What Causes Delays in Admission to Rehabilitation Care? A Structural Estimation Approach

Jing Dong
Decision, Risk, and Operations, Columbia Business School, jing.dong@gsb.columbia.edu
Berk Görgülü, Vahid Sarhangian
Department of Mechanical and Industrial Engineering, University of Toronto, {bgorgulu, sarhangian}@mie.utoronto.ca

Problem definition: Delays in admission to rehabilitation care can adversely impact patient outcomes and are costly for the healthcare system as delayed patients keep occupying their acute care beds, making them unavailable for incoming patients. Existing evidence suggests that admission delays are mainly caused by two sources: lack of rehabilitation bed capacity and the time required to plan for rehabilitation activities, which we refer to as *processing times*. However, due to the complex care transition process, non-standard bed allocation decisions, and data limitations in practice, quantifying the magnitude of the two sources of delays can be technically challenging yet critical to the design of evidence-based interventions to reduce delays. In this paper, we propose an empirical approach to understanding the contributions of the two sources of delays when only a single (combined) measure of admission delay is available. **Methodology/Results**: We propose a Hidden Markov Model (HMM) to estimate the unobserved processing requirements and the status-quo bed allocation policy, where the utility of allocating a bed to a patient depends on the patient's characteristics, the system's state, and various other factors. We employ a simulation-based approach with importance sampling to estimate the parameters of the structural model. Our estimation results quantify the magnitude of processing requirements versus capacity-driven delays and provide insights into factors impacting the bed allocation decision. We validate our estimated policy using a queueing model of patient flow, and find that ignoring processing delays or using simple bed allocation policies such as First-Come First-Served or strict priority can lead to highly inaccurate estimates. In contrast, our estimated policy matches the empirical delay distributions well and allows for accurate evaluation of different operational interventions. Through counterfactual experiments, we examine interventions targeted at addressing different sources of delays. We find that reducing processing times can be highly effective in reducing admission delays and bed-blocking costs. In addition, allowing early transfer – whereby patients can complete some of their processing requirements in the rehabilitation unit – can significantly reduce admission delays, with only a small increase in rehab LOS. **Managerial implications**: Our study demonstrates the importance of quantifying different sources of delays in design of effective operational interventions for reducing delays in admission to rehabilitation care. The proposed estimation framework can be applied in other transition-of-care settings with personalized capacity allocation decisions and hidden processing delays.

Key words: Structural estimation, hidden Markov model, capacity allocation, rehabilitation care

1. Introduction

Rehabilitation care (rehab for short) is an essential stage of treatment to improve the physical ability of patients after their acute care is completed. The continuing growth of the older population

has led to a 63% increase in demand for rehab globally over the past two decades (Cieza et al. 2020). Due to limited capacity, rehab facilities often operate near or at capacity, leading to long waiting times for admission (Cieza et al. 2020). These delays have been found to be associated with worse rehab outcomes, for instance, for stroke (Wang et al. 2011) and severe trauma (Spettell et al. 1991) patients. In addition, while waiting to be admitted into rehab, patients continue to occupy the acute care beds, leading to increased admission delays for incoming acute care patients.

Besides inadequate bed capacity, other factors may also contribute to delays in admission to rehab. When the acute care physician decides that the patient is stable and can be discharged to rehab, a rehab admission request is submitted to either an on- or off-site rehab facility. Before the patient can be physically transferred, rehab activities need to be planned in coordination with the acute care team. This process could take hours or days depending on the patient's condition and the efficiency of communication of different care teams. We refer to such delays as *processing delays* to distinguish them from capacity-driven delays (i.e., delays due to a lack of rehab bed capacity). In addition, we refer to a patient whose processing requirement is completed as *available*. Once a patient is available for rehab admission, she/he joins the waiting list until a bed becomes available. When there is not enough capacity, the bed allocation decision, i.e., which patients on the waiting list are selected for admission next, also impacts delays. As in many other healthcare settings, patients are not admitted on a First-Come First-Served (FCFS) basis. Instead, various medical and operational factors may affect the bed allocation decisions.

Our work was initiated as part of a collaboration with a large community hospital in the Greater Toronto Area (GTA) of Ontario, Canada. The hospital offers on-site Low-Tolerance, Long-Duration (LTLD) rehab care (see Section 2 for more details). In addition to a long average rehab admission delay of 7 days, there are significant disparities in the delays among different types of patients. Specifically, the average admission delays for the two largest acute categories of Medicine and Neurology / Musculoskeletal (Neuro/MSK), are 11.51 and 4.57 days, respectively. To prescribe effective operational interventions to reduce admission delays, the true determinants of delays and their respective effects should be identified and quantified. Processing delays can be reduced by standardizing the rehab planning process and improving the communication between care teams, whereas capacity-driven delays can be alleviated by adding extra rehab beds. To this end, we develop a structural model of rehab admission. The model provides insights into capacity and non-capacity related causes of admission delays.

In our study, we address two modeling and estimation challenges that are not only relevant to our setting, but can also arise in other multi-stage care settings. The first challenge is related to data limitations. Standard data collected by the Canadian Institute for Health Information (CIHI) includes a single measure of rehab admission delay, referred to as the Alternative Level of

Care (ALC) length of stay (LOS) (Canadian Institute for Health Information 2022), i.e., the time spent in acute care when the patient no longer needs that level of care. ALC LOS measures the time between when the acute care physician decides the patient is ready to be discharged from acute care and when the patient is admitted into rehab. As such, it is not possible to ascertain what portion of the delay is due to processing requirements and what portion is due to a lack of rehab capacity. We note that similar data limitations may arise in other care transition settings. For example, when transferring a patient from the Emergency Department (ED) to the inpatient wards, delays can be due to a lack of inpatient bed capacity or processing requirements, and detailed time stamps of various processing requirements may not be available (see, e.g., Chan et al. 2022). Second, the bed allocation decisions may not follow a standard policy and can depend on various patient- and system-level factors. This stands in contrast to the common assumptions in the Operations Management literature, where customers are classified into well-defined priority classes and assumed to be served FCFS within each priority class. Because the bed allocation decisions are made by clinical experts who take multiple factors into account, we focus on estimating the status-quo decisions assuming a utility-maximizing decision maker, as opposed to prescribing an optimal allocation policy based on an imposed objective function.

We propose a Hidden Markov Model (HMM) for the rehab admission process, where the hidden state indicates whether patients are available for admission. The estimation results (based on data from our partner hospital) provide insights into the processing times for different categories of patients and the status-quo bed allocation policy. We then combine the estimated rehab admission process with a queueing model of patient flow to validate the estimation results. Finally, we conduct counterfactual experiments to analyze various operational interventions to reduce admission delays.

1.1. Main Findings and Contributions

- **A structural model of rehab admission:** We propose a HMM for the transition from acute care to rehab. The model has two key components: a hidden component that helps separate processing delays from capacity-driven delays, and a multinomial logit model component that captures the bed allocation decision when there are multiple patients waiting to be admitted and not enough beds available.
- **Estimation results:** We find that processing delays and bed allocation decisions both contribute to admission delays and the disparities in delays between Medicine and Neuro/MSK patients. Processing times account for 59% of the observed delay, and take approximately 2.6 times longer on average for Medicine patients than for Neuro/MSK patients. Moreover, Neuro/MSK patients are on average 12.3% more likely to be allocated the next available bed compared to Medicine patients.

- **Model validation and counterfactual experiments:** Using a carefully calibrated queueing model of patient flow, we show that the distributions of delays under our estimated policy closely match the empirical distributions for each patient category. Failure to account for processing times or using simplistic bed allocation policies (e.g., FCFS or Strict Priority (SP)) can however lead to inaccurate waiting time distributions. We also conduct counterfactual experiments to evaluate the effectiveness of different operational interventions to reduce admission delays. Our key findings are summarized next.

(1) Reducing processing times (e.g., through improving the efficiency of the rehab planning process) can lead to a substantial reduction in admission delays and bed blocking costs. For our partner hospital, a 1% reduction in the average processing time leads to a saving of 11.74 acute patient days, which amounts to C$16,576 savings per year. (2) Through a combination of rehab capacity expansion and processing time reduction, hospitals can design practically feasible solutions to achieve various admission delay reduction targets. For example, a 2.5-day reduction in the average delay is possible at our partner hospital by adding two beds and reducing processing times by 25%. For the same level of reduction, if we use only one intervention, it requires adding 6 beds or reducing the processing time by 70%, which might be practically infeasible. (3) Allowing early transfer – whereby patients can complete their processing requirements after being transferred to the rehab – can reduce bed blocking costs significantly, with only a small increase in rehab LOS. At our partner hospital, an early transfer policy can reduce admission delays by 1.56 days on average with only a 4% increase in rehab utilization, leading to a saving of C$103,150 per year in bed blocking costs.

1.2. Related Literature

Our work mainly relates to four streams of research in the literature.

Capacity allocation in queueing systems: Scheduling policies have been extensively studied in the queueing literature when facing multiple classes of customers, see e.g., Cox and Smith (1991), Van Mieghem (1995) and Mandelbaum and Stolyar (2004). The objective is typically to minimize the holding cost or to satisfy certain service level constraints (Gurvich and Whitt 2010, Soh and Gurvich 2017). In the case of minimizing holding costs, policies that balance the holding cost and the service-time requirement such as the $c\mu$ rule tend to perform well (Mandelbaum and Stolyar 2004). In the healthcare setting, it is common to deviate from these simple policies since they may not capture important practical considerations, see, e.g., Carew et al. (2021) for an example in allocating surgical capacity. In general, there may not be a clear definition of customer classes and patients may belong to infinitely many classes based on various patient characteristics. Master et al. (2017, 2018) study the performance of queueing systems with continuous-priority classes.

Argon and Ziya (2009) and Singh et al. (2022) study priority assignment when perfect information on customers' types is not available. In our setting, when deciding how to allocate the rehab beds to different patients, the objective function of the decision maker can be very complex and have to account for various clinical and non-clinical factors. Thus, we focus on understanding the status-quo bed allocation policy by assuming a utility maximizing decision maker.

Empirical studies of capacity allocation decisions in healthcare and service systems: Several studies have empirically investigated capacity allocation decisions in service and healthcare settings. Tan and Staats (2020) estimate the effect of deviating from a standard table assignment policy (round-robin rule) in restaurants on productivity. Ibanez et al. (2018) analyze the effect of discretionary task ordering on productivity in radiology services, and show that doctors tend to prioritize shorter tasks. Kc et al. (2020) study how physicians select tasks in the ED under different load conditions, and show that physicians tend to prioritize easier tasks as load increases. These studies utilize reduced-form estimation and focus on measuring deviation from a benchmark policy. There are also studies that use the multinomial logit model by assuming utility maximizing decision makers. Ding et al. (2019) study how ED physicians prioritize patients based on different triage levels and their experienced delays up to the decision epoch. They find that the decision makers do not use the case complexity as a major criterion in the prioritization decision. Hathaway et al. (2022) study personalized prioritization policies that utilize past customer interaction information related to abandonment and redialing behaviors. They combine a multinomial logit model with a Bayesian learning framework to dynamically update customers' types. Li et al. (2021) study the prioritization decision between discharged versus admitted patients in the ED. They show that among the high acuity patients, those who are more likely to be admitted are prioritized over those who are more likely to be discharged when the ED is not crowded, and vice-versa when the ED is crowded with a large number of boarding patients. Jiang et al. (2021) examine the long-term capacity allocation decisions for different types of nursing homes: for-profit versus non-profit. Their study also combines the utility maximization behavior with a queueing network model. Similar to Ibanez et al. (2018), Kc et al. (2020) and Li et al. (2021), we find that patients who require less resource-intensive care or those who have waited for longer tend to be prioritized for rehab admission. Meanwhile, our work differs from the existing empirical studies in four main aspects. First, we focus on a different setting: rehab admissions. This leads to different policy implications as illustrated in our counterfactual study. Second, the previous structural models do not have a hidden component. The hidden component is crucial in accurately capturing the dynamics of our system, but poses substantial estimation challenges. Third, in addition to capacity-driven delays, we also study a different mechanism of delay: processing delays. Lastly, leveraging a well-calibrated

queueing model, we conduct counterfactual experiments to quantify the effectiveness of different delay reduction strategies.

Queueing models of patient flow: Our work relates to the growing body of literature on queueing models of patient flow (Armony et al. 2015, Dai and Shi 2021), especially the multi-stage healthcare settings. Bretthauer et al. (2011) study the blocking effect when patients transition between different hospital units and propose a heuristic to evaluate the blocking effect. Armony et al. (2018) focus on the patient flow from the intensive care unit to the step-down unit and study the capacity allocation decision between the two units. Zychlinski et al. (2020) propose a tandem queue with blocking to model patient flow from hospital wards to geriatric institutions, and study the optimal capacity level for the geriatric institutions. These works focus on long-term capacity planning decisions and assume beds are allocated FCFS. In this work, we focus on real-time controls. Our goal is to understand how rehab beds are dynamically allocated to different patients based on patients' characteristics and system status.

Hidden Markov models: We build a HMM for the rehab admission process. In the healthcare setting, HMMs have been used extensively to model disease progression; see, for example, Sukkar et al. (2012), Gupta et al. (2020), Kwon et al. (2020), Severson et al. (2020). Lim et al. (2021) propose a HMM to evaluate the Medicare yardstick incentive program on improving the quality of care. We study a different context (and thus model) from these previous studies, and our estimation results generate new insights on different sources of rehab admission delays.

2. Background and Data

In this section, we provide a summary of the process of admitting patients from acute care to rehab and an overview of the data used in our study.

2.1. Patient Flow from Acute Care to Rehab Care

We study two interconnected healthcare delivery entities: acute care and rehab, with a focus on the latter. Acute care units provide short-term treatment for a severe injury, an episode of illness, an urgent medical condition, or during recovery from surgery. Rehab units provide rehab services to improve patients' physical abilities, mostly after acute care. Our partner hospital, which operates in Ontario's publicly funded healthcare system, provides on-site rehab care, that is patients requiring rehab can be transferred to rehab units in the same hospital system. There are two types of rehab services: Low-Tolerance, Long-Duration care (LTLD) and High-Tolerance, Short-Duration care (HTSD). They provide distinct care targeted for different patient capabilities and needs. In this work, we focus on LTLD rehab due to its relatively simple structure of operation, i.e., it is offered in a single unit with 47 beds and has a high level of congestion (the average admission delay is 7.0 days). LTLD is suitable for patients who are capable of participating in less frequent rehab

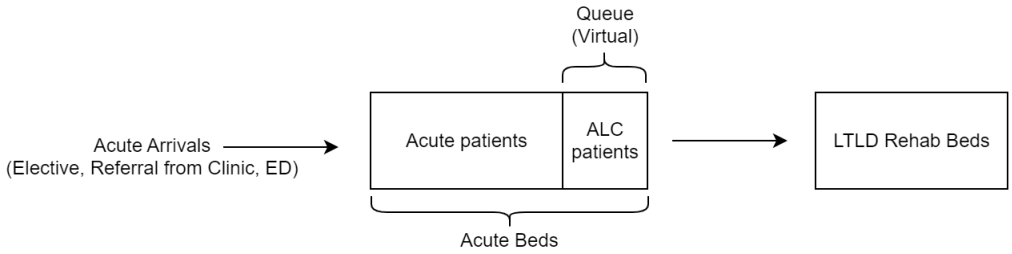


Figure 1 **A schematic representation of patient flow from acute to rehab care.**

activities with a relatively long duration, and these patients usually have other medical conditions (e.g., diabetes) besides the reason for which they are hospitalized. The average LOS is 67.6 days. Consistent with funding rates in Ontario, operating an acute bed costs approximately twice as much as a rehab bed: C$1,412 per day for an acute bed versus C$774 per day for a rehab bed.

Figure 1 provides a schematic representation of the patient flow from acute care to LTLD rehab. Patients arrive at acute care from different sources: elective surgeries, referrals from outpatient clinics, or the ED. They are categorized into different acute categories based on their care needs and are admitted to the corresponding acute wards. Upon completion of acute care, some patients require subsequent rehab care. LTLD rehab patients come from five acute categories (the values reported in brackets are percentages of LTLD rehab admissions from each acute category): Cardiology (1.0%), Cancer (0.8%), Family Practice (2.6%), Medicine (32.1%), Neurological/Musculoskeletal (Neuro/MSK) (60.2%), and Surgery (3.3%). Patients in need of rehab service are referred by their attending acute care physicians. Rehab admission requests are then processed by rehab coordinators. Coordinators evaluate the rehab request and plan the rehab activities in coordination with the rehab providers (physician, occupational therapist, etc.) and the acute care physicians. Patients can not be moved to the rehab unit before their rehab planning is completed. The time required for the planning varies for different patients and could take multiple days. While the exact processing time is not recorded, our collaborators believed that Medicine patients, who have more diverse care needs, tend to require longer processing times, whereas Neuro/MSK patients have more standardized rehab needs and thus shorter processing times. After completing the processing requirements, if there are no beds available, there could be further delays. Patients awaiting rehab admission (either due to a lack of bed capacity or ongoing processing) keep occupying the acute beds and maintain ALC status. We use the term *queue* to refer to the set of *patients with ALC status* in acute care. When a rehab bed becomes available, the rehab coordinators decide how to assign the bed to patients who are available for being transferred. Finally, rehab beds are identical and accommodate all LTLD patients.

2.2. Data

Our data set contains records for patients who were admitted to LTLD rehab after April 2014 and subsequently discharged before August 2019. Patients are admitted from two sources: transfer from acute care (1067 patients) and direct admit from outside the hospital (328 patients). For each patient record, we have the patient's demographic information, e.g., age, sex, comorbidity information, characteristics related to rehab care and acute care (if transferred from acute care), and timestamps for the patient's care trajectory.

In our data, comorbidity is measured with five levels: no comorbidity, levels 1, 2, 3 and 4. For patients who are admitted from the acute care of the same hospital, acute care information includes acute admission source, acute inpatient ward, acute category and subcategory, and Resource Intensity Weight (RIW). RIW is introduced by the Canadian Institute for Health Information (CIR) for planning, monitoring, and estimating acute care cost (Canadian Institute for Health Information 2020). It is a combined score based on comorbidity, LOS, and clinical interventions required. A higher RIW corresponds to a more resource-intensive care episode, which is typically associated with more severe patient conditions and longer acute LOS. Rehab information includes reasons for rehab. The timestamps include acute admission time, acute ALC LOS, rehab admission time, and rehab discharge time. The ALC LOS can be viewed as the time from rehab admission request to acute discharge. It contains both the rehab planning time, which we refer to as the processing time, and the waiting time due to the unavailability of rehab beds. A detailed description of the data fields and their sources can be found in Section EC.1 of the E-Companion.

We use a subset of data from April 2017 to April 2019 when there is no change in bed capacity and the occupancy can be accurately calculated (i.e., there is no censoring of patients due to the time cut-off since we have data well before April 2017 and after April 2019). We use the timestamps of all records to calculate the occupancy level of the LTLD unit (number of patients in service) and the number of patients on ALC status in acute care (queue length) each day during this period. Figure 2 plots these quantities for the studied period. We observe that the rehab unit is operating near capacity (47 beds) and there is typically a non-empty queue.

For our main estimation task, namely the structural estimation of the processing times and the bed allocation policy, we remove some records and make some assumptions. First, we eliminate direct admissions from outside the hospital (17% of the records) and focus on internal transfers only. This is because we do not have the acute care and waiting time information of these patients. Based on discussions with our medical collaborators, we assume internal patients are prioritized over external admissions (see also the discussion in Section 7). Second, we eliminate records with ALC LOS longer than 38 days (97.5% percentile). The exceedingly long ALC LOS of these patients is generally due to special medical circumstances (e.g., mental health, morbid obesity). Therefore,

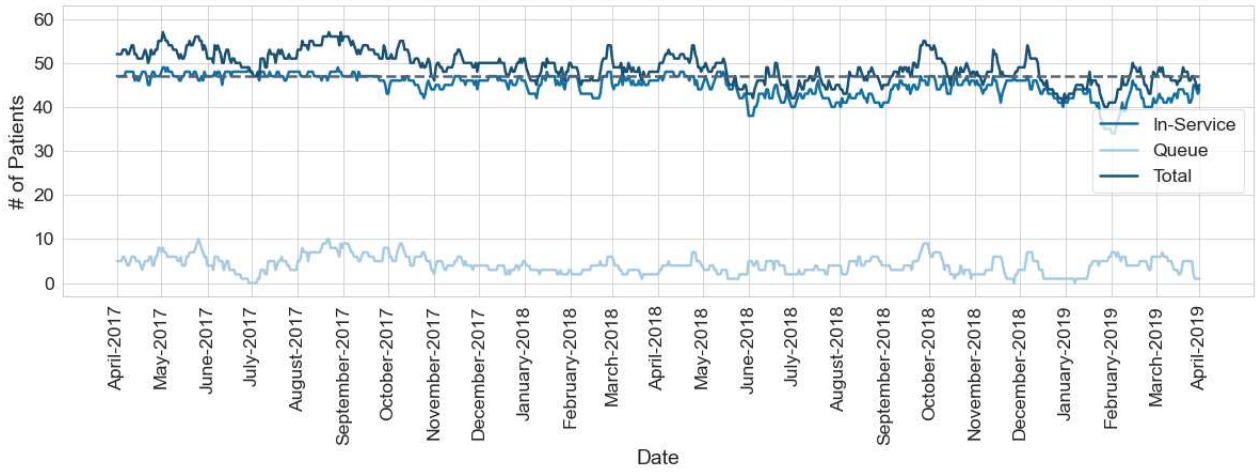


Figure 2 **Historical trajectory of the LTLD rehab. In-Service: Number of patients in the LTLD unit, Queue: Number of patients with ALC status in acute care, Total: Total number of patients in queue or in service.**

Table 1 **Summary statistics of Acute LOS, ALC LOS, and Rehab LOS (days) after outlier elimination.**

		Acute Category	
		Medicine	**Neuro/MSK**
Acute LOS	**Mean**	30.69	22.88
	Std	21.81	16.53
	Median	26.00	19.00
ALC LOS	**Mean**	11.51	4.57
	Std	10.43	7.28
	Median	10.00	0.00
Rehab LOS	**Mean**	69.81	66.90
	Std	38.62	35.39
	Median	63.00	58.50

highly specialized interventions are required to reduce their admission delays, which is beyond the scope of this study. Third, we only include admissions on weekdays in our estimation. This is because rehab admissions rarely happen over the weekends (3.5% of all admissions) and the hospital may operate with reduced resources during weekends. Lastly, we group patients in categories other than Medicine and Neuro/MSK patients into a single category called "Other" (comprising 5% of rehab admissions from acute care). See Appendix A for more details on data processing.

Table 1 provides some summary statistics of LOS-related performance metrics for Medicine and Neuro/MSK patients who require LTLD rehab (see Appendix A for box plots). Comparing the two acute categories, Medicine patients have a longer acute LOS, ALC LOS, and rehab LOS on average. Most noticeably, Medicine patients spend on average 7 more days on ALC status than Neuro/MSK patients.

3. The Model and Estimation Strategy

To understand the rehab admission process, we build a structural model to estimate the processing time distribution and the bed allocation policy for rehab care. Since the processing times and capacity-driven delays are not fully observable (only the sum is observed), we model the system as a HMM, i.e., a Markov chain with unobservable states. Estimating the model parameters of a HMM when the state space is large can be computationally challenging. We use an importance sampling-based estimation strategy to overcome the challenge.

3.1. The Hidden Markov Model

We consider the HMM $Y(t) = (N(t), F(t), C(t))$, $t \in \{0, 1, \ldots, T\}$. Each time period is a day. $Y(t)$ has three components: $N(t) \in \mathbb{N}_0$ denotes the number of patients in the queue (on ALC status); $C(t) \in \mathbb{N}_0$ denotes the number of available beds in the rehab unit; and $F(t) \in \mathbb{R}^{(f+1)\times\infty}$ contains patient-specific information for patients in the queue. More specifically, $F(t)$ consists of $N(t)$ non-zero column vectors, where the jth column, $F_j(t)$, contains the information of the jth patient in the queue (the order of the patients in the queue can be arbitrary). The first element of $F_j(t)$ for $j \in \{1, \ldots, N(t)\}$, denoted by $F_j(t)_1$, is a binary variable indicating the availability of the patient for rehab admission. It takes the value 1 if the patient is available, i.e., her/his rehab planning is completed, and 0 otherwise. Note that this element is not directly observable. The remaining elements of $F_j(t)$, denoted by $F_j(t)_{2:(f+1)}$, are observable and include various patient characteristics such as sex, age, acute category, and ALC LOS up to time t.

At each period, $Y(t)$ evolves as a result of four possible events that we assume occur in the following order: (1) arrivals of new patients to the queue, i.e., acute patients request LTLD rehab; (2) completion of processing requirements, i.e., the patient becomes available for rehab admission; (3) discharge from the rehab unit; and (4) bed allocation and admission to the rehab unit. To simplify the exposition, we assume that at each period, arrivals to the queue are Poisson distributed with time-dependent rates and are independent across periods. We also assume rehab LOS's are independent and identically distributed (iid) Geometric random variables. These assumptions ensure that $Y(t)$ is a Markov Chain. In our estimation, we estimate the model parameters conditional on the realized sample path as in the data, i.e., we take the arrivals and departures as given from the data. Thus, these distributional assumptions are not utilized. $C(t)$ includes admissions from outside the hospital to accurately capture the occupancy level of the rehab unit. Let $A(t)$ denote the number of arrivals, and $\tilde{N}(t)$ denote the number of available patients in the queue, i.e., $\tilde{N}(t) = \sum_{j=1}^{N(t)} 1\{F_j(t)_1 = 1\}$. Then,

$$N(t+1) = N(t) + A(t) - \min(C(t), \tilde{N}(t)).$$

We assume that processing times for patients of acute category $k \in \mathcal{K}$ follow a zero-inflated Geometric distribution with parameter $\phi_k = (\phi_{k,1}, \phi_{k,2}) \in [0,1]^2$, where $\mathcal{K} = \{M(\text{Medicine}), N(\text{Neuro/MSK}), O(\text{Other})\}$ denote the set of acute categories. In particular, for a patient of acute category k, her/his processing time S_k has the following probability mass function:

$$\mathbb{P}(S_k = 0) = \phi_{k,1}, \text{ and } \mathbb{P}(S_k = s) = (1-\phi_{k,1})\phi_{k,2}(1-\phi_{k,2})^{s-1} \text{ for } s \geq 1.$$

In other words, a type k patient has zero processing time (processing requirements are complete within the same period) with probability $\phi_{k,1}$; during each period, a type k patient who is not available at the previous period becomes available with probability $\phi_{k,2}$; and once a patient becomes available, she/he remains available until admitted, i.e., for $t \in \{1, 2, \}$,

$$\mathbb{P}(F_j(t)_1 = 1 | F_j(t-1)_1 = 0) = \phi_{k,2}, \quad \mathbb{P}(F_j(t)_1 = 1 | F_j(t-1)_1 = 1) = 1.$$

For the bed allocation decisions in period t, only the patients who are available can be admitted into the rehab. If the number of available beds is greater than the number of available patients in the queue, i.e., $\tilde{N}(t) \leq C(t)$, all available patients are admitted to the rehab unit. If there is not enough capacity available, i.e., $\tilde{N}(t) > C(t)$, available patients are selected according to a utility maximization rule which leads to a multinomial logit model. More specifically, let $U_{j,t}$ denote the utility of selecting the jth patient in the queue. We assume $U_{j,t}$ takes the following form

$$U_{j,t} = \begin{cases} \beta \cdot F_j(t)_{2:(f+1)} + \epsilon_{j,t}, & \text{for } F_j(t)_1 = 1, \\ -\infty, & \text{for } F_j(t)_1 = 0, \end{cases} \tag{1}$$

where $\beta \in \mathbb{R}^f$ is a vector of coefficients for the observable patient characteristics and $\epsilon_{i,t}$ is the unobservable idiosyncratic determinants of the bed allocation decision. We assume that $\{\epsilon_{j,t}\}$ are iid type-I extreme value distributed. Then, conditional on the characteristics of patients on the waiting list at time t, i.e., $Y(t)$, the jth patient is selected for admission over the other patients with probability (McFadden et al. 1973),

$$p_{j,t} = \frac{\exp(U_{j,t})}{\sum_{i=1}^{N(t)} \exp(U_{i,t})}.$$

We conclude this section with a discussion of our main modeling assumptions and the identification strategy. We consider a discrete-time model with each period being a day. This is reasonable since care-transition decisions are typically made on a daily basis during the inspection round. We model the processing times as zero-inflated Geometric distributions, and assume the processing times do not depend on the congestion level of the system. We also assume the utility of selecting a patient for rehab admission takes the form (1). These assumptions are imposed for analytical

tractability. Similar utility functions are commonly used in the empirical literature (see, for example, Akşin et al. 2013, Ding et al. 2019, Dong et al. 2021). If multiple patients are admitted to the rehab on the same day (11.6% of the days), we do not take the order at which they are admitted into account. We only assume that their admission utilities are higher than the patients who are available but not selected. We assume the nominal capacity of the rehab is 46 beds rather than 47 beds. This is to approximately account for the fact that beds can sometimes be closed (unavailable) due to maintenance, cleaning, or other issues.

Finally, we assume that patients can only be admitted to rehab after their processing requirements are completed. This is consistent with the practice at our partner hospital. Furthermore, we assume a non-idling bed assignment policy, i.e., beds are not held idle if there are available patients waiting to be admitted. This allows us to differentiate processing delays from capacity-driven delays by utilizing the variations in $C(t)$ and $N(t)$, and the observed bed allocation decisions. For example, if $C(t) > 0$ and $N(t) > 0$ after the bed allocation (47.6% of the days), then all patients remaining in the queue have not finished their processing requirements yet. On the other hand, if a patient is selected for admission in period t, she/he must have completed the processing requirements at or before t. Based on the observable states and bed allocation decisions, we can estimate the HMM by maximizing the expected likelihood function. We detail the estimation procedure in the next subsection.

3.2. Importance Sampling-Based Maximum Likelihood Estimation

In the data, we observe partial information of the queue, i.e., $N(t)$ and $F_j(t)_{2:(f+1)}$ for $j \in \{1,\dots,N(t)\}$, the available rehab capacity $C(t)$, and which patients (if any) are selected for rehab admission. Denote the bed allocation decisions in period t by $O(t)$. We estimate β and ϕ by maximizing the expected likelihood of observing $O(t)$ conditional on the observable patient and system information. Let $\tilde{Y}(t)$ denote the observable part of $Y(t)$, i.e., $\tilde{Y}(t) = (N(t), \{F_j(t)_{2:(f+1)}, j = 1,\dots,N(t)\}, C(t))$. Then, we aim to solve

$$\arg\max_{\beta,\phi} \mathbb{E}\left[\prod_{t=0}^{T} \mathbb{P}(O(t)|Y(t)) \middle| \tilde{Y}(t)\right] := \arg\max_{\beta,\phi} \sum_{\omega\in\Omega} \mathbb{P}(\omega) \prod_{t=0}^{T} \mathbb{P}(O(t)|\tilde{Y}(t),\omega), \tag{2}$$

where the expectation is with respect to the unobservable states $\{F_j(t)_1; t \in \{0,\dots,T\}\}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $\omega \in \Omega$ denotes a sample path of the unobservable states, i.e., it records when each patient becomes available. Note that $\mathbb{P}(\omega)$ depends on the value of ϕ. Given $\tilde{Y}(t)$ and ω, $\mathbb{P}(O(t)|\tilde{Y}(t),\omega)$ only depends on the value of β.

If the size of Ω is small, the optimization problem (2) can be easily solved. However, in our model, Ω grows exponentially with the number of patients. In this case, solving (2) exactly is

computationally prohibitive. Thus, we employ a simulation-optimization based approach. For a given ϕ, we consider the following unbiased estimator of the expected likelihood,

$$\frac{1}{n}\sum_{i=1}^{n}\prod_{t=0}^{T}\mathbb{P}(O(t)|\tilde{Y}(t),\hat{\omega}_i),$$

where $\hat{\omega}_1,\hat{\omega}_2,\ldots,\hat{\omega}_n$ are n realizations of ω given ϕ. In particular, for each patient in the data set, we generate their rehab processing times according to the zero-inflated Geometric distributions with parameter ϕ.

One challenge in implementing the above estimation scheme is that most of the generated sample paths are not "feasible", i.e., $\mathbb{P}(O(t)|\tilde{Y}(t),\hat{\omega}_i)=0$. To improve the estimation efficiency, we use importance sampling (Owen 2013). Let $\tilde{\Omega}$ denote the set of feasible paths given the observed bed allocation decisions $O(t)$, i.e., $\tilde{\Omega}=\{\omega:\prod_{t=0}^{T}\mathbb{P}(O(t)|\tilde{Y}(t),\omega)>0\}$. We consider the following unbiased estimator of the expected likelihood given ϕ,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(\hat{\omega}_i|\hat{\omega}_i\in\Omega)}{\mathbb{P}(\hat{\omega}_i|\hat{\omega}_i\in\tilde{\Omega})}\prod_{t=0}^{T}\mathbb{P}(O(t)|\tilde{Y}(t),\hat{\omega}_i)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(\hat{\omega}_i\in\tilde{\Omega})\prod_{t=0}^{T}\mathbb{P}(O(t)|\tilde{Y}(t),\hat{\omega}_i). \tag{3}$$

Note that $\mathbb{P}(\hat{\omega}_i\in\tilde{\Omega})$ is a constant that depends on ϕ. In particular, let s_j^u denote the time that patient j is assigned to a rehab bed. Let s_j^l denote the last time that patient j is not assigned to an available bed despite that there are still available beds. Then, patient j must have become available at some time during $[s_j^u, s_j^l]$. Let s_j^ω denote the time at which patient j becomes available, i.e., the processing time of patient j is completed under sample path ω. Then,

$$\mathbb{P}(\hat{\omega}_i\in\tilde{\Omega})=\prod_{j\in\mathcal{M}}\mathbb{P}(s_j^l\le s_j^\omega\le s_j^u),$$

where $\mathcal{M}$ is the collection of all the patients in the data set.

For a given value of ϕ, we first find β that maximizes (3). Then, we use grid search to find ϕ that maximizes the expected likelihood. Note that for a fixed ϕ, (3) reduces to an affine combination of products of multinomial logit probabilities. Thus, it is a concave function of β (Pratt 1981), and the optimal β can be found using the standard gradient ascent method. When implementing the estimation scheme, the expected likelihood is estimated based on 1000 sample paths. We also calculate the standard errors using parametric Bootstrap (Fuh and Hu 2007). Details of the standard error calculation are provided in Appendix B.

4. Estimation Results

In this section, we present our estimation results. Recall that $F_j(t)_{2:(f+1)}$ contains various characteristics of patient j who is with ALC status at time t. We consider the following features: age, RIW, sex, acute category, whether the patients' ALC LOS up to time t exceeds 15 days, and

the interaction between RIW and the congestion of the system measured by whether the number of patients in the queue exceeds 3 or 6 (corresponding to the 25th and 75th percentiles of the queue-length respectively). We also assume $\phi_{O,1} = 1$, i.e., patients in the Other category (5% of the observation) have zero processing times. We refer to the model with the above features as our main model (Model 1).

To examine the robustness of the estimation results, we examine several alternative model specifications. First, we consider models whose processing times can also depend on the severity/complexity of the patients. In Model 2, we assume the processing time distributions depend on the comorbidity level: High (H) versus Low (L). High comorbidity represents a comorbidity level at or above 3. In Model 3, we assume the processing time distributions depend on both the acute category and the comorbidity level. Second, we examine different measures of ALC LOS and system congestion. In our main model, we measure the effect of ALC LOS through an indicator, indicating whether ALC LOS exceeds 15 days. In Model 4, we treat ALC LOS as a numerical variable. In Models 5–7, we try different threshold values for long ALC LOS. In addition, in our main model, we measure congestion by two indicators based on the queue length: whether the queue length is above 3 and 6, which correspond to the 25th and 75th percentile of the empirical queue length respectively. In Model 8, we treat the queue length as a numerical variable. The estimation results for Models 1–3 are summarized in Table 2, while the full summary including Models 4–8, can be found in Appendix C.

We make a number of observations from the main model (Model 1). First, the processing time distributions for Medicine and Neuro/MSK patients are quite different. Medicine patients tend to have longer processing times. More specifically, 35% of the Medicine patients require zero processing time, while this percentage is 70% for Neuro/MSK patients. (Note that zero processing time means the processing time is less than a day.) Among patients who have non-zero processing times, the average processing time is 11.1 days for Medicine patients, and 9.1 days for Neuro/MSK patients. Based on conversations with our medical collaborators, this observation can be explained by the fact that Neuro/MSK patients have a more streamlined rehab planning process compared to Medicine patients. Since Medicine patients tend to have heterogeneous and complex medical conditions, they require more coordination and planning.

Second, we note that the coefficient for Neuro/MSK in the bed allocation model is positive and significant. Since Medicine is the baseline category, this suggests that after processing, Neuro/MSK patients are prioritized for rehab admission over Medicine. Specifically, the average partial effect (see Wooldridge 2002, Section 15.9) obtained from the estimated coefficients indicates that when there is not enough capacity, Neuro/MSK patients are on average 12.3% more likely to be chosen for admission than Medicine patients. This can be attributed to the utility maximization behavior

Table 2 Estimated coefficients of the model. The upper panel shows the estimated coefficients for the bed allocation decision, i.e., the coefficient for different patient-level characteristics. The lower panel shows the estimated parameters of the processing time distribution, i.e., the probability of having zero processing time and the geometric success probability for the zero-inflated Geometric distribution. Standard errors are provided inside brackets. Stars indicate statistical significance at different levels (ˆ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$).

Covariates	Model 1	Model 2	Model 3
Age	-0.008	-0.025ˆ	-0.018
	(0.011)	(0.012)	(0.012)
RIW	-0.130**	-0.097**	-0.066*
	(0.043)	(0.036)	(0.035)
Sex: Male	-0.371ˆ	-0.385	0.023
	(0.214)	(0.258)	(0.267)
Acute Category: Neuro/MSK	1.143***	1.201***	1.308***
	(0.235)	(0.274)	(0.292)
Acute Category: Other	1.365*	1.968	0.696
	(0.547)	(1.035)	(0.675)
Wait (ALC LOS) > 15	0.500ˆ	1.018**	0.835*
	(0.286)	(0.381)	(0.356)
RIW × (Queue-Length>3)	0.081ˆ	0.019	-0.081*
	(0.049)	(0.044)	(0.041)
RIW × (Queue-Length>6)	0.032	0.052	0.122
	(0.037)	(0.091)	(0.087)
ϕ_M	0.35***, 0.09***	-	-
	(0.067), (0.022)		
ϕ_N	0.7***, 0.11***	-	-
	(0.040),(0.026)		
ϕ_L	-	0.61***, 0.11***	-
		(0.030), (0.019)	
ϕ_H	-	0.47***, 0.07***	-
		(0.031), (0.021)	
ϕ_{ML}	-	-	0.34***, 0.09***
			(0.058), (0.023)
ϕ_{MH}	-	-	0.20*** , 0.07*
			(0.058), (0.030)
ϕ_{NL}	-	-	0.66***, 0.14***
			(0.053), (0.028)
ϕ_{NH}	-	-	0.74***, 0.08***
			(0.058), (0.024)

of the rehab coordinators. Since Medicine patients tend to require a longer rehab LOS and have more complicated care needs, there could be a tendency to prioritize the "easier" Neuro/MSK patients to reduce the overall system congestion. The potential tendency to prioritize "easier" cases can also be inferred from the coefficient for RIW, which is negative and significant, suggesting that the less resource-intensive patients are prioritized. Prioritization based on the difficulty of the tasks has been observed in other healthcare settings (see, e.g., Ibanez et al. 2018 for radiological services, and Kc et al. 2020 for the ED). The bed allocation policy together with the difference

in the processing times explains why Medicine patients have a much longer ALC LOS on average than Neuro/MSK patients (see Table 1).

The estimations based on Models 2 and 3 are in general consistent with that of Model 1. In particular, for both models, we observe that even after controlling for comorbidity-dependent processing times, the coefficient for Neuro/MSK is still positive and significant, suggesting Neuro/MSK patients are prioritized over Medicine patients. In addition, the coefficient for RIW is still negative and significant. Compared to Model 1, we observe a slight decrease in the magnitude of the effect of RIW in Models 2 and 3. This can be attributed to the ability of the later models in capturing severity/complexity-dependent processing times. When taking comorbidity into account in the processing time estimation in Model 2, we observe that patients with a lower comorbidity level tend to have a shorter processing time: there is a larger proportion of patients who require zero processing time (e.g., 61% versus 47%) and the average processing time conditional on the processing time being non-zero is smaller (e.g., 9 versus 14.2 days). We also observe from Model 3 that even after taking comorbidity into account, Medicine patients continue to have longer processing times compared to Neuro/MSK patients. For example, for patients with a low level of comorbidity, 34% of Medicine patients require zero processing time, compared to 66% for Neuro/MSK patients, and among patients who require positive processing times, it takes on average 11.1 days for Medicine patients, compared to 7.1 days for Neuro/MSK. The estimation results for Models 4–8 are also consistent with those of Model 1, see Table 4 in Appendix C) for more details.

5. Patient Flow Model

In this section, we propose a queueing model of patient flow from acute care to rehab. We then use simulation to compare the performance of the queueing model under our estimated bed allocation policy (EP) to the empirical data. Our estimated policy has two key elements: the processing time and the feature-based bed allocation decision. To quantify the effect of these two elements, we also compare our policy to commonly used benchmark policies with or without processing times.

5.1. The Queueing Model and Calibration

We consider a multi-server queue with time-varying arrival rates. Arrivals correspond to rehab bed requests from acute care patients and servers correspond to rehab beds (see Figure 3 for an illustration). We use data from April 2017 to April 2019 for model calibration (e.g., estimating the arrival rates, service time distribution, and distributions of patient characteristics) and validation.

Before proceeding with the details of our model, we introduce some terminology:

- *Queue-length* is the number of patients on ALC status in acute care, i.e., in the queue;
- *Waiting list* is the collection of patients who are in the queue and have already completed their processing requirements, i.e., these are the patients who are available for rehab admission;

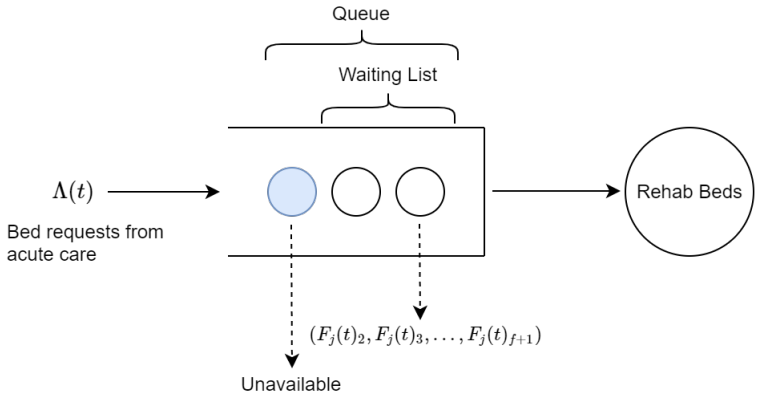


Figure 3 **An illustration of the proposed patient flow queueing model from acute care to rehab.**

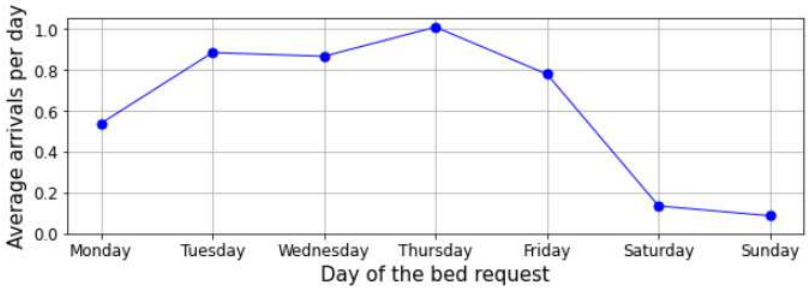


Figure 4 **Daily arrival rate of the bed requests for each day of the week.**

- *Queueing time* is ALC LOS, i.e., the time spent in the queue, including the processing times;
- *Waiting time* is the time spent on the waiting list due to a lack of bed capacity.

Arrivals. We assume a non-homogeneous Poisson arrival process with a piecewise-constant rate $\Lambda(t)$. In the data, we observe a strong day-of-the-week effect in arrivals. Meanwhile, the day-of-the-week pattern is fairly consistent across different weeks. As such, we assume that the arrival rates are periodic with a period equal to 7 days, i.e., $\Lambda(t+7) = \Lambda(t)$ for all $t \geq 0$. We then estimate the arrival rate for each day of the week using the corresponding sample average. We include both internal and external arrivals to match the load of the system as in the data. Figure 4 plots the estimated arrival rate for each day of the week. We observe that there is a large difference between weekdays and weekends, with almost no bed requests initiated during the weekends.

Feature vector and service times. Each arriving patient (bed request) is associated with a time-dependent $(f+1)$-dimensional random vector $X(t) := (F_j(t)_{2:(f+1)}, LOS)$. The first f elements of the vector correspond to patient characteristics, and the last element is her/his rehab LOS. We assume $X(t)$'s at patients' arrival times are iid draws from a joint distribution, and use the empirical distribution to estimate it. In particular, for each arrival, we draw a random sample from the patient records in the data.

Processing times. Processing times are generated from the zero-inflated Geometric distributions estimated based on Model 1. When a patient arrives at the queue (when a bed request is

Table 3 **Average queueing time (days) under EP, benchmark policies without (FCFS, SP) and with processing times (FCFSwP, SPwP). Standard errors are less than 0.5% of the estimates.**

	Average Queueing Time (Days)		
Policy	**All**	**Medicine**	**Neuro/MSK**
Empirical	7.00	11.51	4.57
EP	7.02	11.44	4.84
FCFS	2.88	2.88	2.88
SP	2.84	5.91	1.11
FCFSwP (with Processing)	7.03	10.05	5.61
SPwP (with Processing)	6.99	13.24	3.75

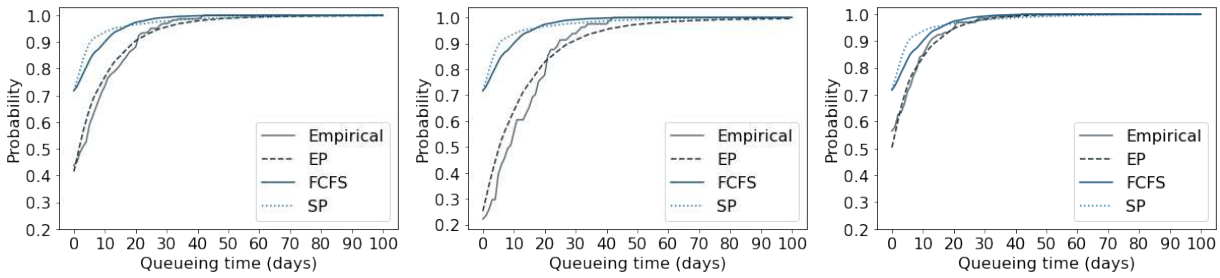
generated), she/he is not eligible for rehab admission until the processing requirements are completed. Recall that for patients in the Other category, we assume zero processing times.

Bed allocation policy. We assume the bed allocation policy is non-preemptive and non-idling. That is, when a patient enters service, she/he stays there until service completion, and we do not allow the server to idle if there are (available) patients waiting to be admitted. Only patients who have completed the processing requirements can be admitted to the rehab. When the number of patients on the waiting list exceeds the number of available beds, the bed allocation policy specifies which patients are to be admitted next. We consider three policies: EP, First-come-first-served (FCFS), and Strict Priority (SP) in favor of Neuro/MSK patients. For EP, the probability of assigning an available bed to a patient on the waiting list is determined by the estimated multinomial logit model based on patients' characteristics.

Data trimming. When calibrating the queueing model, we remove records with very long (97.5th percentile) or short (2.5th percentile) rehab LOSs. The details can be found in Appendix A. We assume rehab discharges cannot happen over the weekends. In particular, if a patient's admission time plus LOS is a weekend, we delay the discharge until Monday. This is because only 6% of the rehab discharges took place over the weekend in the data. As in our estimation, we assume there are 46 beds. Lastly, we scale the estimated arrival rate down by 3.5% to match the observed average waiting time.

5.2. Model Validation

We use simulation to compare the outputs of our queueing model under EP with the data. In addition, we compare the performance of our estimated policy with four benchmark policies: FCFS with processing time (FCFSwP), FCFS without processing time (FCFS), SP with processing time (SPwP), and SP without processing time (SP). In particular, FCFS and SP assume zero processing times, while FCFSwP and SPwP assume the processing times follow zero-inflated Geometric distributions as estimated in our main model. The goal is to quantify the effects of the processing times and bed allocation decisions on rehab admission delays.



(a) All patients (b) Medicine (c) Neuro/MSK

Figure 5 Queueing time distribution under EP, FCFS, and SP in comparison to the empirical queueing time distribution. Figures illustrate the overall, Medicine and Neuro/MSK queueing times.

Figure 5 compares the cumulative distribution functions (CDF's) of the simulated queueing time (ALC LOS) under different bed allocation policies to the empirical CDF's estimated directly from data and Table 3 summarizes the long-run average queueing times (we generate a large enough sample such that the standard errors of the estimates are less than 0.5% of the point estimates).

For the queueing time distributions and the average queueing times, the model under EP matches the empirical observations well, both in aggregate and for each patient category. However, this is not the case for the benchmarks. When we do not take the processing times into account (FCFS and SP), the proportion of patients with zero queueing time is severely overestimated (see Figure 5). Moreover, the average queueing time is 41% lower than the average observed in the data. This indicates that processing times are an important driver of admission delays. When taking the processing times into account, while the aggregated average queueing time under both FCFSwP and SPwP are fairly close to the empirical average, the category-specific averages do not match the data well. EP tends to prioritize Neuro/MSK patients. Thus, FCFSwP overestimates the queueing time for Neuro/MSK patients by 1.04 days and underestimates that for Medicine patients by 1.46 days. On the other hand, EP does not give strict priority to Neuro/MSK patients. Thus, SPwP underestimates the queueing time for Neuro/MSK patients by 0.82 days and overestimates that for Medicine by 1.73 days. In the E-Companion (Section EC.2), we further illustrate the importance of accurately modeling the bed allocation policy under different system congestion levels.

6. Counterfactual Experiments

In this section, we use the queueing model of patient flow under EP to evaluate various operational interventions to reduce admission delays. In Section 6.1, we focus on bed capacity planning. In Section 6.2, we examine the impact of reducing processing times. In Section 6.3, we study the combined effect of capacity expansion and processing time reduction. In Section 6.4, we examine an early-transfer strategy where patients are transferred to rehab as soon as a bed becomes available, which can happen before the processing requirements are completed. Throughout this section, we focus on the long-run average queueing/waiting times as the performance measure.

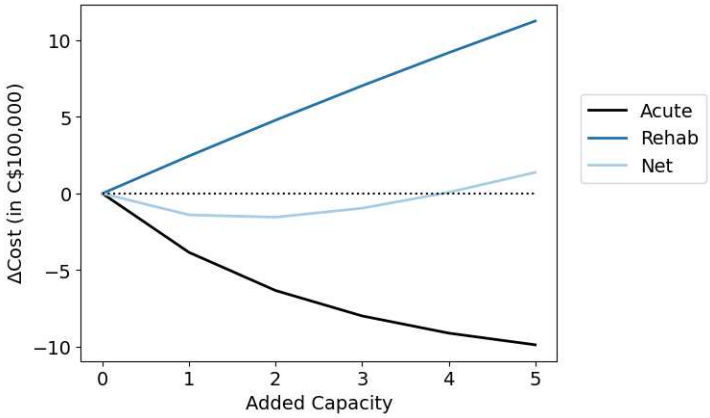


Figure 6 **Change in acute, rehab, and net costs with number of additional rehab beds.**

6.1. Capacity Expansion

Determining the bed capacity of the rehab unit is an important strategic decision. Our patient flow model can help optimize such decisions. We use the estimated model to evaluate the cost and benefit of increasing rehab bed capacity. Adding rehab beds incurs the corresponding operating costs, but also reduces rehab admission delays and hence saves acute patient days.

Figure 6 demonstrates how various costs change as the number of added rehab beds increases – the operating cost of the additional rehab beds, the saving due to saved acute patient days, and the net change in costs by combining the two – under the assumption that the patient demand for rehab does not change with increased bed capacity. For example, by increasing the number of rehab beds from 46 to 48, we can save 449 acute patient days per year. The cost of operating two additional rehab beds is C$1,548 per day which translates to C$478,881 per year if these beds are fully occupied throughout the year. On the other hand, the saving in acute patient days amounts to C$633,988 per year. Thus, the net cost saving is C$155,012 per year. In addition to the cost reduction, adding rehab beds also improve service quality by reducing the rehab admission delay and rehab occupancy. For example, adding two rehab beds reduces the admission delay by 1.57 days (2.41 days for Medicine, 1.12 days for Neuro/MSK) on average and reduces rehab occupancy by 8.7%, i.e., from 88.4% to 79.7%.

6.2. Reducing Processing Times

As discussed earlier, the admission delay (queueing time) contains two parts: the processing time and the waiting time. Adding capacity can only reduce the waiting time. Reducing the processing time requires efforts such as standardizing the rehab planning process and improving the coordination between acute and rehab providers. In this section, we examine the effect of reducing processing times on system performance and evaluate potential cost savings.

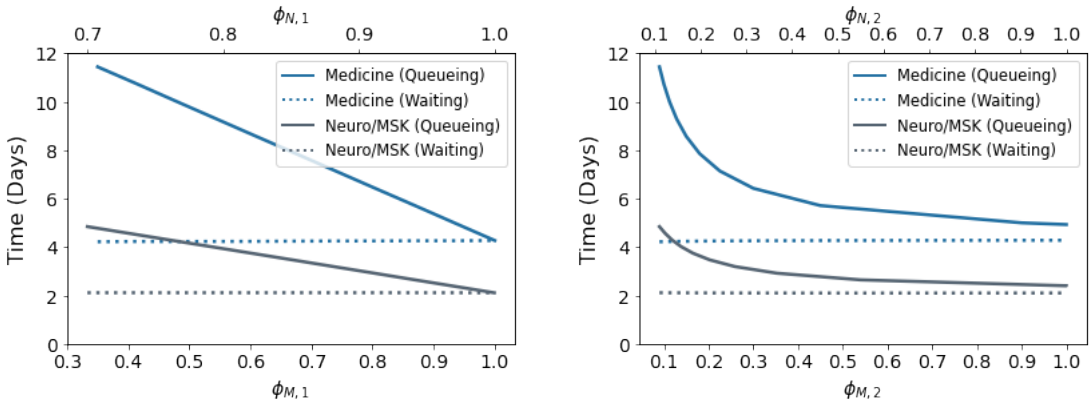


Figure 7 Average queueing and waiting times under EP for different values of ϕ.

We fix the bed allocation policy to the estimated one and vary the distributions of the processing times. Recall that we assume that processing times follow a zero-inflated Geometric distribution with class-dependent parameters, i.e., $\phi_k = (\phi_{k,1}, \phi_{k,2})$ where $\phi_{k,1}$ denotes the probability of requiring zero processing time (i.e., less than a day) and $\phi_{k,2}$ denotes the success probability. A larger value of ϕ_k leads to a shorter processing time on average.

Figure 7 illustrates the average queueing and waiting times for Medicine and Neuro/MSK patients under different values of ϕ_k, $k \in \{\text{M, N}\}$. We vary the values of $\phi_{M,1}$ and $\phi_{N,1}$ together in the left figure and vary the values of $\phi_{M,2}$ and $\phi_{N,2}$ together in the right figure. We observe that increasing $\phi_{k,1}$ leads to a linear decrease in the average queueing time. When $\phi_{k,1} = 1$ all rehab requests can be processed within a day, resulting in zero processing times. Meanwhile, the average queueing time is decreasing at a diminishing rate as $\phi_{k,2}$ increases due to its diminishing effect on the average processing time, i.e., $(1 - \phi_{k,1})(1/\phi_{k,2})$. On the other hand, the average waiting times do not change as $\phi_{k,1}$ or $\phi_{k,2}$ increases.

Similar to reducing capacity-driven delays, reducing processing times also helps reduce ALC LOS, which in turn frees up acute bed capacity. Our experiments reveal an approximately linear relationship between the average processing times and the admission delays. In particular, a 1% reduction in average processing time saves 11.83 acute patient days per year. This indicates that reducing the average processing time by 50% can save 592 acute patient days which amounts to C$835,904 cost savings per year.

We also investigate the effect of reducing the processing times of Medicine patients only, i.e., without changing the processing times of Neuro/MSK patients. In addition to reducing the average queueing time for Medicine patients, we also observe a smoothed patient arrival pattern, which helps reduce system idleness and thus queueing times for Neuro/MSK patients as well. However, the effect is quite small in our setting due to the long rehab LOS. Thus, we relegate its analysis and discussion to the E-Companion (Section EC.3).

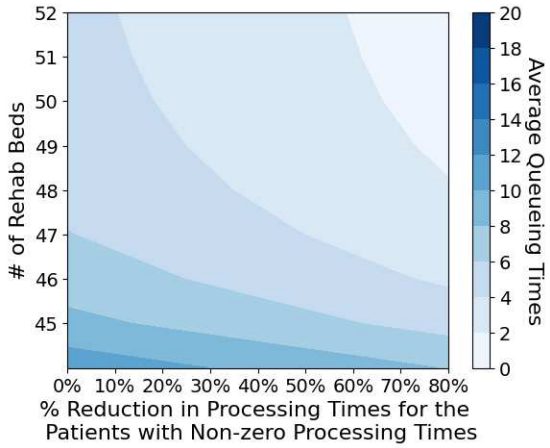


Figure 8 **Average queueing times as a function of the number of rehab beds and percentage reduction in processing times for patients with non-zero processing times.**

6.3. Combining Capacity Expansion with Processing Time Reduction

In this section, we examine the combined effect of capacity expansion and processing time reduction. We focus on the average queueing time as the performance metric. We reduce average processing time in our experiments by increasing $\phi_{k,2}$.

Figure 8 provides the heatmap of the average queueing time as a function of rehab capacity and the percentage reduction in the average processing time for the patients with non-zero processing times. (We also conduct the analysis for each patient category; see Section EC.3 of the E-Companion.) Consistent with our observations in Section 6.1, we observe that when the average queueing times are large, adding rehab beds initially leads to a large decrease in average queueing times but the effect diminishes as the number of beds increases. In contrast, reducing processing times has a linear effect on reducing queueing times. In addition, capacity expansion has approximately the same effect at different levels of processing time reduction. Similarly, reducing processing times has approximately the same effect at different capacity levels. This is because patients are not allowed to transfer to rehab prior to completion of their processing times, which prevents the two sources of delays from interacting with each other.

Figure 8 allows us to identify combinations of the two interventions to achieve a certain performance target. We observe that combining the two interventions can lead to more practically feasible solutions to reducing admission delays. For example, suppose our target is to reduce the average queueing time from 7 days to 4.5 days. We can do so by adding six beds or reducing the average processing time by 70%. If only one intervention is considered, this level of reduction may be infeasible in practice. On the other hand, this reduction can also be achieved by the combined intervention of adding two rehab beds and reducing the processing delay by 25%.

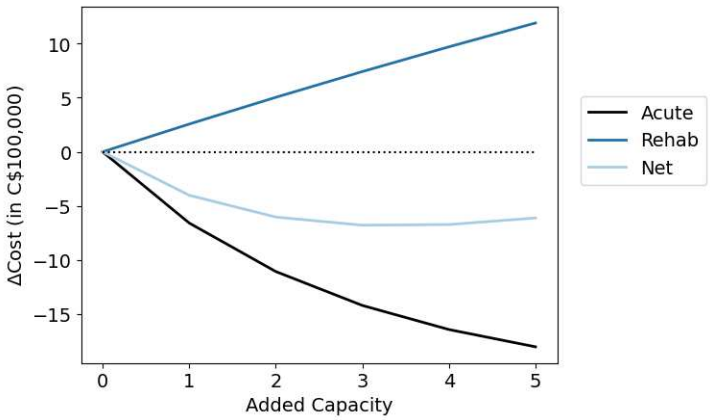


Figure 9 **Change in acute, rehab, and net costs as a function of the number of additional beds under early patient transfer.**

6.4. Early Patient Transfer

Currently, patients can only be transferred to rehab after their processing requirements are completed. In this section, we consider a counterfactual scenario where patients can be transferred to rehab before completing their processing requirements, if rehab beds are available. These patients complete their remaining processing times in rehab beds. As a result, their LOS in rehab is increased, but they spend less time blocking acute beds. We find that allowing early transfer leads to a 1.56-day reduction in the average queueing time, while only increasing the rehab bed utilization by 4%. This translates to a cost saving of $C\$103,150$ per year.

Figure 9 illustrates the effect of rehab capacity expansion with early transfers. Compared to those under the status-quo (Figure 6), the cost curves have a similar structure but with larger cost savings. For example, adding two more beds yields a net cost-saving of $C\$603,060$ per year compared to $C\$155,012$ per year in the base system.

Figure 10 provides a heatmap of average queueing time as a function of the number of rehab beds and the percentage reduction in average processing time for patients with non-zero processing times when early transfers are allowed (See Section EC.3 of the E-Companion for heatmaps of individual categories.) We observe a different structure compared to that under the status-quo (Figure 8). Because processing times no longer contribute to admission delays and only increase the LOS of patients in rehab, reducing them has a different effect at different capacity levels. Reducing processing times creates a larger reduction in queueing times when rehab has fewer beds. Similarly, adding rehab beds creates a larger reduction in queueing times when the average processing time is long. For example, increasing the number of beds from 46 to 48 leads to an additional 2.57-day and 2.00-day reduction when the average processing time is reduced by 10% and 50%, respectively, and reducing the average processing time by 30% leads to an additional 1.05-day and 0.31-day reduction when there are 46 and 50 rehab beds, respectively.

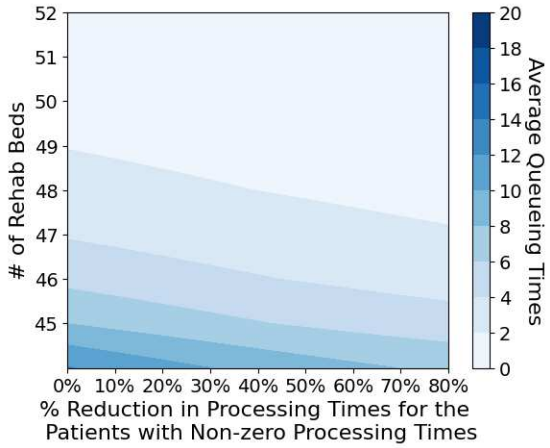


Figure 10 **Average queueing times as a function of the # of rehab beds and % reduction in processing times for the patients with non-zero processing times when early transfer to rehab is allowed.**

So far, we have assumed that rehab beds are allocated to patients regardless of whether their processing requirements are complete. We further examine an alternative policy where patients who have completed their processing requirements are prioritized for rehab admission. This policy yields an additional 0.72-day reduction in the average queueing time. See E-Companion (Section EC.5) for more details. Overall, our analysis suggests that early transfer can be highly beneficial in reducing rehab admission delays and acute care bed blocking. However, it should be noted that early transfer may also have drawbacks. For example, after being admitted to rehab, the acute care physician may be less responsive in rehab planning as they are occupied with newly admitted acute patients. This may further lengthen the required processing time. In this case, one needs to carefully evaluate the benefits against the potential drawbacks of early transfer.

7. Conclusion

Summary: We examine different sources of delays in transition from acute to rehab care. In addition to capacity-driven delays – delays caused by limited rehab bed capacity, we also identify and quantify delays driven by processing requirements - time required to coordinate and plan for the rehab activities. From an operations standpoint, reducing the two types of delays requires different interventions. Hence, it is important to distinguish the two. This is however challenging because (1) processing times are not directly observable in the data and (2) the bed allocation decisions are determined by various competing factors in practice. To address these challenges, we propose a HMM of the rehab admission process, which allows us to jointly estimate the processing time distributions and the status-quo bed allocation policy.

Our estimation results reveal that both the processing times and the bed allocation policy have considerable contributions to the long admission delays and the disparity in delays experienced

by different patient categories. In particular, Neuro/MSK patients tend to have shorter processing times and are also more likely to be prioritized for rehab admission compared to Medicine patients. We validate the estimated admission process using a detailed queueing model of patient flow and find that the output of the model matches the empirical delay distributions well. We then use the estimated model to evaluate various operational interventions to reduce admission delays.

We find that reducing processing delays, e.g., through improving the coordination between care teams and standardizing the rehab planning process, can lead to a significant reduction in admission delays and bed blocking costs. Further, through combining capacity expansion and processing time reductions, it is possible to construct practically feasible interventions to reduce admission delays.

These findings are enabled through the proposed HMM and estimation strategy. Our framework is more broadly applicable to other transition-of-care settings where only a combined measure of delay is typically available (e.g., from the ED to inpatient wards of a hospital, or from acute care to long-term care homes). Our findings further indicate that hospital information systems should collect granular time-stamps on start and completion times of processing times in such care-transition settings.

Limitations and future work: Our study has some limitations. First, we assume that the processing times are exogenous and do not depend on the state of the system. This assumption simplifies our model and identification strategy. Based on conversations with our medical collaborators, the decision makers (rehab coordinator or acute care physicians) do not respond strategically to the system state (e.g., by slowing down the rehab planning when beds are not available). That being said, speed-up or slow-down in rehab planning based on the load of the system is possible. Examining the impact of system load on processing times would be an interesting topic for future work.

Second, we focus on reducing admission delays under the status-quo bed allocation policy. A prescriptive approach that designs the optimal bed allocation policy under a practically relevant objective function would be an interesting future research direction. Defining an appropriate objective function can be challenging and requires carefully accounting for various considerations faced by practitioners. In addition, admission delays can also have heterogeneous effects on patient outcomes for different patients (Görgülü et al. 2023). It is important to take these heterogeneous effects into account when designing bed allocation policies.

Third, although we include various patient-level and system-level covariates, there can still be unobservable factors that affect the bed allocation decisions. For example, we may miss some important patient severity information in our data. Controlling for additional patient severity information may provide a more clear explanation of why Neuro/MSK patients are prioritized over Medicine patients. However, redistributing the weights of Neuro/MSK to other severity-related factors is unlikely to affect the overall waiting time estimates. In addition, given that our partner

hospital operates in Ontario's publicly funded healthcare system, we can rule out potential financial incentives. However, financial incentives might play an important role in other private healthcare systems. This may lead to a larger degree of prioritization of Neuro/MSK patients.

Finally, we focus on the bed allocation policy for patients transferred from acute care of the same hospital (internal admissions). We assume internal admissions are prioritized over admissions from outside the hospital (direct admissions). However, if external admissions have a higher priority, the hospital may hold available beds in anticipation of future external admission requests. In this case, our approach may lead to an overestimation of the processing times. Based on discussions with our medical collaborators, this is unlikely to be the case in our partner hospital. This assumption is further supported by additional regression analysis presented in the E-companion (Section EC.6).

Appendix A: Description of Variables and Data Processing

In this section, we provide a detailed description of the data used in our analysis and the related assumptions. The data was extracted from the Discharge Abstract Database (DAD) of the Canadian Institute for Health Information (CIHI) (Canadian Institute for Health Information 2020), and the hospitals' Electronic Health Records (EHR). Table EC.1 in E-Companion summarizes the variables with their descriptions, and sources.

The data are used for three main tasks: (i) calculating the rehab occupancy level (i.e., the number of available beds in the rehab unit); (ii) estimating the rehab admission process; and (iii) calibrating the queueing model for counterfactual experiments. Figure 11 summarizes the data used for each task.

To calculate the rehab occupancy level, we consider patients who stay in the rehab units any time during the period between April 2017 to April 2019. We focus on this period of time because the rehab bed capacity remained equal to 47 beds, and we can calculate the rehab occupancy level accurately during this period.

To estimate the rehab bed allocation process, we similarly consider patients who stay in the rehab units at any time during the period from April 2017 to April 2019. We eliminate direct admissions from outside the hospital (17% of the records) and patients with an ALC LOS longer than the 97.5-th percentile. We also exclude rehab admissions over the weekends. Note that we only exclude weekend admissions, but keep the patient records when deciding who is prioritized for admission during the weekdays. In particular, a patient who is admitted over the weekend but is on the waiting list before then is considered in the choice sets during weekdays prior to her/his admission.

To calibrate the queueing model, we estimate the arrival rates and use the empirical distribution for service times of different patient types. We again consider patients who stay in the rehab units any time during the period April 2017 to April 2019. We exclude patients whose rehab LOSs are longer than the 97.5th percentile or shorter than the 2.5th percentile.

Figure 12 presents three box plots summarizing Acute LOS, ALC LOS, and Rehab LOS for Medicine and Neuro/MSK patients.

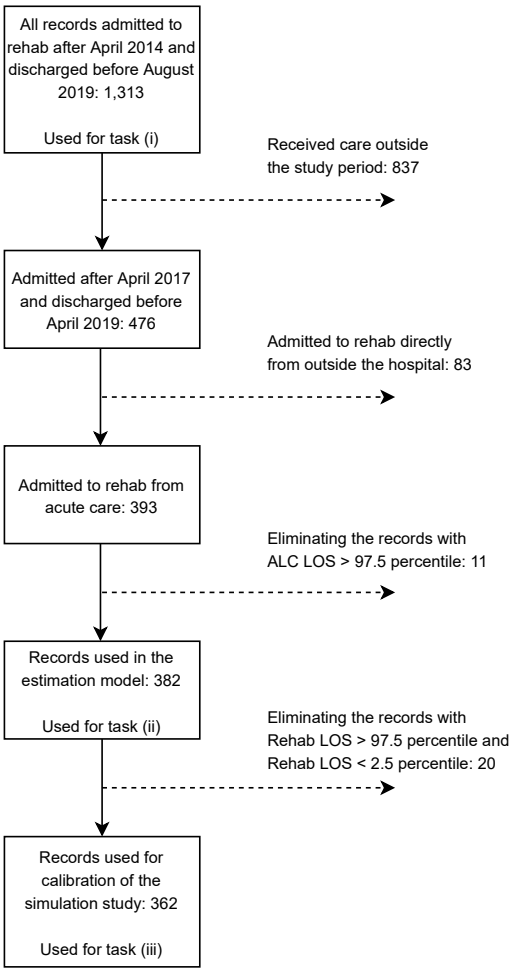


Figure 11 **Data selection for the estimation and simulation tasks.**

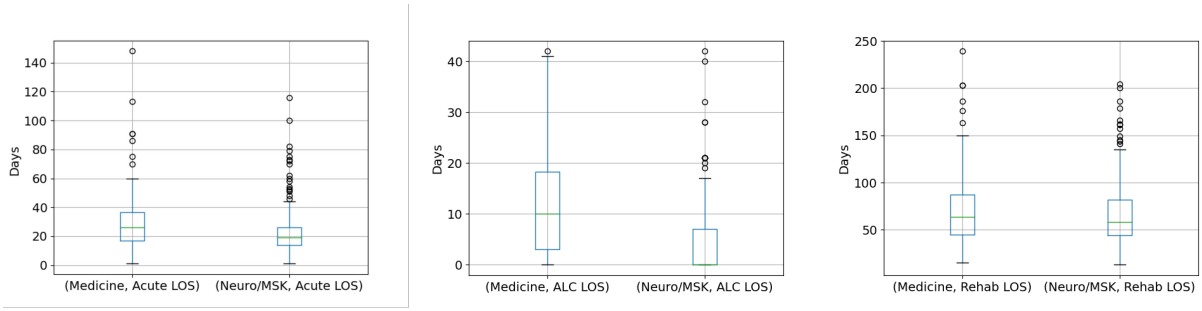


Figure 12 **Box plots summarizing Acute LOS, ALC LOS and Rehab LOS (days) for Medicine and Neuro/MSK patients.**

Appendix B: Standard Error Calculation

We use the parametric Bootstrap method to calculate the standard errors. Given the estimated coefficients $(\hat{\beta}, \hat{\phi})$, we re-sample patients' processing times and the bed allocation decisions. Then, based on the new sample of observed patient characteristics and bed allocation decisions, we re-estimate the coefficient (β, ϕ).

Different samples differ only by the processing times and the bed allocation decisions. Patients' arrival times, observable characteristics (e.g., sex, age, acute category, etc), and rehab LOS's are fixed as in the data.

More specifically, for each arriving patient, the processing time is generated from a zero-inflated Geometric distribution with category-dependent parameter $\hat{\phi}_k$. At each decision epoch, conditional on patients' availability and observable characteristics, bed allocation decisions are sampled according to the probability of the corresponding multinomial logit model with coefficient $\hat{\beta}$. When a patient is selected, that patient is removed from the queue and starts receiving rehab service. A re-generated sample contains all the patients who arrive after April 1, 2017 and are subsequently discharged before April 1, 2019. In order to obtain the same sample information as in the data, we omit the processing time information and treat patient availability as hidden. We re-estimate (β, ϕ) for each re-sampled path using the simulation optimization method with 100 samples. We repeat the re-sampling and re-estimation procedure 500 times. Let $(\hat{\beta}^i, \hat{\phi}^i)$, $i = 1, 2, \ldots, 500$ denote the re-estimated parameters. Then, the standard errors are calculated as

$$\hat{s}_{\beta_j} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\beta}_j^i - \hat{\beta}_j)^2}, \quad \hat{s}_{\phi_{k,j}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\phi}_{k,j}^i - \hat{\phi}_{k,j})^2}.$$

Appendix C: Robustness Checks

Table 4 provides the complete estimation results for different model specifications described in Section 4. Model 1 denotes our main model. Models 2 and 3 assume that processing time distribution depends on patient severity. Specifically, Model 2 assumes the processing time distribution depends on the comorbidity level, and Model 3 assumes the processing time distribution depends on both the acute category and comorbidity level.

Models 4-7 are concerned with the effect of ALC LOS (Wait). Model 4 treats Wait as a numerical variable, i.e., it assumes Wait has a linear effect on the selection utility. Model 5 uses the indicator: Wait>12. Model 6 uses the indicator: Wait>17. Model 7 includes three indicators for Wait: Wait>12, Wait>15, and Wait>17. We observe that the coefficient for Wait is not significant in Model 4. Compared to Model 1, this suggests that Wait is likely to have a nonlinear effect on patient selection utility. Further, Wait> 12 in Model 5 does not have a significant effect while Wait>17 in Model 6 has a significantly positive effect. This suggests that only very long waits have a significant effect. When Wait>12, Wait>15, and Wait>17 are all included in Model 7, Wait>15 stands out as the only significant one. This supports the choice of Wait>15 in our main model.

Model 8 includes Queue-Length $\times$ RIW by treating Queue-Length as a numerical variable. Comparing Model 8 to Model 1, we confirm that the queue length does not have a significant impact on the patient prioritization decision.

Table 4 Estimated coefficients of the model. The upper panel shows the estimated coefficients for the bed allocation decision, i.e., the coefficient for different patient-level characteristics. The lower panel shows the estimated parameters of the processing time distribution, i.e., the probability of having zero processing time and the geometric success probability for the zero-inflated geometric distribution. Standard errors are provided inside brackets. Stars indicate statistical significance at different levels (ˆ $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$).

Covariates	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Age	-0.008	-0.025ˆ	-0.018	-0.014	-0.011	-0.008	-0.023*	-0.018ˆ
	(0.011)	(0.012)	(0.012)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
RIW	-0.130**	-0.097**	-0.066*	-0.085*	-0.095*	-0.114**	-0.118**	-0.127**
	(0.043)	(0.036)	(0.035)	(0.042)	(0.045)	(0.043)	(0.044)	(0.048)
Sex: Male	-0.371ˆ	-0.385	0.023	-0.300	-0.300	-0.307	-0.464*	-0.399ˆ
	(0.214)	(0.258)	(0.267)	(0.205)	(0.212)	(0.216)	(0.222)	(0.064)
Acute Category: Neuro/MSK	1.143***	1.201***	1.308***	1.041***	1.107***	1.158***	0.938***	1.046***
	(0.235)	(0.274)	(0.292)	(0.232)	(0.233)	(0.246)	(0.238)	(0.237)
Acute Category: Other	1.365*	1.968	0.696	1.326*	1.372*	1.395**	1.593ˆ	1.306*
	(0.547)	(1.035)	(0.675)	(0.604)	(0.598)	(0.533)	(0.843)	(0.551)
Wait (ALC LOS)	-	-	-	0.001	-	-	-	-
				(0.011)				
Wai (ALC LOS) > 12	-	-	-	-	0.266	-	-0.471	-
					(0.249)		(0.503)	
Wait (ALC LOS) > 15	0.500ˆ	1.018**	0.835*	-	-	-	1.235ˆ	0.622*
	(0.286)	(0.381)	(0.356)				(0.617)	(0.281)
Wait (ALC LOS) > 17	-	-	-	-	-	0.722*	-0.137	-
						(0.313)	(0.456)	
RIW × Queue Length	-	-	-	-	-	-	-	0.012
								(0.008)
RIW × (Queue Length>3)	0.081ˆ	0.019	-0.081*	0.067	0.062	0.068	0.052	-
	(0.049)	(0.044)	(0.041)	(0.050)	(0.047)	(0.051)	(0.049)	
RIW × (Queue Length>6)	0.032	0.052	0.122	0.021	0.027	0.032	0.040	-
	(0.037)	(0.091)	(0.087)	(0.038)	(0.038)	(0.037)	(0.037)	
ϕ_M	0.35***, 0.09***	-	-	0.35***, 0.09***	0.35***, 0.09***	0.35***, 0.09***	0.36***, 0.08***	0.35***, 0.09***
	(0.067), (0.022)			(0.025), (0.015)	(0.055), (0.017)	(0.046), (0.019)	(0.037), (0.014)	(0.028), (0.015)
ϕ_N	0.7***, 0.11***	-	-	0.7***, 0.11***	0.7***, 0.11***	0.7***, 0.11***	0.7***, 0.11***	0.7***, 0.11***
	(0.040),(0.026)			(0.018),(0.030)	(0.032),(0.031)	(0.027),(0.030)	(0.020),(0.028)	(0.019),(0.030)
ϕ_L	-	0.61***, 0.11***	-	-	-	-	-	-
		(0.030), (0.019)						
ϕ_H	-	0.47***, 0.07***	-	-	-	-	-	-
		(0.031), (0.021)						
ϕ_{ML}	-	-	0.34***, 0.09***	-	-	-	-	-
			(0.058), (0.023)					
ϕ_{MH}	-	-	0.20*** , 0.07*	-	-	-	-	-
			(0.058), (0.030)					
ϕ_{NL}	-	-	0.66***, 0.14***	-	-	-	-	-
			(0.053), (0.028)					
ϕ_{NH}	-	-	0.74***, 0.08***	-	-	-	-	-
			(0.058), (0.024)					

References

Akşin Z, Ata B, Emadi SM, Su CL (2013) Structural estimation of callers' delay sensitivity in call centers. *Management Science* 59(12):2727–2746.

Argon NT, Ziya S (2009) Priority assignment under imperfect information on customer type identities. *Manufacturing & Service Operations Management* 11(4):674–693.

Armony M, Chan CW, Zhu B (2018) Critical care capacity management: Understanding the role of a step down unit. *Production and Operations Management* 27(5):859–883.

Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1):146–194.

Bretthauer KM, Heese HS, Pun H, Coe E (2011) Blocking in healthcare operations: A new heuristic and an application. *Production and Operations Management* 20(3):375–391.

Canadian Institute for Health Information (2020) CMG+. https://www.cihi.ca/en/cmg, Accessed: 2020-08-02.

Canadian Institute for Health Information (2022) Guidelines to support alc designation. https://www.cihi.ca/en/guidlines-to-support-alc-designation, accessed: 2022-18-02.

Carew S, Nagarajan M, Shechter S, Arneja J, Skarsgard E (2021) Dynamic capacity allocation for elective surgeries: Reducing urgency-weighted wait times. *Manufacturing & Service Operations Management* 23(2):407–424.

Chan CW, Sarhangian V, Talwai P, Gogia K (2022) Utilizing partial flexibility to improve emergency department flow: Theory and implementation. *Naval Research Logistics (NRL)* 69(8):1047–1062.

Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T (2020) Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019. *The Lancet* 396(10267):2006–2017.

Cox DR, Smith W (1991) *Queues*, volume 2 (CRC Press).

Dai JG, Shi P (2021) Recent modeling and analytical advances in hospital inpatient flow management. *Production and Operations Management* 30(6):1838–1862.

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

Dong J, Shi P, Zheng F, Jin X (2021) Structural estimation of load balancing behavior in inpatient ward network. Technical report, Working paper.

Eick SG, Massey WA, Whitt W (1993) $M(t)/G/\infty$ queues with sinusoidal arrival rates. *Management Science* 39(2):241–252.

Fuh CD, Hu I (2007) Estimation in hidden markov models via efficient importance sampling. *Bernoulli* 13(2):492–513.

Görgülü B, Dong J, Hunter K, Bettio KM, Vukusic B, Ranisau J, Spencer G, Tang T, Sarhangian V (2023) Association between delayed discharge from acute care and rehabilitation outcomes and length of stay: A retrospective cohort study. *Archives of Physical Medicine and Rehabilitation* 104(1):43–51.

Green L, Kolesar P, Svoronos A (1991) Some effects of nonstationarity on multiserver markovian queueing systems. *Operations Research* 39(3):502–511.

Gupta A, Liu T, Crick C (2020) Utilizing time series data embedded in electronic health records to develop continuous mortality risk prediction models using hidden markov models: A sepsis case study. *Statistical Methods in Medical Research* 29(11):3409–3423.

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

Hathaway BA, Emadi SM, Deshpande V (2022) Personalized priority policies in call centers using past customer interaction information. *Management Science* 68(4):2806–2823.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. *Management Science* 64(9):4389–4407.

Jiang Y, Lu LX, Van Mieghem JA (2021) Nonprofit vs. for-profit: Allocation of beds and access to care in us nursing homes. *For-Profit: Allocation of Beds and Access to Care in US Nursing Homes (July 7, 2021)* .

Kc DS, Staats BR, Kouchaki M, Gino F (2020) Task selection and workload: a focus on completing easy tasks hurts performance. *Management Science* 66(10):4397–4416.

Kwon BC, Achenbach P, Dunne JL, Hagopian W, Lundgren M, Ng K, Veijola R, Frohnert BI, Anand V, Group TS, et al. (2020) Modeling disease progression trajectories from longitudinal observational data. *AMIA Annual Symposium Proceedings*, volume 2020, 668 (American Medical Informatics Association).

Li W, Sun Z, Hong LJ (2021) Who is next: Patient prioritization under emergency department blocking. *Operations Research* 0(0).

Lim JM, Moon K, Savin S (2021) Searching for the best yardstick: Cost of quality improvements in the us hospital industry. *Available at SSRN 3885132* .

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

Master N, Reiman MI, Wang C, Wein LM (2018) A continuous-class queueing model with proportional hazards-based routing. *Available at SSRN 3390476* .

Master N, Zhou Z, Bambos N (2017) An infinite dimensional model for a many server priority queue. *2017 51st Annual Conference on Information Sciences and Systems (CISS)*, 1–6 (IEEE).

McFadden D, et al. (1973) Conditional logit analysis of qualitative choice behavior .

Owen AB (2013) Monte carlo theory, methods and examples. https://artowen.su.domains/mc.

Pratt JW (1981) Concavity of the log likelihood. *Journal of the American Statistical Association* 76(373):103–106.

Severson KA, Chahine LM, Smolensky L, Ng K, Hu J, Ghosh S (2020) Personalized input-output hidden markov models for disease progression modeling. *Machine Learning for Healthcare Conference*, 309–330 (PMLR).

Singh S, Gurvich I, Van Mieghem JA (2022) Feature-based priority queuing. *Available at SSRN 3731865* .

Soh SB, Gurvich I (2017) Call center staffing: Service-level constraints and index priorities. *Operations Research* 65(2):537–555.

Spettell CM, Ellis DW, Ross SE, Sandel ME, O'Malley KF, Stein SC, Spivack G, Hurley KE (1991) Time of rehabilitation admission and severity of trauma: effect on brain injury outcome. *Archives of Physical Medicine and Rehabilitation* 72(5):320–325.

Sukkar R, Katz E, Zhang Y, Raunig D, Wyman BT (2012) Disease progression modeling using hidden markov models. *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2845–2848 (IEEE).

Tan TF, Staats BR (2020) Behavioral drivers of routing decisions: Evidence from restaurant table assignment. *Production and Operations Management* 29(4):1050–1070.

Van Mieghem JA (1995) Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *The Annals of Applied Probability* 809–833.

Wang H, Camicia M, Terdiman J, Hung YY, Sandel ME (2011) Time to inpatient rehabilitation hospital admission and functional outcomes of stroke patients. *PM&R* 3(4):296–304.

Whitt W (2014) The steady-state distribution of the $M_t/M/\infty$ queue with a sinusoidal arrival rate function. *Operations Research Letters* 42(5):311–318.

Wooldridge JM (2002) Econometric analysis of cross section and panel data mit press. *Cambridge, MA* 108.

Zychlinski N, Mandelbaum A, Momčilović P, Cohen I (2020) Bed blocking in hospitals due to scarce capacity in geriatric institutions—cost minimization via fluid models. *Manufacturing & Service Operations Management* 22(2):396–411.

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

E-Companion

Appendix EC.1: Description of variables used in the study

Table EC.1 Description of variables used in the study.

Variables	Description	Type	Data Source
Acute Admission Source	Patient's admission source to the acute care	Categorical	EHR
Acute Admit Date	Timestamp of patient's admission to acute care	Datetime	EHR
Acute Discharge Date	Timestamp of patient's discharge from acute care	Datetime	EHR
Acute Inpatient Ward	Name of the acute ward that the patient is admitted	Categorical	EHR
Acute Total LOS	Patient's length of stay in acute care (Acute Discharge Date - Acute Admit Date) (days)	Integer	EHR
Acute ALC LOS	Patient's length of stay in ALC status (days)	Integer	EHR
Reason for Rehab	Patient's reason for receiving rehabilitation care	Free text	EHR
Rehab Admit Date	Timestamp of patient's admission to rehabilitation	Datetime	EHR
Rehab Discharge Date	Timestamp of patient's discharge from rehabilitation	Datetime	EHR
Rehab Length of Stay (Days)	Patient's length of stay in rehabilitation (days)	Integer	EHR
Acute MRDiagnosis Category	Most responsible diagnosis category of the patient	Categorical	DAD
Resource Intensity Weight (RIW)	A score that measures how resource intensive the patient's care is	Float	DAD
Sex	Patient's sex (male / female)	Categorical	DAD
Acute Comorbidity Level	Number of comorbidities a patient has	Categorical	DAD
Acute Category	Acute provider program that the patient belongs to	Categorical	DAD
Acute Subcategory	Acute provider subprogram that the patient belongs to	Categorical	DAD

Appendix EC.2: The Impact of Bed Allocation Policy and Processing Time on Delays: The Importance of Using the "Correct" Model

We first analyze the importance of using the "correct" bed allocation policy in the model. Figure EC.1 illustrates how the average queueing time varies with different capacity levels under EP, FCFS, SP, FCFSwP, SPwP. Figure EC.2 illustrates how the probability of waiting less than 30 days varies with different capacity levels under different policies. In general, there is a diminishing return of adding extra rehab capacity. We make a few important observations.

First, we observe that not accounting for the processing time can lead to a substantial underestimation of the required bed capacity. For example, in order to reduce the average queueing times for both acute categories to below 8 days, EP requires at least 51 beds, while FCFS and SP (which do not take the processing time into account) already achieve this performance target with the current capacity level, i.e., 46 beds.

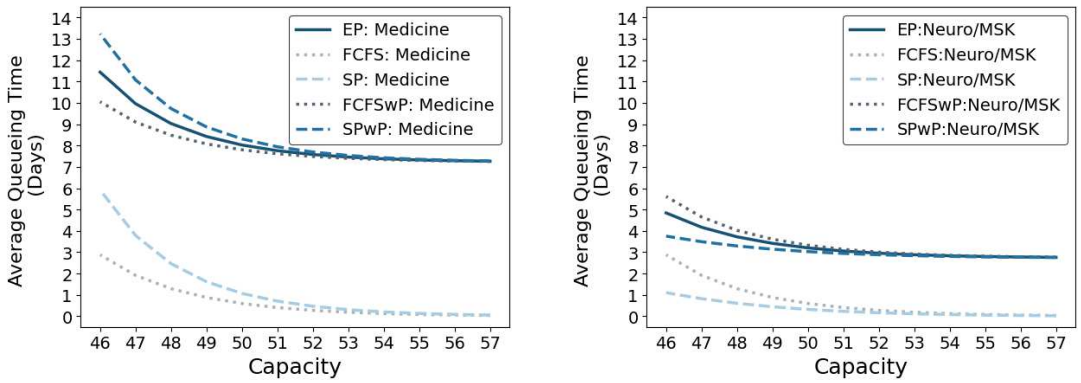


Figure EC.1 **Average queueing times under EP, FCFS, SP, FCFSwP and SPwP policies at different capacity levels.**

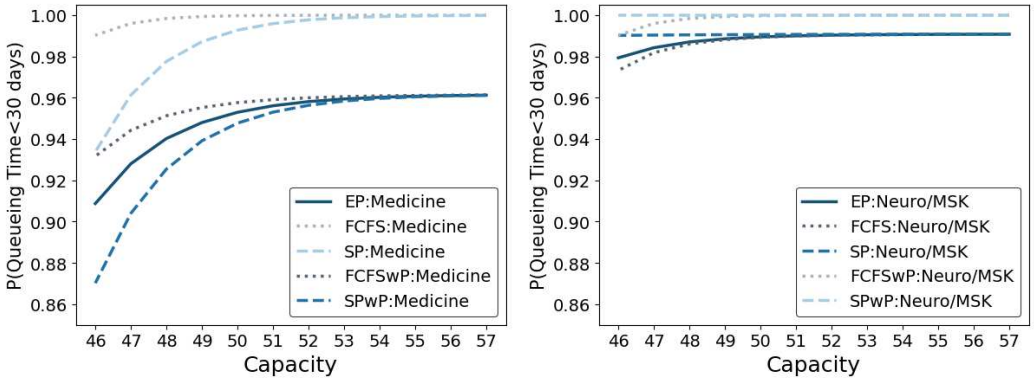


Figure EC.2 **$\mathbb{P}$(Wait <30 days) under EP, FCFS, SP, FCFSwP and SPwP policies at different capacity levels.**

Second, we note that adding extra capacity helps reduce the waiting time but not the processing time. For example, by adding extra beds, the average queueing time for Medicine patients cannot be reduced to below 7.22 days which is the average processing time for Medicine patients.

Lastly, using the correct allocation policy is also important in determining the right capacity level, especially for the less prioritized category – Medicine. To demonstrate this, we compare EP to FCFSwP and SPwP. To achieve an average queueing time of 8 days (or less) for Medicine patients, FCFSwP requires one less bed than EP, while SPwP requires one more bed than EP. These seemingly small differences in the number of required beds can still have significant operational and financial implications due to the high operating costs of acute and rehab beds.

Based on the calibrated arrival rates, the current average utilization of the system, which we denote by r, is 85%. In this regime, processing times account for 67% the observed difference in admission delays between Medicine and Neuro/MSK patients. Meanwhile, demand for rehab service is projected to increase, which can lead to a higher level of system utilization, i.e., a more congested rehab unit. To study the effect of processing times and bed allocation policies in more congested systems, we increase r by scaling up the arrival rates.

Figure EC.3 illustrates how the average queueing times under different policies change when the arrival rate increases. We observe that when the system utilization is high, the effect of the bed allocation policy becomes more pronounced. For example, when the arrival rate increases by 10%, the difference in average queueing times between Medicine and Neuro/MSK under EP increases by 28.8 days. In addition, the difference in average queueing times for Medicine patients between EP and FCFSwP increases from 1.4 to 20.0 days, with the additional 18.6-day difference solely explained by the bed allocation policy. Similarly, the difference in the average queueing times for the Medicine patients between EP and SPwP increases from 1.8 days to 24.6 days.

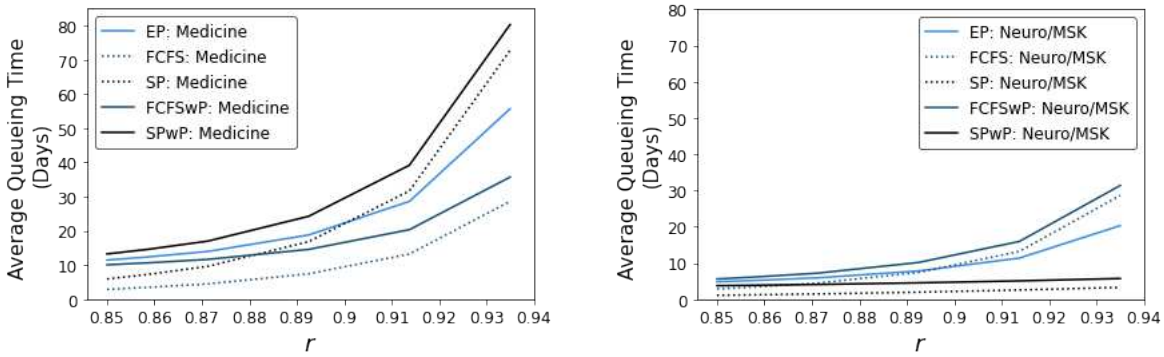


Figure EC.3 Average queueing times under EP, FCFS and SP policies with increasing arrival rates.

Appendix EC.3: The smoothing effect of processing times

Recall that our estimation reveals significantly longer processing times for Medicine patients compared to Neuro/MSK patients. Hence, we investigate the effect of reducing the processing times of Medicine patients without changing the processing times of Neuro/MSK patients. We do this by changing $\phi_{M,1}$ while keeping all other parameters fixed. The results of the experiment are summarized in Table EC.2.

Table EC.2 **Average queueing and waiting times under EP for different values of $\phi_{M,1}$ while keeping $\phi_{M,2}$, $\phi_{N,1}$ and $\phi_{N,2}$ the same. Standard errors are less than 0.5% of the estimates.**

	Queueing Times			Waiting Times		
$\phi_{M,1}$	All	Medicine	Neuro/MSK	All	Medicine	Neuro/MSK
0.350	7.02	11.44	4.84	3.01	4.22	2.11
0.480	6.51	10.03	4.82	3.00	4.25	2.09
0.545	6.26	9.32	4.81	2.99	4.27	2.08
0.610	6.01	8.62	4.81	2.98	4.28	2.07
0.675	5.75	7.91	4.80	2.97	4.30	2.06
0.740	5.50	7.21	4.79	2.96	4.32	2.05
0.805	5.25	6.50	4.78	2.95	4.34	2.04
0.870	4.99	5.80	4.77	2.94	4.35	2.04
0.935	4.74	5.09	4.77	2.94	4.37	2.04
1.000	4.49	4.38	4.76	2.93	4.38	2.03

We make two observations. First, as $\phi_{M,1}$ increases, the average queueing time for Medicine patients decreases as expected. Second, as $\phi_{M,1}$ increases, the waiting time for Medicine patients increases slightly while the queueing and waiting time for Neuro/MSK patients decreases. We note that although the magnitudes of the changes in waiting/queueing times are small, they are statistically significant (the standard errors of the estimations are less than 0.5% of the point estimates). This observation can be attributed to two reasons. First, changing the processing time leads to a more smoothed patient arrival pattern, which helps reduce system idleness and results in shorter queues overall. However, the magnitude of the smoothing benefit is relatively small. Since the average service time in our system (60.7 days) is much larger than the period of the arrival rate function (7 days), fluctuation in the arrival rate function has a limited impact on system performance. Second, when the system is less congested, Neuro/MSK patients, who tend to have a smaller RIW, gain more priority over Medicine patients under our estimated bed allocation policy.

We further examine the smoothing effect of processing times observed in Section 6.2 of the paper through a stylized queueing model. Reducing the processing time has the obvious benefit of reducing the queueing time. On the other hand, changing the processing time can change the rate at which patients enter the waiting list. In some cases, the processing time can help smooth the demand, which leads to reduced waiting time (capacity-driven delays).

We consider an $M_t/M/\infty$ queue followed by a $\cdot/M/n$ queue with two classes of customers where the infinite server queue models the processing time. We consider a sinusoidal total arrival rate function $\Lambda(t) = 0.72 + A\sin(2\pi t/p)$, where A is the amplitude of fluctuation and p is the period. We assume patients belong to the two categories with equal probabilities. Let μ_i and γ_i denote the service rate and processing rate of class i customers, $i = 1, 2$. Note that to analyze the congestion in the second queue, we only need the departure rate of the first $M_t/M/\infty$ queue, which is available in closed form and is a Poisson process (see, Eick et al. 1993, Whitt 2014). In particular, the second queue in the tandem queue is an $M_t/M/n$ queue.

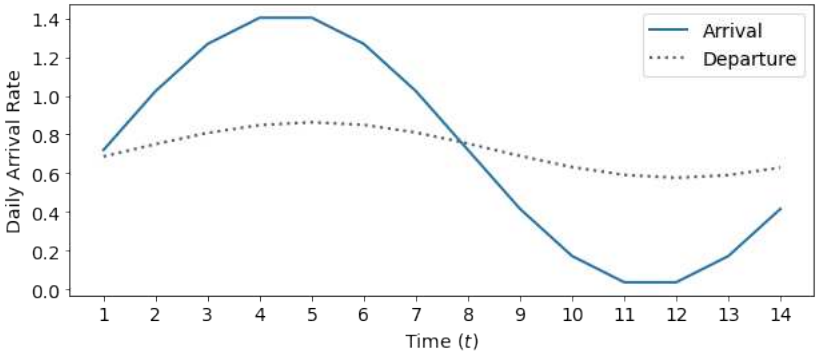


Figure EC.4 **Arrival and departure rate functions of the $M_t/M/\infty$ queue ($A = 0.7$, $\psi = 50$, $p = 14$, $\gamma_1 = 0.1$, $\gamma_2 = 0.5$).**

To demonstrate the smoothing effect of the processing time, Figure EC.4 plots the total arrival rate and departure rate of the first $M_t/M/\infty$ queue. We set $A = 0.7$, $p = 14$, $\gamma_1 = 0.1$ and $\gamma_2 = 0.5$. We observe that due to the heterogeneity in the service rates (processing rates) of the two classes, the departure rate curve is smoother (has an smaller amplitude) than the arrival rate curve.

Next, we investigate the effect of demand smoothing on system performance for a system similar to our setting. In Table EC.3, we provide simulation estimates of the average waiting time in an $M_t/M/n$ queue with two classes of customers where class 1 accounts for 35% of the arrivals. Set $\mu_1 = \mu_2 = 1/60$ and $n = 47$. We consider the total arrival rate function $\Lambda(t) = 0.72 + A\sin(2\pi t/p)$ and vary the values of A and p. We also consider two bed allocation policies: FCFS and SP in favor of class 1.

We make the following observations. First, as expected (e.g., Green et al. 1991), as the amplitude of the sinusoidal function A increases, the average waiting time also increases. The deterioration in performance with the increased amplitude is more severe in systems with a larger period relative to the service time. Note that the average service time is around 60 days in our system. When $p \leq 15$, the performance does not change significantly as A increases. On the other hand, when $p = 60$, the overall average waiting time increases by 47% when A increases from 0 to 0.7. Second,

Table EC.3 **Average waiting times under $M_t/M/n$ queue under varying p and A.**

		Avg. Waiting Times					
		FCFS			SP		
p	A	All	Class 1	Class 2	All	Class 1	Class 2
7	**0.0**	2.76	2.76	2.76	2.76	0.81	6.38
	0.1	2.76	2.76	2.76	2.76	0.81	6.38
	0.3	2.76	2.76	2.76	2.76	0.81	6.37
	0.5	2.77	2.77	2.77	2.77	0.82	6.38
	0.7	2.77	2.77	2.77	2.77	0.84	6.36
15	**0.0**	2.76	2.76	2.76	2.76	0.81	6.38
	0.1	2.76	2.76	2.76	2.76	0.81	6.38
	0.3	2.77	2.77	2.77	2.77	0.83	6.38
	0.5	2.79	2.79	2.79	2.79	0.86	6.37
	0.7	2.82	2.82	2.82	2.82	0.92	6.35
30	**0.0**	2.76	2.76	2.76	2.76	0.81	6.38
	0.1	2.76	2.76	2.76	2.76	0.81	6.37
	0.3	2.81	2.81	2.81	2.81	0.86	6.42
	0.5	2.91	2.91	2.91	2.91	0.97	6.53
	0.7	3.11	3.11	3.11	3.11	1.12	6.80
60	**0.0**	2.76	2.76	2.76	2.76	0.81	6.38
	0.1	2.78	2.78	2.78	2.78	0.82	6.42
	0.3	2.99	2.99	2.99	2.99	0.95	6.76
	0.5	3.40	3.40	3.40	3.40	1.21	7.45
	0.7	4.06	4.06	4.06	4.06	1.62	8.59

when comparing FCFS with SP, the performance of both high- and low-priority queue deteriorates as A increases. In addition, the high-priority class incurs a larger percentage increase in average waiting time than the low priority class.

We now use the above insights to explain the observations made in Table EC.2 of the paper. First, the average service time in our system is much larger than the period for the arrival rate function. In particular, the average service time in our system is 60.7 days, while the period for the arrival rate function is 7 days. In this regime, fluctuations in the arrival rate function do not have a significant impact on system performance. Second, varying the processing time does not give rise to a substantially smoother arrival rate function as illustrated in Figure EC.5. Third, recall that we do not allow discharges on weekends. This pushes more admissions to weekdays, which reduces the effect of smoothing.

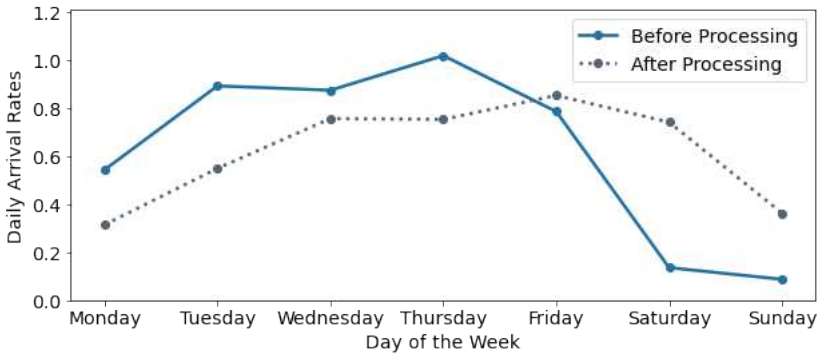
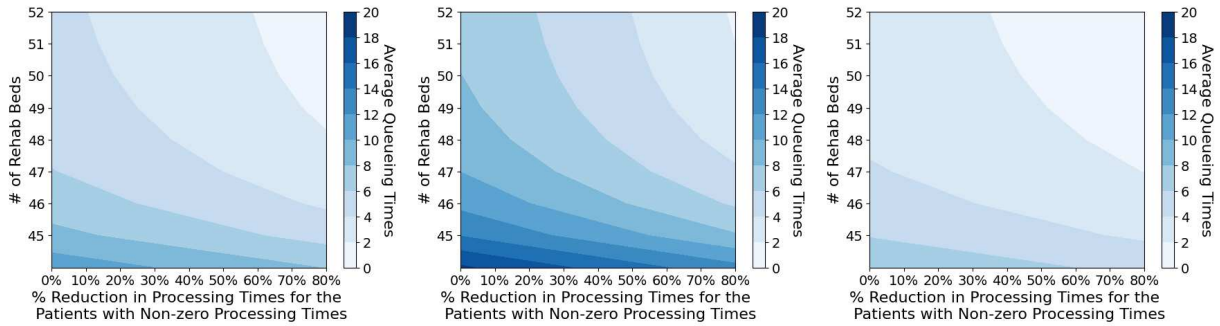


Figure EC.5 **Arrival rate function before and after processing.**

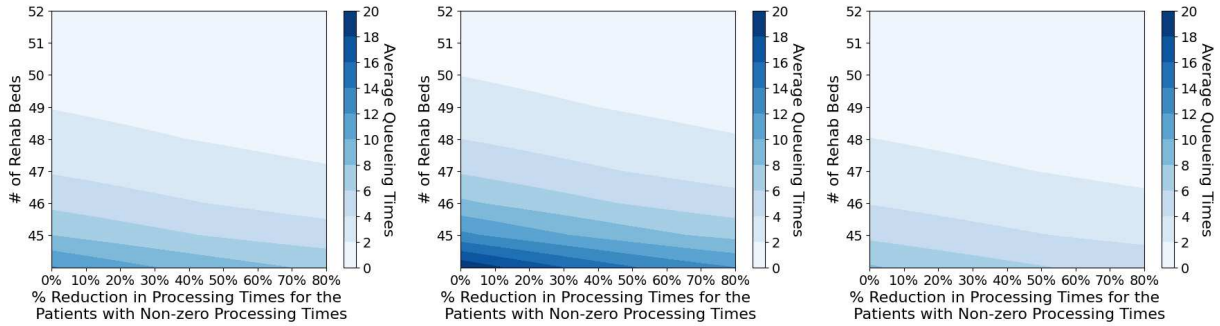
Appendix EC.4: Heatmaps of Average Queueing Times

In this section, we provide the heatmaps of average queueing time as a function of rehab capacity and the percentage reduction in average processing times for all patients as well as Medicine and Neuro/MSK patients separately. Figures EC.6 and EC.7 respectively provide the heatmaps for the original system and the system with the early patient transfer. The observations are consistent with those presented in Sections 6.3 and 6.4.



(a) All patients (b) Medicine (c) Neuro/MSK

Figure EC.6 Average queueing times as a function of the # of rehab beds and % reduction in processing times for the patients with non-zero processing times.



(a) All patients (b) Medicine (c) Neuro/MSK

Figure EC.7 Average queueing times as a function of the # of rehab beds and % reduction in processing times for the patients with non-zero processing times when early transfer to rehab is allowed.

Appendix EC.5: Early Patient Transfer with Priority to Available Patients

In this section, we present an alternative version of the early patient transfer scenario introduced in Section 6.4. In this scenario, patients can be transferred to rehab before they complete their processing times. However, available patients are prioritized. Compared to the original early transfer scheme, we observe an additional 0.72-day reduction in the average queueing time. This is because by prioritizing the available patients, we can more efficiently utilize the rehab beds and reduce the rehab LOS.

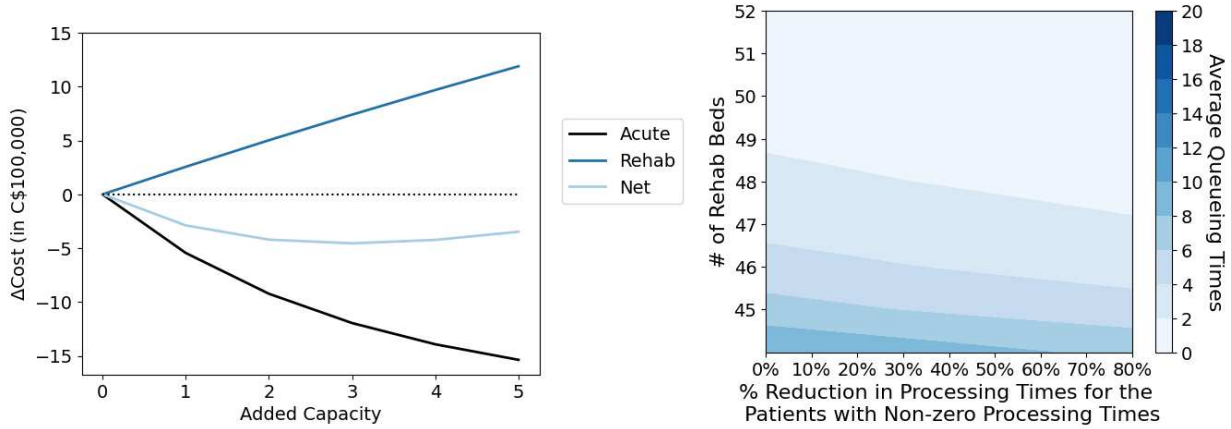


Figure EC.8 Change in costs as a function of the number of additional beds (left); average queueing times as a function of the # of rehab beds and % reduction in processing times for the patients with non-zero processing times when early transfer to rehab is allowed and priority is given to available patients (right).

Figure EC.8 illustrates the effect of capacity expansion and provides a heatmap of the average queueing time as a function of the number of rehab beds and the percentage reduction in average processing time for patients with non-zero processing times. First, we observe that adding additional rehab beds creates smaller cost savings than the original (without prioritization of available patients) early patient transfer policy. For example, adding two more beds yields a net cost-saving of $C\$420,632$ per year compared to $C\$603,060$ per year in the original early patient transfer policy. This is because queueing times are already low in the new early transfer scheme. Adding additional rehab beds leads to lower reductions in queueing times. Second, similar to the original early patient transfer scheme, we observe a linear structure in the heatmap. Reducing processing times creates a larger reduction in queueing times when rehab is heavily loaded. Similarly, adding rehab beds create a larger reduction in queueing times when the processing times are long.

Appendix EC.6: Analyzing the Effect of External Admissions on the Waiting Times

In this section, we estimate the impact of external admissions, i.e., patients that are admitted outside the hospital on the ALC LOS of the patients transferring to rehab from the same hospital's acute care. To this extent, we consider two exogenous variables: (1) the total number of external admissions in rehab on the day that the patient received ALC status (*ExInRehab*) and (2) the number of external admissions that occurred on the day that patient received ALC status (*ExAdmRehab*). ExInRehab corresponds to the total number of external admissions in rehab whereas ExAdmRehab counts the external admissions that occurred on that day.

Consider the following models each associated with one of our exogenous variables:

$$\textbf{Model 1}: \quad ALC\ LOS_i = 1 + Sex_i + Acute\ Category_i + RIW_i + Age_i + Congestion_i + ExInRehab_i + \epsilon_i,$$

$$\textbf{Model 2}: \quad ALC\ LOS_i = 1 + Sex_i + Acute\ Category_i + RIW_i + Age_i + Congestion_i + ExAdm_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma)$ and Congestion is defined as the total number of patients in rehab and waiting to be admitted to rehab. Table EC.4 illustrates the estimation results. The results indicate that ExInRehab and ExAdmRehab do not have a significant effect on ALC LOS. This suggests that external admissions do not delay the admission of internal patients, supporting the assumption that internal patients are prioritized for rehab admission.

Table EC.4 Effect of ExInRehab and ExAdmRehab on ALC LOS.

Covariates	Estimates	
	Model 1	**Model 2**
Intercept	-7.14 (7.35)	-8.88 (7.15)
Sex: Male	-0.79 (0.85)	-0.78 (0.85)
Acute Category: Medicine	-6.93*** (0.87)	-7.0*** (0.87)
Acute Category: Others	-7.81*** (1.91)	-7.91*** (1.92)
RIW	0.7*** (0.12)	0.7*** (0.12)
Age	0.01 (0.03)	0.01 (0.03)
Congestion	0.32*** (0.12)	0.33*** (0.12)
ExInRehab	-0.22 (0.23)	-
ExAdmRehab	-	0.05 (1.06)