LOS for ED admissions. Recall that we have three subclasses of patients within the ED admissions. We perform the outcome analysis only for the focus-group patients for the following reasons. Because the medical conditions of observational patients are relatively mild, off-service placement is less likely to affect their LOS. Indeed, most of them undergo simple evidence-based protocols. Long-stay patients account for less than 6% of overall admissions, and a relatively small proportion of them are placed off-service. Moreover, their longer LOS is likely to be caused by rare medical conditions or non-medical reasons, e.g., some discharges are delayed due to social reasons (Lim et al. 2006).

The challenge for the outcome estimation is the omitted variable bias. In particular, there are unobserved patient characteristics, such as the severity of the patient's medical condition, that are correlated with both the routing decision and the outcome. We tackle this challenge using an instrumental variable approach, for which we provide more details in Section 5. We emphasize that it is important to estimate the causal effect of off-service placement on those off-service placed patients' LOS without bias. This is because that, to construct the tradeoff curve, one needs to vary the level of off-service placement and evaluate its impact on the system performance of the inpatient ward network. In other words, it is key to estimate the impact of off-service placement on the LOS of the off-sevice placed patients as a first step, in order to analyze the overall tradeoff between off-service placement and admission delay of the inpatient ward network.

Calibration results. The output from the calibrated stochastic model matches the empirical performances remarkably well. Figure 4 plots the bed occupancy (utilization) for each ward and the average occupancy across all wards from simulating the model and from data. The bed occupancy is calculated by the daily average number of patients occupying a bed (all types of patients) divided

by the number of beds for each ward. Figure 5a compares the time-dependent average admission delay for the focus-group patients by their bed-request hours. Figure 5b compares the off-service placement proportion in each specialty, as well as the average across all specialties for the focus-group patients. For the simulation output in Figure 5a, an additional delay of one hour is added to our simulation results to capture the extra amount of delay patients experience after a bed is allocated. This "extra delay" is caused by delays in preparing beds and transporting patients from the ED to the inpatient wards and is often referred to as post-allocation delay (Shi et al. 2016).

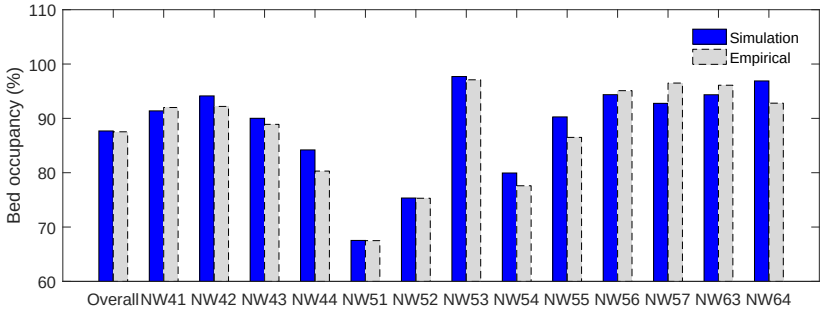
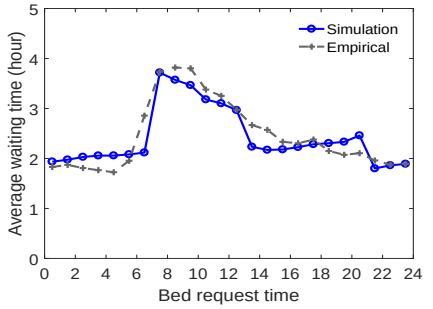
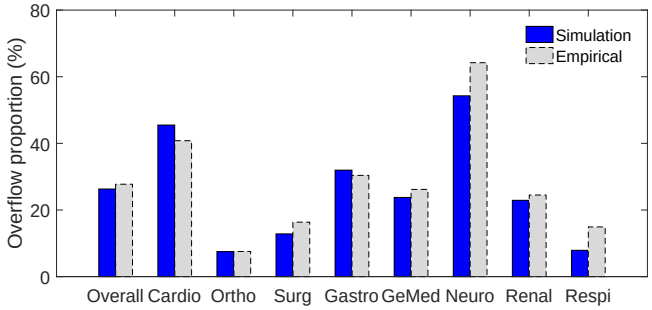


Figure 4 **Comparing bed occupancy rate from simulation output and empirical data.**



(a) Average admission delay by bed-request hour



(b) Average and specialty-level off-service placement proportion

Figure 5 **Comparing average admission delay and off-service placement proportion from simulation output and empirical data. For plot (a), one-hour of extra "administrative" delay is included.**

3.3. Comparison with stylized routing policies

In this section, we compare our estimated policy to other stylized routing policies. To make the stylized policies competitive, we incorporate all the key factors identified from our choice model estimation, such as load balancing, different preferences for different wards, and anticipation for future bed availability, into an index-based routing policy. Our analysis shows that even with all these complications, the stylized routing policies are still insufficient to capture the real system

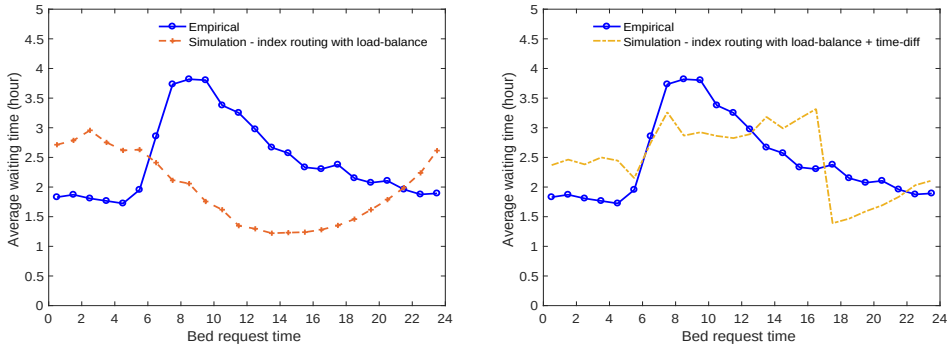
dynamics and match the empirical performances. This justifies the necessity of using a sophisticated choice-model based routing policy, one of the main contributions of this paper.

More specifically, the index-based policies we considered are commonly used in the queueing literature. Each medical specialty is assigned a preference list for different wards. In the basic version, when a new patient arrives, we start from the most preferred ward. If there is an available bed in that ward, we assign the patient to that bed; otherwise, we go sequentially down the preference list until we find an available bed. If no bed is available when we reach the end of the list, we keep the patient waiting. A potential problem with this basic index policy is that we tend to place too many patients off-service. Under the occupancy level of our partner hospital (88%), this policy would allow most patients to get a bed immediately upon arrival, which is not consistent with the empirical average admission delays. To give enough benefit of the doubt to the stylized routing policies, we consider two modifications to the basic index policy that incorporate two important insights from our choice model estimation in Section 4.

- Load balancing: we impose a threshold $U < 100\%$ on the occupancy of off-service wards to strengthen the "undesirability" of off-service placement. When searching through the preference list, if the occupancy of a ward exceeds U, we will skip this ward and move on to the next ward on the list.
- Time differentiation: we further impose two different thresholds, U^m and U^e, for the morning period (7am-7pm) and the evening period (7pm-7am next day), respectively. Our choice model estimation suggests that the off-service placement is used less in the morning due to the anticipation of more beds becoming available soon (in early afternoon). Thus, we set $U^m < U^e$.

To calibrate the index-based policies, we use the baseline utility estimated from our choice model to rank the wards for the preference list. This is also in accordance with the bed allocation guidelines in our partner hospital. For the thresholds on occupancy, we fine-tune their values such that the simulated performance metrics, such as admission delay, are close to the empirical values.

Figure 6 shows the average admission delay for the focus-group patients by their bed-request hours, using the modified index-based policies with load balancing (a), and both load balancing and time differentiation (b). We observe that without time differentiation, (a), the performance curve constructed using the index-based policy has a completely different shape than the empirical performance curve. In (b), after adding the time differentiation, the performance curve gets closer to the empirical one, but it is still not as good as the curve constructed using our estimated policy in Figure 5a. Moreover, the proportion of off-service placement in each ward using the index-based policies also substantially deviates from the empirical one. That is because the occupancy thresholds, even with time-differentiation, are not as dynamically adapted to the overall system load as the choice model.



(a) Index-based rule with load-balancing

(b) Index-based rule with load-balancing and time differentiation

Figure 6 Average admission delay by bed-request hour.

We conclude this section with two remarks. First, in addition to the insufficiency we demonstrated in Figure 6, we will later show, in Section 6.1.1, that the stylized routing policies also generate different and potentially misleading tradeoff curves between admission delay and off-service placement. Second, the improvement in calibration from the basic index policy in Figure 6a to Figure 6b suggests that the insights from our choice model analysis can help derive better stylized policies.

4. Ward assignment decision

In this section, we investigate the determinants of the routing decisions using data for the focus-group patients. We model the ward assignment decisions using a discrete choice model, and estimate the importance of each determinant from the data. There are two objectives. One is to understand, empirically, how important each potential determinant is, from the point of view of the decision maker. The other is to estimate the routing policy for our stochastic model.

We first provide a brief description of the bed assignment process in our partner hospital. All bed assignments in this hospital (including bed-requests from both ED patients and non-ED patients) are managed by the central bed management unit. The unit has two shifts: a regular shift from 7am to 7pm, and a night shift from 7pm to 7am the next day. When the unit's team receives a bed request from ED, they will start searching for an appropriate bed (could be either primary or non-primary) and make a tentative bed allocation. After the bed allocation is confirmed, the team communicates with the ED and the transport team to physically move the patient to the allocated bed. There are internal guidelines for the team to make bed assignments. We highlight two points from the guidelines that motivate the setup of our choice model. First, the team members are required to try their best to find a primary bed that matches the patient's medical specialty; if no primary bed is available, they may start searching for a non-primary bed. Second, the hospital has an internal goal (also required by law) of not keeping patients waiting for more than six hours in the ED. Thus, patients who have waited for a longer time are more likely to be placed off-service.

4.1. Choice model

As described above, one objective of this study is to provide a model that captures the main determinants of the ward assignment decision using data. In particular, we estimate a decision rule that summarizes the behavior of the decision maker, which can then be applied to analyze the tradeoff between off-service placement and admission delay in Section 6.

We emphasize that we do not intend to recover causal parameters in this analysis. The reasons are as follows. First, the decision maker's objective is likely to be complex. Because we do not directly observe that objective, we do not impose a precise objective function or cost function and assume that the bed management team is making the optimal decisions given the objective function. Second, conversations with the bed management team and the hospital management team indicate that the decision makers are following simple ad-hoc rules in practice. Therefore, we choose to use a simple multinomial logit model to capture these rules instead of estimating causal parameters in a full structural model. Finally, this choice is consistent with the overall goal of the paper. Our goal is to evaluate the impact of different off-service levels on the system level admission delay, instead of providing an explanation to the observed off-service level in the data. As a result, we only need a decision rule that captures the main determinants of the ward assignment decision, which we can then apply in Section 6 to generate different levels of off-service placement.

We use the following discrete choice model to study the determinants of ward assignment decisions. For patient i in specialty l, the "utility" (or incurred cost) u_{ijlt} for admitting her into ward j in period t is

$$u_{ijlt} = \alpha_{jl} + X'_{it}\delta_{jl} + Z'_{jt}\gamma_l + W'_t\eta_{jl} + \varepsilon_{ijlt}, \tag{1}$$

where we include waiting as one of the options j; X_{it} is a vector of patient characteristics that could change over time; Z_{jt} is a vector of ward characteristics that are also allowed to change over time; η_{jl} is a vector of time fixed effects; and ε_{ijlt} are i.i.d. type I extreme value errors. In our main specification, X_{it} includes patient i's triage classes, gender, and the amount of time patient i has waited in the ED up to time t (Delay). Z_{jt} includes three indicators for how busy ward j is at time t: Busy 1 indicates whether the occupancy level of the current ward is above 99%; similarly, Busy 2 and Busy 3 indicate whether the occupancy level of the current ward is above 95% and 90%, respectively. W_t includes two indicators for two time windows, Morning and Evening, defined as 7am to 12pm and 9pm to 6am, respectively. These capture the differences in the bed management team's behavior before and after 12 pm (when the discharge of most patients begins) and at night. We allow all coefficients in Equation (1) to be specialty-specific.

Ward j belongs to the set of possible ward choices C_l, which is specific to the specialty l and includes a waiting option (Wait). Specifically, we include a ward in the choice set if we observe in the

data that more than 1% of type l patients are admitted to that ward. For each specialty l, without loss of generality, we choose one ward l_0 to be the reference ward – i.e., $\alpha_{jl_0} = \delta_{jl_0} = \gamma_{l_0} = \eta_{jl_0}$=0. The admission probability of patient i in specialty l to ward j is

$$P_{ijlt} = \frac{\exp(\alpha_{jl} + X'_{it}\delta_{jl} + Z'_{jt}\gamma_l + W'_t\eta_{jl})}{1 + \sum_{k \in C_l} \exp(\alpha_{kl} + X'_{it}\delta_{kl} + Z'_{kt}\gamma_l + W'_t\eta_{kl})}. \tag{2}$$

Each period t is two hours long. For patients who wait in the ED before being admitted to a ward unit, we do not observe the time when a "Wait" decision is made, but only the time when the admission decision is made. For those patients, we assume that a decision is made at their bed request time, and every two hours after that. For instance, if a bed request is made for patient i at 6am, and the patient is admitted into ward j at 9am, we assume that the bed management team decided to make her wait at 6am and to admit her to ward j at 8am, but the actual transfer takes place at 9am. In other words, for patients who wait less than two hours, we treat the waiting time as "negligible." We believe that this assumption is reasonable because previous studies in the literature using the same dataset indicate that the time for preparing the bed and transporting the patient could take up to two hours (Shi et al. 2016).

4.2. Estimation and results

We estimate the choice model for each patient's specialty l separately, and summarize the estimation results in Table 2. For each patient's specialty l, the set of possible ward choices C_l is different. We categorize the set of possible wards in C_l into primary and non-primary wards. As expected, the choice-specific coefficients are similar for those wards in the same category for each specialty. Therefore, for each specialty, we report the result for only one ward from each of the categories. We use a non-primary ward in each specialty as the reference alternative, or outside option. As a result, all estimated coefficients should be interpreted as relative to a non-primary ward. We report the four most important findings from Table 2 and describe them below.

First, the intercept indicates the baseline utility that the bed management team receives from keeping the patient waiting and admitting the patient into the primary ward, relative to admitting the patient into a non-primary ward. The results suggest that the primary ward is, in general, preferred to the non-primary ward. Waiting is also an attractive option since most primary wards often have high occupancy.

Second, the coefficients on the busyness levels of the wards suggest how much the bed management team takes into account current occupancy level of each ward when they make the assignment decisions. The estimates show that, in general, the busier a ward is, the less likely it is that the bed management team would assign the patient to that ward (Busy 1 is the busiest). This result suggests that load balancing is an important concern in the ward assignment decisions.

Third, the coefficients on Delay are, in general, negative and significant. This suggests that the longer the patient waits, the less likely the patient is to wait more. Moreover, the longer the patient waits, the more likely it is that the patient is going to be off-placed to a non-primary ward since the primary wards are more likely to be very busy. Our finding is consistent with part of the bed management team's general objective to reduce patients' admission delay.

Finally, the time of the day indicators are, in general, statistically significant. The evening variable indicates the time period after the main discharge window in the hospital. During the discharge period, most beds in preferred wards are assigned to patients as soon as or even before they become available. In the evening period, after discharges, there are often few beds available in primary wards. As expected, the primary wards generally have negative coefficients due to limited bed availability. Waiting also has a negative and significant coefficient and often greater magnitude. This is because the bed management team is aware that the number of discharges at night is extremely low, which means that waiting is likely the least attractive option in this period. The pattern in the morning period is the opposite because discharges will start after physicians check on patients on their morning rounds and beds in primary wards will soon become available. As a result, the coefficients of primary wards are positive and statistically significant. Moreover, in anticipation of the discharge peak in the early afternoon, waiting is an attractive option in the morning as well; that is, the coefficients of waiting are generally positive with higher magnitude and statistical significance.

5. Outcome analysis

In this section, we estimate the causal effect of ED admission delay and off-service placement on patient outcome for the focus-group patients. This allows us to provide a partial quantification of the cost of off-service placement. From the operational perspective, we are especially interested in the impact of off-service placement on patients' LOS. In particular, patients are often placed off-service during congested periods to reduce excessive delay. However, off-service patients may require longer LOS's, adding more workload to the already congested system. We emphasize that the estimated effect in this section is only the immediate effect of off-service placement on the misplaced patients. To assess the overall impact on the system, we need our stochastic model to quantify how this immediate effect is propagated through the inpatient ward network.

We assume that different factors would affect the log of the medical LOS, $\log(d_{los})$, through a linear model:

$$\log(d_{los,i}) = \beta_0 + \beta_{11} O_i + \beta_{12} D_i + Y_i' \beta_2 + T_i' \beta_3 + C_i' \beta_4 + \epsilon_i + \nu_i, \tag{3}$$

where $d_{los,i}$ is the medical length of stay for patient i. O_i is a binary variable for off-service placement, with $O_i = 1$ denoting that the patient is assigned to a non-primary unit. D_i is a binary

Table 2 Determinants of ward assignment decisions

	Cardio	GenMed	Surgical	Neuro	Gastro	Resp	Renal
Primary	2.454***	4.799 ***	1.662***	3.478***	3.847***	4.062***	4.924***
	(0.398)	(0.576)	(0.498)	(0.848)	(0.712)	(0.830)	(1.082)
Wait	3.310***	4.927***	3.075***	4.487 ***	3.659***	3.788***	4.714***
	(0.392)	(0.576)	(0.396)	(0.830)	(0.718)	(0.840)	(1.088)
Busy 1	-0.798***	-1.226***	-0.592***	-0.222	0.266	-0.117	-0.332†
	(0.126)	(0.110)	(0.135)	(0.166)	(0.200)	(0.287)	(0.180)
Busy 2	-0.577***	-0.585***	0.086	-0.249†	-0.736***	-0.679**	-0.178
	(0.090)	(0.067)	(0.110)	(0.135)	(0.132)	(0.237)	(0.159)
Busy 3	-0.466***	-0.310***	0.005	-0.371***	-0.443***	0.004	-0.378**
	(0.069)	(0.058)	(0.083)	(0.111)	(0.101)	(0.172)	(0.133)
Delay×Prim	-0.284**	-0.442***	-0.402***	0.002	-0.300***	-0.602***	-0.486**
	(0.094)	(0.066)	(0.076)	(0.170)	(0.122)	(0.159)	(0.148)
Delay×Wait	-0.850***	-0.858***	-0.935***	-0.438*	-0.830***	-0.875***	-0.955***
	(0.095)	(0.069)	(0.067)	(0.168)	(0.126)	(0.169)	(0.153)
Evening×Prim	-0.352	-1.557***	-0.667***	-1.123*	-0.521	-1.321**	-0.394
	(0.305)	(0.239)	(0.246)	(0.492)	(0.431)	(0.626)	(0.577)
Evening×Wait	-1.294***	-2.064***	-1.220***	-1.696***	-1.303**	-2.236***	-1.366*
	(0.302)	(0.240)	(0.202)	(0.468)	(0.436)	(0.646)	(0.592)
Morning×Prim	2.595***	1.263*	1.310***	1.775†	2.513*	-	1.946*
	(0.754)	(0.498)	(0.391)	(1.069)	(1.058)		(0.824)
Morning×Wait	3.437***	1.919***	2.273***	2.071	3.303**	-	2.884
	(0.754)	(0.499)	(0.353)	(1.062)	(1.060)		(0.829)
No. of obs.	3369	4826	2623	1444	1967	757	1368
Log-Likelihood	-4810.8	-5962.1	-3746.1	-2237.6	-2158.2	-804.13	-1328.5
Pseudo R^2	0.132	0.102	0.080	0.100	0.113	0.081	0.106

† : $0.05 < p \leq 0.1$,* : $0.01 < p \leq 0.05$, ** : $0.001 < p \leq 0.01$, *** : $p \leq 0.001$
The standard errors are reported in parentheses.

variable for admission delay. Specifically, $D_i = 1$ denotes that the admission delay is longer than four hours, with admission delay calculated as the time between when the decision to hospitalize the patient is made and when the patient is admitted into an inpatient ward unit. We choose a binary variable because we expect that the admission delay is likely to have a nonlinear effect on patient outcome. We test different threshold values for D_i in the online supplement. To differentiate this binary variable with the continuous admission delay, we refer to D_i as 'ED delay' for the rest of this section. Y_i is a vector of patient characteristics, including age, gender, ED triage score and medical specialty. T_i is a vector of variables related to admission and discharge times, including a binary indicator of whether the admission is in the evening (defined as 6pm to 6am the next day), a binary indicator of whether the admission is during the weekend, and the day of week on which the patient is discharged. C_i is a vector of variables capturing the system congestion and the physician workload during a patient's LOS. It includes the average occupancy of the assigned unit during the patient's LOS (DestAvgOccu), the attending physician's normalized workload during the patient's LOS (PhyAvgLoad), and the attending physician's normalized workload the day before discharge (PhyMinus1Load). Here, the workload is defined as the number of patients that the attending physician is treating, and the normalization is to divide the workload by the average workload of that physician in the entire year. ϵ_i is an error term that captures the effect of unobserved variables

that are correlated with both the LOS and the routing decision (i.e., O_i and D_i). ν_i is an error term that is uncorrelated with the observable variables.

In addition, we have a smaller dataset (five months of patient data) that contains more detailed patient diagnostic information. This includes primary and secondary DRG codes, number of operations, etc. Based on the additional information, we calculate the number of diagnostic codes, the number of operations and the van Walraven score (van Walraven et al. 2009). We include these comorbidity-related variables in Y_i as additional patient features when using the smaller dataset. The goal is to provide a more complete analysis (see Model 3 in Table 4).

5.1. Estimation strategy

The key estimation challenge comes from the ϵ_i term. We first note that the routing decisions are likely to be endogenous: some aspects of patient severity (e.g., complication of the case), which are not fully captured in the data, are likely to affect both the routing decision and patient LOS. For example, the bed management unit is more likely to keep a less complicated patient waiting or place her off-service, while a less complicated patient is also more likely to have a shorter LOS. If we are to estimate (3) directly, the ϵ_i term is likely to impose a negative estimation bias for β_{11} and β_{12} – i.e., we underestimate the effect of off-service placement and ED delay on LOS.

To solve the problem of omitted variable bias, we apply an instrumental variable (IV) approach. The IVs we propose are the primary ward occupancy and the hospital occupancy one hour before the admission hour. In particular, we define

$$Z_{i1} = \begin{cases} 1, \text{ if primary ward occupancy } > 0.97 \text{ one hour before the admission hour,} \\ 0, \text{ otherwise,} \end{cases}$$

and Z_{i2} is the average occupancy across all 13 wards one hour before the admission hour. Note that we also tried other thresholds for Z_{i1}, such as 0.95 and 0.99. The details can be found in the online supplement. The main insight is that the primary ward occupancy is likely to have a nonlinear effect on the off-service placement decision.

The first-stage and reduced-form regression equations take the form

$$\begin{aligned} \hat{O}_i &= \beta_0^O + \beta_{11}^O Z_{i1} + \beta_{12}^O Z_{i2} + Y_i'\beta_2^O + T_i'\beta_3^O + C_i'\beta_4^O + \epsilon_i^O, \\ \hat{D}_i &= \beta_0^D + \beta_{11}^D Z_{i1} + \beta_{12}^D Z_{i2} + Y_i'\beta_2^D + T_i'\beta_3^D + C_i'\beta_4^D + \epsilon_i^D, \\ \log(d_{los,i}) &= \beta_0 + \beta_{11}\hat{O}_i + \beta_{12}\hat{D}_i + Y_i'\beta_2 + T_i'\beta_3 + C_i'\beta_4 + \epsilon_i + \nu_i. \end{aligned} \tag{4}$$

A valid IV needs to satisfy two conditions: (C1) It must be correlated with the off-service placement decision O_i and/or ED delay D_i; (C2) It has no direct effect on $d_{los,i}$ other than through the the off-service placement decision O_i and/or ED delay D_i, conditional on the other covariates. We next discuss the validity of our IVs with respect to the two conditions.

When deciding whether to route a patient to her primary unit, the bed management unit needs to balance the benefit of primary admission with the opportunity cost of admitting a more complicated patient in the future. This tradeoff is especially relevant when the primary ward occupancy is high; thus, when it is high, the patients are more likely to be placed in off-service units. We use an indicator function to capture the business level of the primary ward since the effect of primary ward occupancy on off-service placement is highly nonlinear.

The overall hospital busyness level will likely affect both the routing decision and the ED delay. For ED delay, the more crowded the hospital is, the longer the delay is likely to be. The reason is twofold. First, it is hard to find a bed for the patient, as most wards are busy. Second, the transportation staff and other resources that are shared across different specialties are also likely to be busy, leading to longer delay.

Table 3 summarizes the estimation results from the first-stage linear regressions. Note that although O_i (Off-service) and D_i (ED delay) are binary variables, we use linear probability models for simplicity and interpretability, as well as for the correctness of inference in the second stage. We observe that the two IVs are positively correlated with the off-service placement decision and are very significant. The second IV is also positively correlated with ED delay.

Table 3 Fitted results for first-stage regression

Variables	Off-service (SE)	ED Delay (SE)
PriAdmOccuHigh	0.139 ***	−0.012
	(0.011)	(0.011)
HospAdmOccu	1.619 ***	0.619 ***
	(0.096)	(0.090)
R^2	0.268	0.044
No. of obs.	8642	8642

$^{+}: 0.05 < p \leq 0.1, *: 0.01 < p \leq 0.05, **: 0.001 < p \leq 0.01, ***: p \leq 0.001.$

The standard errors are reported in parentheses.

Next, we discuss the exclusion restrictions. System-level busyness measures have often been used as IVs in hospital settings (Chan et al. 2016, Song et al. 2018). The exclusion restriction relies on the randomness in patient arrivals. One potential concern is that ward occupancy may affect patient LOS directly (KC and Terwiesch 2009, 2012). Therefore, we control for the attending physician's workload during both the patient's stay and the day before discharge (PhyAvgLoad_i and PhyMinus1Load_i). We also control for the average occupancy of the assigned (destination) ward during the patient's stay (DestAvgOccu_i). Lastly, we focus on patients who spend at least two days in the hospital ($d_{los} > 2$), which helps further reduce the correlation between the load an hour before admission and the load during the patient's entire hospital stay. As the occupancy level typically varies on the time scale of hours, we observe a very low correlation between our IVs and each of the covariates listed above (PhyAvgLoad, PhyMinus1Load, DestAvgOccu).

5.2. Estimation results

The estimation results of the two-stage least squares regression (2SLS) decribed in Equation (4) are reported in Table 4. We also report the results from the direct estimation of (3) (Model 1). When comparing Model 1 to Model 2, we observe that neglecting the endogeneity of the routing decision introduces substantial negative bias – e.g., -0.004 versus 0.172, for off-service placement on $\log(d_{los})$. After correcting for the omitted variable bias using IV, we find a significant positive effect of off-service placement on $\log(d_{los})$ for the patients who are placed off-service. With an average LOS of four days, off-service placement increases d_{los} by 0.75 days, on average. However, we find no statistically significant effect of ED delay on $\log(d_{los})$. One explanation is that patients who are admitted to inpatient wards are, in general, not in urgent or critical conditions. Given that care has already been provided in the ED, delay in admission to the inpatient wards would not have a significant impact on their outcome. Indeed, even for certain ICU patients, previous work has shown that ED delay does not have an effect on their LOS (Chan et al. 2016).

Model 3 is fitted using the smaller dataset, in which we have more comorbidity information. We observe estimates for β_{11} and β_{12} that are similar to those in Model 2. We also note that, as expected, the variables related to comorbidity have a significant positive impact on LOS.

Table 4 2SLS models with different covariates

Variables	Model 1 OLS (without IV)	Model 2 2SLS (with IV)	Model 3 2SLS (with IV)
OffService (Fitted)	−0.004	0.172 *	0.207 *
	(0.011)	(0.069)	(0.089)
EDDelay (Fitted)	−0.024 $^{+}$	−0.157	−0.240
	(0.012)	(0.271)	(0.423)
van Walraven Score	–	–	0.006 ***
			(0.002)
No. of codes	–	–	0.027 ***
			(0.002)
No. of ops.	–	–	0.107 ***
			(0.014)
No. of obs.	8642	8642	4311

$^{+}: 0.05 < p \leq 0.1$,$*: 0.01 < p \leq 0.05$, $**: 0.001 < p \leq 0.01$, $***: p \leq 0.001$.

The robust standard errors are reported in parentheses.

6. The tradeoff between admission delay and off-service placement

To evaluate the tradeoff between admission delay and off-service placement, we vary the coefficients in the estimated choice model to generate different preferences for off-service placement in patient routing decisions and to evaluate their impact on system performance using our model. In other words, we vary the level of off-service placement that the hospital is willing to tolerate, and compute the implied average admission delay in the system taking into account the impact of off-service

slowdown on the entire inpatient ward network. Using the results, we construct a tradeoff curve (or efficiency frontier) that describes the full set of options of the off-service placement proportion and the average admission delay combinations that hospital managers face.

In the rest of this section, we first provide a detailed description about the construction of the tradeoff curves. We also highlight the substantial differences between the tradeoff curve derived using our estimated patient routing policy and that derived from the stylized policies commonly adopted in the literature. We emphasize that the differences in the tradeoff curves also lead to significant differences in the implied managerial insights. Second, we deviate from the current operations of our partner hospital – where there is a substantial mismatch between capacity and demand across specialties – and show that the general shape of the tradeoff curve remains similar under a more balanced capacity allocation. Importantly, by comparing different capacity reallocation strategies, we highlight the impact of network structure on the tradeoff between admission delay and off-service placement. Finally, we discuss the impact of the slowdown effect on the shape and location of the tradeoff curves. Our results show that the slowdown effect can reduce, and sometimes completely cancel out, the benefit of capacity pooling. It is, thus, important for managers to understand where their hospital currently stands on the tradeoff curve to evaluate whether or not more off-service placements are beneficial.

6.1. The construction of the tradeoff curve

To construct the efficiency frontier, we first multiply the estimated coefficients of Delay$\times$Wait and Delay$\times$Prim for all specialties in the choice model by a common factor c_1 and vary the value of c_1 to allow different preferences of the bed management team for off-service placement. As both coefficients are negative in the estimated choice model, when we increase c_1, the preference for waiting and for assigning the patient to the primary ward becomes weaker. In other words, we increase the hospital's willingness to place the patient off-service, conditional on all the other determinants in the model. Then, for a given c_1, we use the new patient-routing policy function and our stochastic network model to simulate the inpatient flow; based on this, we evaluate the off-service placement proportion and the average admission delay in the network. In subsequent analyses, we refer to the baseline case as the current state of our partner hospital – i.e., the system performance under the originally estimated choice models, where $c_1 = 1$. Note that we apply c_1 for the choice models we estimated for all ED admissions. For the rest of this section, we report the performance measures for the focus group only. The performance changes for the other two groups of ED patients (short-stay observational and long-stay) are similar. On the other hand, as we do not change the routing policy for the other sub-classes of patients, such as Elec, ICU, etc., the performance measures, mainly the off-service placement proportion, for these groups of patients stay the same.

Figure 7 shows the tradeoff curve between average admission delay (x-axis) and off-service placement proportion (y-axis) for the focus group, where the multiplier c_1 varies between 0.05 and 32. When $c_1 = 0.05$, on the extreme right end of the tradeoff curve, the hospital has a strong preference for placing patients in the primary ward. In this case, the off-service placement proportion is reduced from 24% in the baseline scenario to 21%. However, the cost of this reduction is a 56% increase in the average admission delay, from 1.42 hours in the baseline case to 2.21 hours. On the extreme left end of the curve, when $c_1 = 32$, the hospital has a strong preference for off-service placement. In this case, the off-service placement proportion is 47%, but the average admission delay is shortened to 1.07 hours.

The curve illustrates that, when the hospital is operating on the right end of the curve – i.e., when the average admission delay is high and the off-service placement is low – a small increase in the off-service placement proportion can lead to a substantial reduction in admission delay. In other words, off-service placement is a highly effective control for managing patient waiting time in the network in this region.

Meanwhile, the benefit of off-service placement diminishes quickly as we move towards the left end of the curve and the off-service proportion increases. For example, in the baseline scenario, increasing the off-service proportion further does not help the hospital improve the average admission delay significantly. In other words, off-service placement is no longer an effective tool to manage delay in the network.

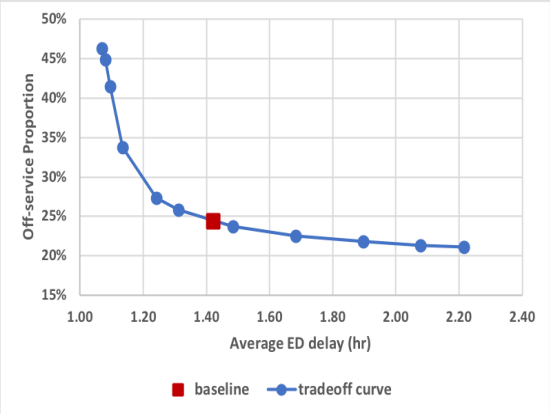


Figure 7 The tradeoff curve

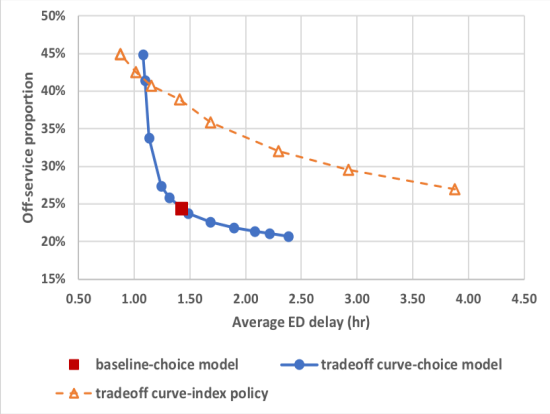


Figure 8 Comparison between estimated choice model policy and index policy

Importantly, the curve provides hospital managers with the full set of possible choices in terms of the combinations of off-service placement proportion and average admission delay in the inpatient ward network. In addition, instead of imposing a particular objective for the hospital managers, the curve allows them to choose the appropriate tradeoffs for different scenarios.

6.1.1. Comparison to tradeoff curve using index policy To emphasize the benefit of our proposed data-driven approach, we compare the tradeoff curves derived using our estimated policy from the choice model and the index-based policies introduced in Section 3.3. The dashed curve in Figure 8 corresponds to the tradeoff curve derived by using a stylized index-based routing policy. Note that we incorporate both the load balancing and the time of day effect in this policy. We incrementally change the utilization thresholds U^m and U^e to generate different levels of off-service placement proportions in the system. We also compute the corresponding average admission delays using our simulation model. The solid curve is the tradeoff curve derived using our estimated choice model.

There are two main differences between the two curves. First, the tradeoff curve for the index-based policy lies above the curve for the estimated choice model. In other words, for any given level of off-service placement proportion, the index-based policy leads to a much higher average admission delay than the estimated choice model policy does. This is because, as described in Section 3.3, the index-based routing policy is not flexible enough to dynamically adjust the thresholds with the overall workload of the system. Second, the dashed curve has a more linear, flatter shape than the solid curve, especially in the region where the off-service placement proportion is relatively high. In particular, the slope of the solid curve around the current state of the hospital (the baseline) is about four times the slope of the dashed curve. This finding suggests that the index-based policy does not do a good job capturing the diminishing returns in terms of the reduction in the average admission delay when increasing the off-service placement proportion. The reason is that the index policy simply reduces the occupancy thresholds for admitting patients across all wards, but ignores the delicate interplay between off-service placement and other determinants of the routing decision. For example, as the estimated choice model suggests, when the occupancy of a ward increases, the disutility of the off-service placement in that ward increases in a nonlinear fashion. The higher the occupancy of a ward, the less likely it is that an off-service patient will be placed in that ward. The index policy is not flexible enough to capture this nonlinear relationship and, thus, predicts a more linear tradeoff between the off-service placement proportion and the admission delay. More importantly, it may mislead hospital managers to overinvest in reducing off-service placement and, thus, suffer excess admission delay.

6.2. Capacity reallocation and network effect

In the previous section, we investigated the tradeoff between admission delay and off-service placement by plotting the tradeoff curve under the current operations of our partner hospital. We acknowledge that the shape of the curve relies heavily on the operations of the particular hospital. To make our findings more applicable to other hospital settings, we study how the tradeoff between

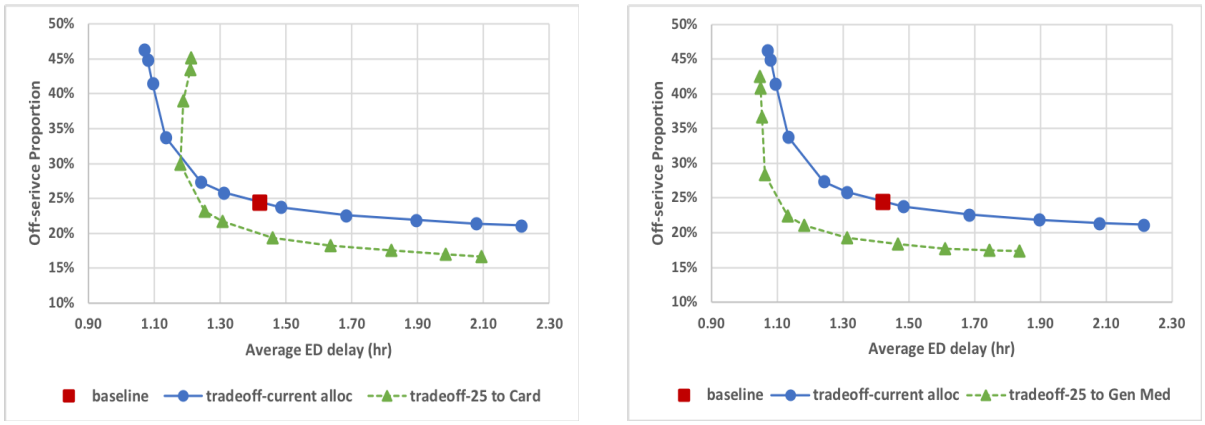
admission delay and off-service placement changes when we deviate from the observed state of our partner hospital. In particular, since there is substantial mismatch between capacity and demand in our partner hospital, off-service placement is applied to cope with both capacity mismatch and stochastic fluctuations in the system. In this section, we study how the tradeoff curve changes when we reallocate the bed capacity across the specialties to better match capacity with demand. More importantly, by comparing different capacity reallocation strategies and their resulting tradeoff curves, we highlight the impact of the ward network structure on determining the tradeoff between admission delay and off-service placement.

Under the current state of our partner hospital, Table 1 shows that Card and Gen Med are both heavily overloaded, while Ortho is underutilized. To gain insights from a more balanced capacity allocation across the specialties, we consider two simple reallocation strategies. First, we reallocate 25 beds from Ortho to Card, which reduces the nominal utilization of Card to 85%. In the second strategy, we reallocate 25 beds from Ortho to Gen Med, which reduces the nominal utilization of Gen Med to 87%. We compute the tradeoff curves under both scenarios and show the results in Figure 9.

First, we find that, in both cases, the general shape of the tradeoff curve stays the same after the capacity reallocation. That is, under balanced capacity allocation, hospital managers face a similar tradeoff between admission delay and off-service placement. This is because, even without capacity mismatch, hospital managers still rely on off-service placement to deal with the stochastic fluctuations in the system. As a result, our findings in Section 6.1 apply to more general settings in which the mismatch between capacity and patient demand across specialties is not as severe as in our partner hospital. Notably, the return to off-service placement in terms of admission delay reduction still diminishes quickly, as the off-service proportion increases, indicating the importance for hospital managers to know where the current operation lies on the tradeoff curve.

Second, we observe that the two reallocation strategies lead to different tradeoff curves. In particular, in Figure 9b reallocating capacity to Gen Med leads to a more inward-positioned efficiency frontier than reallocating capacity to Card, as shown in Figure 9a. This is because the network effect of the two reallocation strategies differs. In particular, Gen Med wards are much better connected than Card in the inpatient ward network. Gen Med wards often receive off-service patients from other medical specialties, while Card wards rarely receive off-service patients. As a result, the 25 beds assigned to Gen Med provide additional benefits to other specialties on the network because Gen Med can now accept more off-service patients. Thus, reallocating beds to Gen Med leads to a greater improvement of the efficiency curve.

Figure 9a also shows that, as the off-service placement proportion increases to above 30%, the tradeoff curve starts to bend – i.e., the average admission delay increases as the off-service proportion further increases. This is the result of reallocating capacity to the less-connected Card



(a) Reallocate capacity to Card

(b) Reallocate capacity to Gen Med

Figure 9 Tradeoff curves for different allocation strategies

wards interacting with the off-service slowdown effect. In particular, the additional beds assigned to Card wards provide little benefit to other wards through the network since Card wards rarely receive off-service patients. Meanwhile, the effect of the longer LOS of off-service patients in other wards is propagated through the inpatient ward network, generating higher overall workload to the system – i.e., higher system congestion. We postpone the more detailed discussion of the impact the off-service slowdown on the tradeoff curve to the next subsection.

We conclude this section with a brief summary of managerial insights. Better matching capacity and demand between specialties can substantially improve the efficiency frontier of the off-service placement proportion and the average admission delay. When reallocating capacity from the underloaded specialty to the overloaded specialty, it is important to take the network structure into account when off-service placement is present. We use the example of Card and Gen Med wards to show that it is more beneficial to allocate capacity to wards of well-connected specialties, as they receive more off-service patients. This results in greater benefit to other specialties and the overall system.

6.3. The off-service slowdown and the tradeoff curve

We analyze the impact of the off-service slowdown on the tradeoff curve in this section. Off-service placement is employed as a control to reduce excess admission delay in general. However, since off-service patients have longer LOS, it leads to higher overall workload for the system, which, in turn, may block the admission of future patients. In other words, it may offset some of the benefits of resource pooling generated by off-service placement. More importantly, this effect of the off-service slowdown can spread through the inpatient ward network. This requires a complete analysis of the entire network, which can not be achieved without our modeling framework.

In this section, we first analyze the effect of the off-service slowdown in the setting of our partner hospital. Then, we investigate the problem in more general settings that are closer to the scenarios

in other large hospitals studied in the literature. Finally, we study the effect of off-service slowdown at the specialty level, which highlights the interplay between the slowdown effect and the network effect.

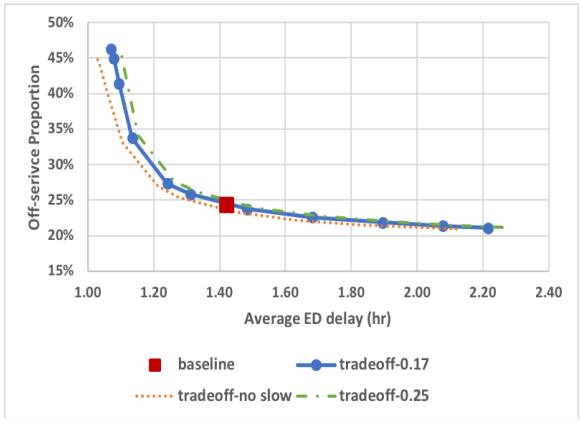
6.3.1. The off-service slowdown in our partner hospital Figure 10a plots the tradeoff curves for different values of the slowdown factor β_{11}, defined in Equation (4). The parameter β_{11} is estimated for our partner hospital to be around 0.17. A similar magnitude of the slowdown factor is also observed in a large US teaching hospital (Song et al. 2018). To understand the effect of off-service slowdown on the tradeoff curve, we investigate two additional scenarios: one in which $\beta_{11} = 0$, which indicates zero slowdown; and one in which $\beta_{11} = 0.25$, which is the estimated value for some specialties with a larger slowdown effect in our partner hospital (see the online supplement).

Since the three tradeoff curves in Figure 10a are very close to each other, the slowdown factor seems not to play an important role in determining the tradeoff curve. This can be due to two factors. First, the overall utilization of our partner hospital is 87%, which is relatively low. Therefore, the additional workload generated by off-service patients may not be substantial enough to affect the tradeoff curve. Second, the group of patients who potentially experience off-service slowdown (the focus group) accounts for only 33% of the total patient population we model. The magnitude of the effect may not be big enough in our partner hospital due to the relatively small size of the focus group.

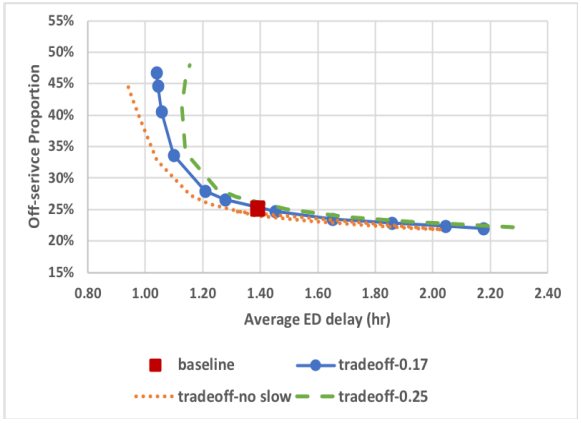
6.3.2. The off-service slowdown in more general settings Given that the utilization and the proportion of the focus group in other hospitals studied in the literature are generally higher (Copenhaver et al. 2019, Song et al. 2018), we deviate from the operating environment of our partner hospital by increasing the utilization and the proportion of patients who are subject to off-service slowdown in the system, which provides a more general analysis of the impact of the slowdown factor.

First, in Figure 10b, we increase the proportion of focus-group patients from 33% in the baseline case to 45%. The three curves in the figure correspond to different values of the slowdown factor. We observe that when the off-service placement proportion is small, there is still not much difference among the three curves because not many patients are affected by off-service slowdown. However, when the off-service proportion is large, we start seeing bigger differences among the three curves. In particular, for $\alpha = 0.25$, when the off-service placement proportion is high ($> 40\%$), we see no improvement in admission delay when further increasing the off-service placement proportion. In fact, the average admission delay even increases slightly.

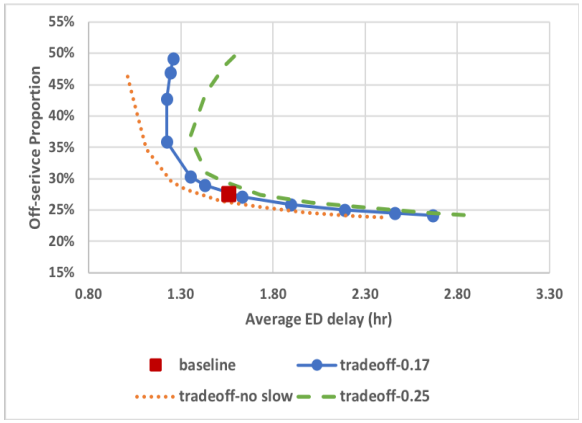
Second, compared with Figure 10b, we increase the utilization of the system in Figure 10c from 87% to 94%, while keeping the focus-group patient population at 45%. We observe that the



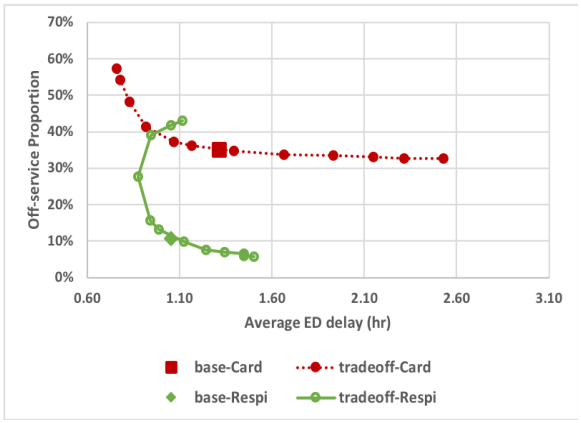
(a) Different slowdown factors



(b) Focus group 45%



(c) Focus group 45%, increase load



(d) Tradeoff curves for Card and Resp; Focus group 45%, slowdown factor 0.25.

Figure 10 The tradeoff curve between the average admission delay and the off-service proportion under different values of the slowdown factor and proportions of focus group.

slowdown factor has a bigger impact on systems with a heavier load. In particular, as the off-service placement proportion increases, the reduction in admission delay diminishes much more significantly compared to that in Figure 10b. Moreover, the two tradeoff curves with positive β_{11} start to bend when the off-service placement is above 35%. This suggests that, in this region, the negative effect of off-service placement on the admission delay, or the higher overall workload on the system generated by the off-service slowdown, completely cancels out the benefit of resource pooling. In other words, the off-service placement can no longer be employed as a control to reduce excess admission delays. In those cases, it is critical for managers to know where the current hospital operation lies on the tradeoff curve in order to avoid inappropriate management decisions.

6.3.3. The interaction of off-service slowdown and network effect To further investigate the negative effect of off-service slowdown on the average admission delay, we construct the tradeoff curves with a high slowdown effect, $\beta_{11} = 0.25$, at the specialty level. Figure 10d plots the

tradeoff curves for two specialties: Card and Resp. We set the focus group of patients who are subject to the off-service slowdown to be 45%. This corresponds to the scenario for the dashed line in Figure 10b. Figure 10d shows that there are significant differences in the tradeoff curves between Card and Resp. In particular, Resp suffers heavily from the negative effect of off-service placement on admission delay when the off-service proportion is above 18%. This is because Resp wards receive a large number of off-service patients from other specialties. When the off-service proportion increases, the Resp wards are heavily affected by the higher workload generated by the longer LOS of the off-service patients. Moreover, this higher workload also leads to more Resp patients being blocked from admission into their primary ward and, thus, being placed off-service, which further increases the workload of the overall system. This snowball effect leads to a substantial negative effect of off-service placement on admission delay, as shown in Figure 10d. On the other hand, Card wards play a completely different role on the inpatient ward network: they rarely receive off-service patients and, thus, are less affected by off-service slowdown. More specifically, many Card patients may be placed in off-service wards (large off-service proportion), but the capacity saving is not used to help other specialties. In other words, patients in the Card wards are mainly primary patients and do not experience any off-service slowdown effect. As a result, Card wards are not affected as much as Resp wards by the slowdown effect.

To support the above argument and highlight the effect of the off-service slowdown, we perform a detailed analysis of the Resp ward NW44 when $\beta_{11} = 0.25$ vs. $\beta_{11} = 0$. When $\beta_{11} = 0.25$, if we increase the overall off-service proportion from 26% to 46% by adjusting c_1, the proportion of off-service patients in NW44 increases from about 30% to more than 60%. Due to the off-service slowdown, the occupancy rate of NW44 increases to 94%. This higher occupancy rate results in more Resp patients blocked from admission to NW44 and placed off-service. The average admission delay for Resp patients increases to 1.05 hours, and the off-service proportion increases to 42%. By contrast, when $\beta_{11} = 0$, if we increase the overall off-service proportion from 26% to 46%, the occupancy of NW44 remains at 85%. The average admission delay for Resp patients is as low as 0.70 hour, and the off-service proportion for Resp patients is 38%.

To summarize, by conducting the specialty-level analysis when the slowdown effect is high, we find that some specialties can be so heavily affected by off-service slowdown that the benefit of resource pooling is completed canceled out by the higher overall workload. This leads to higher admission delay when the off-service placement proportion further increases beyond a certain threshold. We also find that there is a large heterogeneity across specialties in terms of how the off-service slowdown affects the tradeoff curve. These findings highlight the importance of the interplay between the inpatient ward network effect and the off-service slowdown effect. The findings also illustrate the importance of plotting the tradeoff curves using the empirical and modeling

tools we proposed in this paper for hospital managers to make intelligent decisions on managing congestion in the inpatient flow.

7. Conclusion

In this paper, we build a high-fidelity stochastic model to quantify the tradeoff between off-service placement and admission delay in inpatient wards. We note that, even though off-service placement can help create more resource pooling in the network of inpatient wards, the effect is diminishing. Moreover, off-service slowdown can offset the benefit of resource pooling when the off-service proportion is high. We also study the impact of the network structure on the tradeoff, and propose capacity reallocation strategies to improve the efficiency frontier.

In this paper, we developed a methodology framework by building a high-fidelity stochastic model to capture the underlying physics of patient-flow dynamics, and using choice model and outcome analysis to tackle several model estimation challenges. This framework can be applied to other health care delivery systems. For example, Chan et al. (2016) estimate the causal effect of waiting on ICU patient outcomes. If we combine that with ICU admission/scheduling policies estimated from the data and a stochastic model describing patient-flow dynamics through the ICU, we can provide a reliable evaluation of improvement strategies there. Similar examples can be found in KC and Terwiesch (2017) for surgical scheduling, Song et al. (2015) for ED case management, and Chan et al. (2018) for step-down units.

Our work in this paper had some limitations. First, in our partner hospital, we rarely observed patients being moved back to their primary units after they had been placed off-service initially. The reason is that transferring patients between wards can be cost-inefficient (staff to transfer the patient, bed cleaning, etc.); impose safety concerns, such as discontinuity of care due to patient handover; and can hurt patient experience in the hospital. While many hospitals in the U.S. and Europe follow a similar practice of not transferring patients from the off-service unit back to the primary unit when a bed becomes available, we are also aware that a subset of hospitals do transfer patients. For hospitals that frequently practice "transfer-back," our analysis does not directly apply. New routing policies need to be estimated in those settings.

Second, due to the weather conditions in Singapore, patient arrivals in our partner hospital do not exhibit significant seasonality patterns. For other hospitals, at which patient arrivals do experience strong seasonality, one can do a separate estimation and simulation for each season. The potential improvement strategies may also be dependent on seasonality effects.

References

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

Armony, Mor, Shlomo Israelit, Avi Mandelbaum, Yariv Marmor, Yulia Tseytlin, Galit Yom-Tov. 2015. Patient flow in hospitals: A data-based queueing perspective. *Stochastic Systems* **5**(1) 146–194.

Ata, B., J.A. Van Mieghem. 2009. The value of partial resource pooling: Should a service network be integrated or product-focused? *Management Science* **55**(1) 115–131.

Best, Thomas J., Burhaneddin Sandkç, 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.

Carr, Brendan G., Judd E. Hollander, William G. Baxt, Elizabeth M. Datner, Jesse M. Pines. 2010. Trends in boarding of admitted patients in US emergency departments 2003-2005. *Journal of Emergency Medicine* **39**(4) 506–511.

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

Chan, Carri W, Linda V Green, Lijian Lu, Suparerk Lekwijit, Gabriel Escobar. 2018. Assessing the impact of service intensity on customers: An empirical investigation of hospital step-down units. *Management Science* **65**(2) 751–775.

Copenhaver, M., A.V. Berg, A.C. Zenteno Langle, R. Levi, P. Dunn. 2019. Optimizing hospital-wide bed allocation. Presentation at POMS 2019, Washington D.C.

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

Dong, J., O. Perry. 2018. Queueing models for patient-flow dynamics in inpatient wards. 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.

Goulding, L., J. Adamson, I. Watt, J. Wright. 2012. Clinical outcomes in medical outliers admitted to hospital with heart failure. *BMJ Quality and Safety* **21**(3) 218–224.

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

Guajardo, Jose A, Morris A Cohen, Serguei Netessine. 2015. Service competition and product quality in the us automobile industry. *Management Science* **62**(7) 1860–1877.

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.

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

Han, Shasha, Shuangchi He, Hong Choon Oh. 2016. Models for hospital inpatient operations: A data driven optimization approach for reducing ED boarding times. Presentation at INFORMS 2016, Nashville, TN.

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

Hoot, NR, D. Aronsky. 2008. Systematic review of emergency department crowding: Causes, effects, and solutions. *Ann Emerg Med* **52** 126–36.

Jacobson, E.U., N.T. Argon, S. Ziya. 2012. Priority assignment in emergency response. *Operations Research* **60**(4) 813–832.

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

KC, D. S., C. Terwiesch. 2017. Benefits of surgical smoothing and spare capacity: An econometric analysis of patient flow. *Production and Operations Management* **26**(9) 1633–1684.

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.

Kuntz, L., S. Scholtes, S. Sulz. 2019. Separate and concentrate: Accounting for patient complexity in general hospitals. *Management science* **65**(6) 2482–2501.

Lim, SC, V Doshi, B Castasus, JKH Lim, K Mamun. 2006. Factors causing delay in discharge of elderly patients in an acute care hospital. *Annals-Academy of Medicine Singapore* **35**(1) 27.

Phillips, Robert, A Serdar Şimşek, Garrett Van Ryzin. 2015. The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science* **61**(8) 1741–1759.

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, Pengyi, M. Chou, J.G. Dai, D. Ding, J. Sim. 2016. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science* **62**(1) 1–28.

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

Song, H., A.L. Tucker, R. Graue, S. Moravick, J.J. Yang. 2018. Capacity pooling in hospitals: The hidden consequences of off-service placement. Working paper.

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.

Stylianou, N., R. Fackrell, C. Vasilakis. 2017. Are medical outliers associated with worse patient outcomes? a retrospective study within a regional NHS hospital using routine data. *BMJ Open* **7**(5).

van Walraven, C., P.C. Austin, H. Quan, A.J. Forster. 2009. A modification of the elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med Care* **47**(6) 626–633.

Online Appendix for "Off-service Placement in Inpatient Ward Network: Resource Pooling versus Service Slowdown"

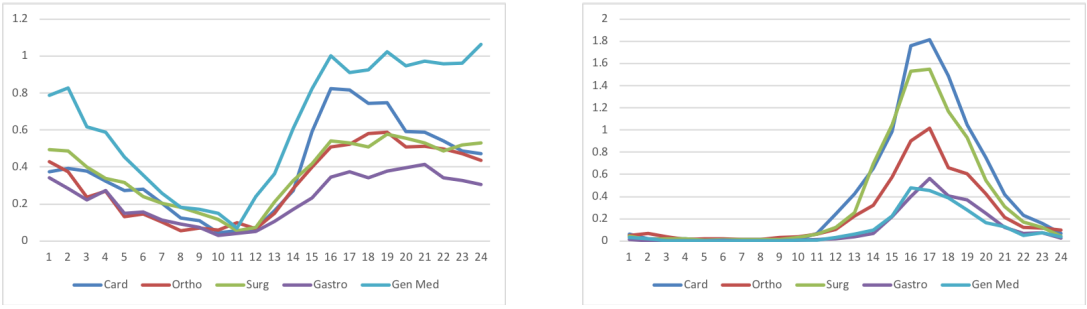
Authors' names blinded for peer review

This document serves as an Online Appendix for the main paper. Section 1 provides more details about the dataset we obtained from our partner hospital. Section 2 provides more details of the stochastic model including how we populate the model parameter. Robustness analysis for the choice model estimation and outcome analysis are provided in Section 3 and 4 respectively.

1 Data description

Our data comes from a large teaching hospital in Singapore, spanning from January 1 to December 31, 2010. The total number of patient admissions is 92081. To study inpatient ward management, we use a subset of the total patient admissions where we exclude admissions to non-inpatient wards and certain highly specialized specialties. In particular, we exclude patients who only visited outpatient centers such as dental clinics, outpatient surgery wards, endoscopy center. We also exclude patients from Obstetrics and Gynaecology (OB/GYN), Oncology, and specialties with a very small inpatient volume such as Dental and Eye. The reason we exclude OB/GYN and Oncology is that their patients population are very different from other specialties. These patients require highly specialized treatment and care. Thus, there is little interaction between these two specialties and other specialties, i.e. these patients are rarely placed off-service, and their primary wards rarely receive off-service placements from other specialties. Lastly, we exclude patients who were admitted to private wards. Admissions to these private wards require more expensive private insurance (instead of the universal insurance provided by the government). In addition, these wards usually have much lower occupancy.

Our selected sample contains 34030 patient admission records (93.2 admissions per day on average) from eight specialties to thirteen inpatient wards. The eight specialties are General Medicine (Gen Med), Gastroenterology (Gastro), Neurology (Neuro), Renal Disease (Renal), and Respiratory (Resp), Surgery (Surg), Cardiology (Card), and Orthopedic (Ortho). The first four specialties all belong to the Medicine cluster. The 13 specialties have different sizes of practice, with Card being the largest (19.8 admissions per day on average) and Resp being the smallest (4.8 admissions per day on average). See Table 1 in the main paper for a summary of the load of different specialties. Patients in this sample are admitted from five different sources: Emergency Department (ED), Elective (Elec), Intensive Care Unit (ICU), Transfer (Trans), and Others (e.g., same-day admissions who go through surgery first and then admitted to inpatient wards). Note that we count each transfer as a separate admission record. As shown in Figure 1 of the main paper, the majority (53%) of patients are admitted from ED – we refer to them as "ED admissions" in the rest of this document. This is mostly due to the public health care structure in Singapore: our partner hospital is state-owned, whose primary objective is to treat the patients with the greatest needs. As a result, the elective surgery waiting lists are usually exceptionally long. Many patients who wish to receive surgery or treatment in a short period of time would rather visit the ED instead of waiting on the elective list. Such phenomena have been discussed



(a) ED admissions

(b) Non-ED admissions

Figure 1: Hourly admission rates for different hours of the day

extensively in countries and regions with similar public funded health care systems, such as Canada, U. K., etc. [1, 2].

1.1 Patient-level information

For each admission to the inpatient wards, we have the following information in our data set.

- **Patient characteristics:** These include age, gender, ED triage score (for ED admissions only), medical specialty, primary diagnosis and billing code, admission source , admission ward ID, attending physician ID, etc.
- **Time stamps for in-hospital activities:** These include admission to a ward, transfer between wards, discharge from a ward, bed request and ED discharge (for ED admissions only).

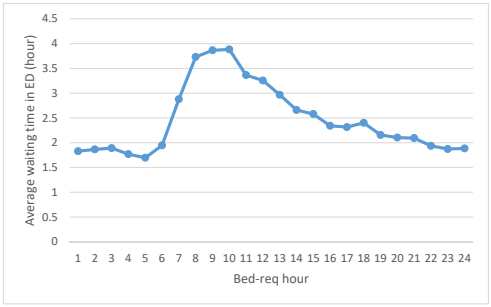
Based on these information, we can calculate **workload related measures**. These include the occupancy of different wards at the hourly level, the attending physician's workload at the daily level. We can also calculate **patient-level performance measures**, such as the admission delay (for ED admissions only) – defined as the time between the bed-request time and the ward admission time, whether the patient is placed off-service, etc.

1.2 Time-dependent inpatient flow dynamics

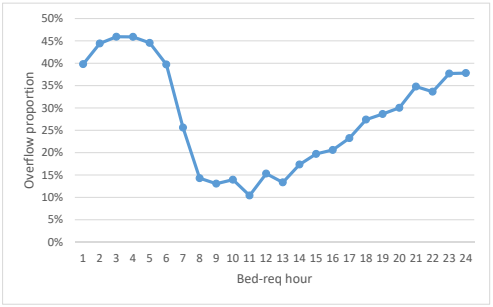
The hospital operations in a highly non-stationary environment. Understanding the time-dependent system dynamics is important to model the inpatient flow.

Figure 1 shows the number of admissions to wards in each hour for the five largest specialties (Card, Ortho, Surg, Gen Med and Gastro) for ED admissions (2a) and non-ED admissions (2b). We can see that the admission rates are time-varying, and the variability patterns for ED admissions and non-ED admissions are different. In particular, non-ED admissions are mostly clustered in the afternoon when beds are becoming available after the block discharges as demonstrated in Figure 4(b) in the main paper.

Figure 2 summarizes the average admission delay (a) and off-service placement proportion (b) by patients' bed request hours for all ED admissions. We observe that these performance measures varies a lot for different hours of the day. This is mostly due to the special discharge pattern in the hospital. As seen from Figure 4(b) in the main paper, most discharges take place in the afternoon and there are almost no discharges late at night or early in the morning. Thus, we observe that patients who request beds in the morning experience a longer waiting time on average. We also observe that the



(a) Average ED delay by bed-request hour



(b) Off-service placement proportion by bed-request hour

Figure 2: Time-dependent performance by bed-request hour for ED admissions.

bed management team places a higher proportion of patients off-service in the late night and early morning (9pm to 7am) hours, while during the day, the off-service proportion is much lower.

2 Stochastic model

In this section, we provide the details of our stochastic network model. We build a multi-class multi-pool queue with time-varying arrival rates. Key components in our model include choice-model based routing and two-time scale service time.

In this model, there are $J = 13$ server pools, representing the 13 inpatient wards in our partner hospital. Pool j has N_j servers (beds), which are estimated from the data; see Table 1 in the main paper. There are $I = 8$ customer classes, representing the 8 medical specialties. Under each specialty, we further divide patients into $S = 7$ sub-classes:

1. Focused group ED patients: patients who are admitted from ED and have a LOS between 2 and 7 days. We call this group the focused group because these are the patients we focus on when changing routing policies. We also focus on estimating the impact of off-service placement on patient's LOS for this group only. This group constitutes 63% of the ED admissions and 33% of all the admissions we consider.

2. Observational ED patients: patients who are admitted from ED and have a short LOS (0 or 1 day). These patients have relatively mild medical conditions. Indeed, most of them undergo simple evidence-based protocols. This group constitutes 26% of the ED admissions and 14% of all the admissions we consider.

3. Long-stay ED patients: patients who are admitted from ED and have a long LOS: more than 7 days. This group constitutes 11% of the ED admissions and 5.8% of all then admissions we consider.

4. Elective patients: patients who are admitted through elective referrals. They account for 14% of all the admissions we consider.

5. ICU patients: patients admitted/transferred from ICU. They account for 9% of all the admissions we consider.

6. Other patients: patients who are admitted from other sources such as same day surgery. They account for 7% of all the admissions we consider.

7. Transfer patients: patients who are initially admitted from ED to wards but later have in-hospital transfer (mostly between the ICU and wards). This group constitutes 17% of all the admissions we consider.

We assume the arrival process of each specialty i ($i = 1, \ldots, 8$) and subclass s ($s = 1, \ldots, 7$) follows a time-inhomogeneous Poisson process with its corresponding arrival rate function $\lambda_{i,s}(t)$. To capture the hour of the day effect, we assume the $\lambda_{i,s}(t)$ is periodic function with period equal to 1 day. For ED patients, we use the bed-request time as the arrival time. For non-ED patients, we use their admission time as the arrival time.

While the arrival rate is rather straightforward to estimate, the calibration of the routing policy and the service times are more involved. We next elaborate on each of them.

2.1 Ward assignment

We fit a ward assignment policy for each specialty and each subclass of patients from data. We use different strategy for different subclasses.

For ED admissions, we fit a choice model for each subclass. We refer to Section 4 in the main paper for details of the choice model. To reduce simulation noise, we impose a 0.15 cutoff for the p-value when using the fitted choice model coefficients, i.e., if the p-value is larger than 0.15, we set the coefficient to be 0. We also tested other cutoff values and found the results are quite robust to different reasonable cutoff values.

We denote the choice probability for specialty i subclass s at time t as

$$p_{i,s}(t) = \left(p^0_{i,s}(t), p^1_{i,s}(t), \ldots, p^J_{i,s}(t)\right).$$

Here, choice j, $j = 1, \ldots, J$, corresponds to ward j, and choice $j = 0$ corresponds to the waiting option. For patient l of specialty i and subclass s, at the decision epoch t, we calculate the set of probabilities $\{p^j_{i,s}(t)\}$ from the fitted choice model, with the following information gathered at t from the simulation model:

- z_{jt}, the utilization of each ward $j = 1, \ldots, J$ at time t; the utilization for the waiting option is always set to be 0;
- x_{lt}, the amount of time that the patient has waited till t;

We then generate a decision from the choice set $\{0, 1, \ldots, J\}$ according to the probability $p_{i,s}(t)$.

If the waiting option is chosen, the patient waits in a buffer dedicated to her specialty and subclass; otherwise, she is admitted to the chosen ward j. There is a little caveat here we need to take special care of. That is assigning a patient to a full ward. We first note that this only happens very occasionally. Specifically, the negative coefficient associated with high ward utilization in the choice model leads to very small probabilities of assigning a patient to a full ward. When this does happen, to ensure the simulated ward assignments are consistent with the choice model, we allow employing "surge capacities" in assigned ward. Surge capacities such as trolley beds not uncommon in practice and are indeed used in our partner hospital. Furthermore, the simulation results show that the average bed utilization for each ward from our stochastic model matches the empirical utilization well (see Figure 5 in the main paper).

Lastly, we discuss how the decision epochs for each patients are specified. Each patient gets two types of decision epochs: The first type is a set of pre-specified epochs, which are 0, 2, 4, ..., hours from the patient's bed-request time (arrival time). These decision times are chosen to be consistent with the estimation of the choice model. The second type of are epochs triggered by the decision times of other patients. Specifically, when a patient is at her decision epoch of the first type, we also

trigger a decision epoch for all other patients who are of the same specialty, same subclass, and arrive before this patient. In this case, we generate a decision for each of these patients in the sequence of their arrival times. The second type of decision times are added to capture the preference of staying close to First-in-First-out for fairness.

For non-ED admissions, we use randomized ward assignment policy according to their empirical ward-assignment distribution fitted from data. We note that these patients do not have a waiting option. This is because most of these admissions are scheduled. In particular, these patients often arrive in the afternoon hours when peak discharge takes place. Thus, very few of them have to wait. In addition, we also do not have waiting time information for these patients. We do note fit a choice model for these sub-classes as we will not change their bed-assignment decisions when constructing the trade-off curve. Indeed, for non-ED admissions, we do not face as much a trade-off between admission delay and off-service placement as those ED admissions.

2.2 Service time

When a patient is admitted to a ward, she stays in the ward until being discharged or transferred. The LOS of the patient is referred to as the *service time* in the simulation model.

We employ a two time-scale (day versus hour) service time model. Specifically, for each patient, we generate d_{los} and h_{dis} upon her admission:

- d_{los} is the integer number of days the patient stay in the ward. It counts the number of 10am's the patient spent in the ward.
- h_{dis} is the discharge delay. It is the time between 10 am on the day of discharge and the actual departure time of the patient.

Patient departures are then generated as the follows: at 10am each day (the end of rounding in our partner hospital), we check all patients who are currently in service. If the number of days she spent (from the admission day till today) reaches d_{los}, this patient will be discharged that day, after a delay of h_{dis} hours; otherwise, the patient stays in service.

We estimate the distribution of d_{los} for each specialty and subclass separately. The d_{los} for the focused group patients is fitted using a subset of data as detailed below to account for the off-service slowdown. The d_{los} for each of the other sub-classes is fitted from all admissions in that specialty and subclass. This is because for these sub-classes, we will not change their bed assignment decisions when constructing the trade-off curve. In addition, for some sub-classes, e.g. the observational ED patients, we suspect that off-service placement will not have a big impact on their LOS.

Incorporate off-service slowdown. We note from our outcome analysis in Section 5 of the main paper that off-service placement is associated with a longer d_{los} for the focused group patients.

To account for the off-service slowdown, for the focused group patients, we fit a baseline d_{los} distribution for each specialty using data from patients who are admitted to their primary ward only. For each patient, we first generate a d_{los} from the baseline distribution. If she is admitted to the primary ward, we keep that d_{los}; if she is placed off-service, we adjust the d_{los} as follows

- We calculate $\tilde{d}_{los} = d_{los} \cdot \exp(\beta_{11})$.
- To ensure that the adjusted d_{los} remains an integer, we set it equal to $\lfloor \tilde{d}_{los} \rfloor + 1$ with probability $p = \tilde{d}_{los} - \lfloor \tilde{d}_{los} \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer that is less than or equal to x; and we set it equal to $\lfloor \tilde{d}_{los} \rfloor$ with probability $1 - p$.

The distribution of h_{dis} is fitted using data across all patients. This is because we see little heterogeneity among different specialties or admission sources in the discharge delay distribution.

3 Robustness check of the choice model

In this section, we test two alternative specifications of the choice model. In the first specification, we include an additional feature, Wkend, which is an indicator with $Wkend = 1$ if the patient is admitted during the weekend. For the second specification, we leverage the smaller dataset (5-month) which contains more detailed patient diagnostic information. In this specification, in addition to Wkend, we also include the van Walraven score (vanWal) which is a comorbidity score calculated based on the detailed diagnostic information. The estimation results are summarized in Table 1. We observe that the estimated preferences for different options are very close to those presented in Table 2 in the main paper, suggesting the robustness of our estimation. In particular, in addition to baseline preference, we also have similar magnitude of load-balancing (i.e. the parameters for Busy 1,2,3), the tendency to avoid delay (i.e. the parameters for Delay×Option), and time-of the day effect (the parameters for Evening×Option and Morning×Option).

Table 1: Robust check: determinants of ward assignment decisions

	1 year dataset			5 month dataset		
	Card	GenMed	Surg	Card	GenMed	Surg
Primary	2.165***	4.948***	1.650**	2.329***	5.464***	3.751***
	(0.405)	(0.586)	(0.504)	(0.670)	(1.142)	(1.019)
Wait	3.267***	5.157***	3.186***	3.281***	5.696***	4.964***
	(0.398)	(0.586)	(0.402)	(0.658)	(1.146)	(0.961)
Busy 1	-0.767***	-1.246***	-0.590***	-1.051***	-0.967***	-0.912***
	(0.126)	(0.109)	(0.136)	(0.186)	(0.165)	(0.193)
Busy 2	-0.571***	-0.637***	-0.064	-0.460**	-0.493***	0.134
	(0.090)	(0.066)	(0.111)	(0.141)	(0.102)	(0.176)
Busy 3	-0.450***	-0.308***	0.046	-0.549***	-0.316***	-0.267†
	(0.070)	(0.057)	(0.083)	(0.112)	(0.087)	(0.140)
Delay×Prim	-0.245**	-0.459***	-0.405***	-0.247*	-0.608***	-0.544***
	(0.095)	(0.067)	(0.077)	(0.123)	(0.111)	(0.155)
Delay×Wait	-0.843***	-0.880***	-0.957***	-0.885***	-1.016***	-1.140***
	(0.096)	(0.069)	(0.068)	(0.127)	(0.115)	(0.152)
Evening×Prim	-0.398	-1.569***	-0.674**	-0.404	-1.504***	-0.784†
	(0.306)	(0.239)	(0.247)	(0.423)	(0.408)	(0.476)
Evening×Wait	-1.309***	-2.080***	-1.205***	-1.539***	-2.237***	-1.168**
	(0.303)	(0.240)	(0.202)	(0.420)	(0.412)	(0.442)
Morning×Prim	2.544***	1.274*	1.421***	1.948*	1.264†	2.483*
	(0.755)	(0.498)	(0.392)	(0.795)	(0.689)	(1.123)
Morning×Wait	3.413***	1.920***	2.316***	2.944***	1.770*	3.381**
	(0.754)	(0.499)	(0.354)	(0.793)	(0.692)	(1.105)
Wkend×Prim	1.074*	-0.192	-0.003	1.232*	-0.052	-0.391
	(0.420)	(0.243)	(0.269)	(0.566)	(0.449)	(0.624)
Wkend×Wait	0.235	-0.460†	-0.371	0.320	-0.221	-0.534
	(0.421)	(0.245)	(0.230)	(0.567)	(0.452)	(0.593)
vanWal×Prim				-0.049	0.101†	0.112
				(0.039)	(0.057)	(0.125)
vanWal×Wait				-0.014	0.102†	0.085
				(0.039)	(0.057)	(0.122)
No. of obs.	3369	4826	2623	1456	2082	1038
Log-Likelihood	-4768	-6364.1	-3704.2	-1918.1	-2623.3	-1369
Pseudo R^2	0.140	0.109	0.091	0.171	0.103	0.109

$^{+}: 0.05 < p \leq 0.1$, $*: 0.01 < p \leq 0.05$, $**: 0.001 < p \leq 0.01$, $***: p \leq 0.001$.
The standard errors are reported in parentheses.

4 Robustness check of the outcome analysis

In this section, we present some of robustness checks for the causal effects of ED delay and off-service placement on patient's medical LOS.

4.1 Specialty level effect

In this section, we compare the effects of ED delay and off-service placement on patient's LOS across different medical specialties. Table 2 compares the estimation results for the Med cluster (including four medicine specialties) and the Surg specialty. We combine all the medicine specialties to reduce the estimation error. We first observe that ED delay still doesn't have a significant effect on patient's LOS. On the other hand, off-service placement has quite different effects for patients of different specialties. Surgical patients suffer more from being placed off-service, i.e. their increase in LOS is the larger.

Table 2: Effects of ED Delay and off-service placement for different clusters of patients

Variables	Med	Surg
OffService (fitted)	0.206 *	0.262 **
	(0.097)	(0.098)
EDDelay (fitted)	−0.423	−0.577
	(0.330)	(0.487)
No. of obs	5621	1377

$^{+}: 0.05 < p \leq 0.1$, $*: 0.01 < p \leq 0.05$, $**: 0.001 < p \leq 0.01$, $***: p \leq 0.001$.
The robust standard errors are reported in parentheses.

4.2 Instrumental variable

In this section, we conduct two sets of sensitivity analysis. Firs, we test different threshold values for PriAdmOccuHigh. The results are summarized in Table 3. We notice that the occupancy level of the primary ward has a highly nonlinear effect on the off-service placement decision. When the primary ward occupancy is below a certain level, i.e. 0.8, the chances of placing a patient off-service is almost negligible, while when the primary ward occupancy grows beyond a certain level, i.e. 0.95, the off-service placement rate increases very fast as the occupancy level further increases.

Table 3: Different threshold values for the instrumental variable

Variables	Threshold=0.99	Threshold=0.95
OffService (fitted)	0.238 *	0.006
	(0.098)	(0.103)
EDDelay (fitted)	−0.385	0.417
	(0.370)	(0.416)
PriAdmOccuHigh for OffService	0.142 ***	0.137 ***
	(0.016)	(0.010)
No. of obs	8642	8642

$^{+}: 0.05 < p \leq 0.1$, $*: 0.01 < p \leq 0.05$, $**: 0.001 < p \leq 0.01$, $***: p \leq 0.001$.
The robust standard errors are reported in parentheses.

We also test different cutoff values for for ED Delay. The results are summarized in 4. We observe that the estimated off-service slowdown is quite robust to reasonable threshold values for ED Delay.

Table 4: Different threshold values for the ED Delay

Variables	Threshold = 3 hours	Threshold = 5 hours
OffService (fitted)	0.186 *	0.185 *
	(0.089)	(0.086)
EDDelay (fitted)	−0.150	−0.316
	(0.259)	(0.545)
No. of obs	8642	8642

$^{+}: 0.05 < p \leq 0.1$, $*: 0.01 < p \leq 0.05$, $**: 0.001 < p \leq 0.01$, $***: p \leq 0.001$.
The robust standard errors are report in parentheses.

4.3 Control for other covariates

In this section, we test a few other model specifications to estimate the effect of off-service placement on patient's LOS. In particular, we include different combinations of covariates related to the destination ward's occupancy and the attending physician's workload. These include the average occupancy of the destination ward during the patient's LOS (DestAvgOccu), the average workload the of attending physician during the patient's LOS (PhyAvgLoad), and the average workload of the attending physician the day before the patient's discharge (PhyMinus1Load). Estimation results are summarized in Table 5. We observe that the estimation results are quite robust across different choice of these covariates. Model III is the model we adopt in the main paper.

Table 5: 2SLS models with different covariates

Variables	Model I	Model II	Model III
OffService (Fitted)	0.192 **	0.183 **	0.172 *
	(0.072)	(0.069)	(0.069)
EDDelay (Fitted)	−0.003	−0.071	-0.157
	(0.254)	(0.273)	(0.271)
DestAvgOccu	—	0.323 **	0.321 **
		(0.116)	(0.115)
PhyAvgLoad	—	—	0.134 ***
			(0.012)
PhyMinus1Load	—	—	−0.128 ***
			(0.011)
No. of obs	8642	8642	8642

$^{+}: 0.05 < p \leq 0.1$, $*: 0.01 < p \leq 0.05$, $**: 0.001 < p \leq 0.01$, $***: p \leq 0.001$.
The robust standard errors are report in parentheses.

References

[1] Jayaprakash, Namita, Ronan O'Sullivan, Tareg Bey, Suleman S Ahmed, Shahram Lotfipour. 2009. Crowding and delivery of healthcare in emergency departments: the european perspective. *Western Journal of Emergency Medicine* **10**(4) 233.

[2] Siciliani, Luigi, Valerie Moran, Michael Borowitz. 2014. Measuring and comparing health care waiting times in oecd countries. *Health policy* **118**(3) 292–303.