Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Who Is Next: Patient Prioritization under Emergency Department Blocking

Wenhao Li, Zhankun Sun
Department of Management Sciences, College of Business, City University of Hong Kong, Kowloon, Hong Kong
wenhaoli6-c@my.cityu.edu.hk, zhankun.sun@cityu.edu.hk

L. Jeff Hong
Department of Management Science, School of Management, Fudan University, Shanghai, China
hong_liu@fudan.edu.cn

Upon arrival at emergency departments (EDs), patients are classified into different triage levels indicating their urgency. Using data from a large hospital in Canada, we find that within the same triage level, the average waiting time (time from triage to initial assessment by a physician) of patients who are discharged is shorter than that of patients who are admitted for middle-to-low acuity patients, suggesting that the order in which patients are served deviates from FCFS (first-come-first-served), and to a certain extent, discharged patients are prioritized over admitted patients. This observation is intriguing as among patients of the same triage level, admitted patients—who need further care in the hospital—should be deemed no less urgent than discharged patients who only need treatment at the ED. To understand how ED decision makers choose the next patient for treatment, we estimate a discrete-choice model and find that ED decision makers apply urgency-specific delay-dependent prioritization. Moreover, we find that when the ED blocking level is sufficiently low, admitted patients are prioritized over discharged patients for high acuity patients, whereas disposition does not affect the prioritization of middle-to-low acuity patients. When the ED blocking level becomes sufficiently high, decision makers start to prioritize discharged patients in an effort to avoid further blocking the ED. We then analyze a stylized model to explain the rationale behind the change in decision makers' prioritization behavior as the ED blocking level increases. Using a simulation study, we demonstrate how policies inspired by our findings improve ED operations by reducing the average patient waiting time and length of stay, resulting in significant cost savings for hospitals. We also show how to leverage our findings to improve the accuracy of ED waiting time predictions. By testing and highlighting the central role of decision makers' patient prioritization behavior, this paper advances our understanding of ED operations and patient flow.

Key words: Patient Prioritization, ED Blocking, Discrete-Choice Model, MDP, Simulation, Waiting Time Prediction

1. Introduction

Emergency department (ED) waiting times—the total time from triage to initial assessment by a physician—is a well-established metric of the timeliness of emergency care. Unfortunately, long waiting times are extremely common in many countries around the world. In the United States, one out of four patients waited more than 2 hours to see a physician in 2006 (United States Government Accountability Office 2009). In 2015–2016,

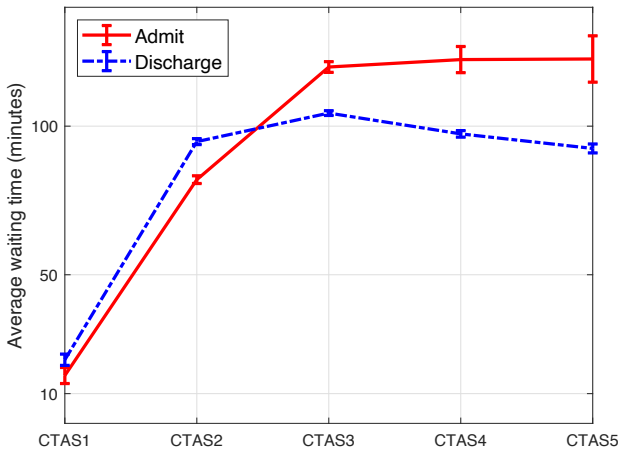
one out of three Canadian patients reported waiting 4 or more hours the last time they sought care at an ED (Canadian Institute for Health Information 2016). Prolonged waiting times are associated with increased morbidity and mortality (Sun et al. 2013), patients left without being seen (Weiss et al. 2005), and increased rate of readmission (Richardson and Bryant 2004), among other consequences. Hence, it is critical to understand the determinants of ED waiting times. One key determinant is the order in which patients are seen by ED healthcare professionals. Upon arrival at an ED, patients are first triaged by a nurse, and a score is assigned based on the patients' acuity (some triage protocols also consider patient resource needs). In Canada, the most prevalent triage protocol is the Canadian Triage and Acuity Scale (CTAS), a 5-point scoring scale (1 to 5) with smaller number indicating a higher level of urgency. Similar to other triage protocols, CTAS focuses on classifying patients rather than providing explicit guidelines as to which patients to prioritize (Murray et al. 2004). CTAS proposes a fractile response objective for each triage level (e.g., 90% of CTAS 3 patients should be seen within 30 minutes of arrival); however, as stated clearly in the implementation guidelines, these time responses are not established care standards (Beveridge et al. 1998). Without explicit guidelines, decision makers use their discretion to select the next patient to receive care, and the order of patients being seen can deviate significantly from the one that sorts patients by their triage levels and arrival times. This is the case in our study hospital and has also been found by Ding et al. (2019) using data from EDs in metro Vancouver. In other words, how exactly ED decision makers (nurses and physicians) select the next patient from all patients waiting to be seen, given their triage scores, waiting times, and resource availability, does not follow an explicit rule.

Despite the importance of understanding how patients are prioritized in EDs, the literature on this topic is limited. A recent study (Ding et al. 2019) finds that when EDs are critically loaded, decision makers (i) apply delay-dependent prioritization in selecting patients across different triage levels; (ii) follow the first-come-first-served (FCFS) rule in general for patients of the same triage level. However, the adherence decreases if patients wait longer than certain threshold. What we observe additionally in our study hospital is that even within the same triage level, decision makers may deliberately deviate from the order of FCFS. We elaborate below.

After treatment in an ED, patients who need further care at inpatient units are admitted. Hereafter, we call them *admit patients* and call those who do *not* need further care at the hospital *discharge patients*. The destination of patients after their treatment in the ED is called their *disposition*. Using patient-level data from an urban tertiary hospital in the Calgary Zone of Alberta, Canada, we observe that the average waiting time of admit patients is less than that of discharge patients for CTAS 1 and 2; however, for CTAS 3, 4, and 5, admit patients wait longer than discharge patients on average (significant at the 5% level), suggesting that discharge patients are, to a certain extent, prioritized over admit patients (see Figure 1).[1] This observation

[1] Our data shows that the number of admit patients delayed by a discharge patient, i.e., who arrives later but gets treated earlier, is higher than that of discharge patients delayed by an admit patient for triage levels 3–5, but lower for levels 1 and 2, which provides more evidence for the conjectured prioritization behavior.

Figure 1 Comparisons between the average waiting times (in minute) of admit patients and discharge patients by triage levels in an urban tertiary hospital in Alberta, Canada. The error bars represent 95% confidence intervals.



is intriguing because among patients of the same triage level, admit patients—who need further care after ED treatment—should be deemed no less urgent than discharge patients. Thus, one may wonder why ED decision makers prioritize discharge patients over admit patients within the same triage level. This question motivates our investigation.[2]

The discrepancy in waiting times for patients within the same triage level is a reflection of the patient prioritization behavior of ED healthcare professionals. To the best of our knowledge, this behavior has not been studied in the literature. Our objective is to shed light on ED patient prioritization. Specifically, we focus on three research questions: How does disposition affect the prioritization of patients from the same triage level? What is the rationale behind ED healthcare professionals' prioritization behavior? Importantly, how can we leverage our knowledge on patient prioritization to improve ED operations?

1.1. Contributions to the Literature

To answer the research questions, we conduct: (i) an empirical study to understand decision makers' patient prioritization behavior; (ii) an analysis of a stylized decision model to explain the rationale behind this behavior; and (iii) a simulation study and a forecasting study to show the value of our findings.

Our study makes the following contributions to the operations management, queueing, and forecasting literature. First, we empirically examine patient prioritization decisions by applying a discrete-choice model to patient-level visit records. We assume that when selecting the next patient, decision makers can estimate the disposition of a patient and are aware of the *ED blocking level*, which measures the extent to which

[2] One possible explanation of Figure 1 is the inclusion of fast-track patients. Fast-track patients are streamed into a different queue after triage and treated by a dedicated team in a separate area. Most fast-track patients are discharged after their treatment in the ED. It is possible that fast-track patients having low waiting time drives this observation. However, the figure barely changes when we remove fast-track patients from the dataset, and the puzzle remains.

the ED's ability to treat new patients is compromised by too many ED beds being occupied by *boarding patients*—admitted patients waiting in ED beds to be transferred. Using a conditional logit model, we find that ED decision makers generally apply urgency-specific delay-dependent prioritization. That is, higher urgency and longer waiting time both lead to higher priority. In addition, we find that (i) when the ED blocking level is sufficiently low, admit patients are prioritized over discharge patients for high acuity patients (triage level 2), and disposition does not affect the prioritization of middle-to-low acuity patients (triage levels 3, 4, and 5) within the same triage level; (ii) as the blocking level increases, decision makers start to prioritize discharge patients. To our knowledge, this is the first work that empirically studies how ED decision makers prioritize patients in response to ED blocking.

Second, we develop a Markov decision process (MDP) model to study how patients are selected from the same triage level under ED blocking. We show that it is optimal for decision makers to dynamically prioritize patients; in particular, prioritize discharge patients when the ED blocking level is high. The structure of the optimal policy provides insights into the rationale behind decision makers' patient prioritization behavior. Our model also captures the interaction between the dual resources that decide ED's capacity of treating patients, i.e., physicians and beds, so as to study patient prioritization in a multi-class two-station queueing network. Hence, we also make contributions to the queueing literature.

Third, by means of a data-calibrated simulation, we provide managerial insights into how to leverage our findings to improve ED operations. Specifically, we devise patient prioritization policies based on our empirical findings and the optimal policy from our MDP model. Our simulation results show that policies that prioritize admit (discharge) patients when ED blocking level is low (high) can reduce the average waiting time and length of stay for both types of patients. As a result, hospitals can save millions of dollars annually. This suggests that we can improve ED operations and achieve cost saving by altering the order in which patients are treated. We believe this has important implications for hospital management. We also demonstrate with real data that our findings can improve the accuracy of leading-edge algorithms used to predict ED waiting time. Hence, we make contributions to ED practice and the forecasting literature.

1.2. Organization

The rest of this paper is organized as follows. We discuss the relevant literature in Section 2 and introduce the study setting in Section 3. We empirically study how decision makers choose the next patient in Section 4. In Section 5, we develop an MDP formulation and connect its optimal policy to our empirical findings. In Sections 6 and 7, we show how to leverage our findings to improve ED operations and waiting time prediction, respectively. In Section 8, we discuss how our results relate to the shortest processing time rule. In Section 9, we conclude this work, discuss the managerial insights, and point to future research directions. All proofs and additional results are given in the e-companion.

2. Literature Review

In recent years, operations research/management tools have been widely applied to improve patients' access to emergency care (see Saghafian et al. (2015) and Dai and Tayur (2020) for overviews). Studies of healthcare professionals' behavior are particularly relevant to our study. Healthcare professionals have been found to respond to system workload by adjusting their service speed and capacity rationing decisions; see, e.g., KC and Terwiesch (2009, 2012), Powell et al. (2012), Kuntz and Sülz (2013), Kim et al. (2014), Batt and Terwiesch (2016), Freeman et al. (2016), Berry Jaeker and Tucker (2017), Ding et al. (2019), Kim et al. (2020). These studies suggest that the decisions of healthcare practitioners are not driven purely by clinical factors, and they identify various mechanisms to explain these behavior. Our work studies the patient prioritization behavior of ED nurses and physicians and thus is relevant.

Among these works, Ding et al. (2019) is the most relevant one. Using data from EDs in metro Vancouver, Ding et al. (2019) empirically show that when EDs are highly loaded, delay-dependent prioritization is applied in selecting new patients. They also find that FCFS is generally followed in the same triage level but the deviation increases when patients wait past certain thresholds. Our work aligns with Ding et al. (2019) in that we also empirically study patient prioritization decisions in Canadian EDs. Ding et al. (2019) state that "physicians are the bottleneck resources in the treatment process" in their study hospitals, whereas both beds and physicians can be bottlenecks in our ED. Hence, in addition to factors such as triage levels and waiting times, bed capacity may impact patient prioritization decisions. This impact is the focus of our study, which differentiates our work from Ding et al. (2019).

Another stream of relevant papers concerns the phenomenon of *ED blocking* or *bed block*, referring to situations that too many beds occupied by admit patients (to be transferred to inpatient units) leads to insufficient ED beds for new patients. Despite being *the most significant factor* causing ED overcrowding in many countries around the world (Olshaker and Rathlev 2006, Pines et al. 2011, Affleck et al. 2013), research on this topic is relatively limited (Saghafian et al. 2015). Two recent papers (Shi et al. 2015, Chan et al. 2016) study this problem by controlling the discharge (inspection) timing in inpatient wards. They model the inpatient flow dynamics through time-varying queues and propose policies that potentially alleviate ED blocking. Our work fills a gap in the literature, as we focus on how ED decision makers can control the demand for inpatient beds so as to relieve ED blocking.

A number of papers have investigated various aspects of EDs, and thus are relevant to our study. The research topics that have been studied include, among others, waiting time prediction (Ibrahim and Whitt 2011, Ang et al. 2015), the impact of delay announcement (Dong et al. 2019), and ambulance diversion decisions (Deo and Gurvich 2011, Allon et al. 2013). Furthermore, there is a large body of empirical studies on EDs and other service systems, see, e.g., Batt and Terwiesch (2015), Song et al. (2015), Ibanez et al. (2018), Tan and Staats (2020), KC et al. (2020). Among them, Ibanez et al. (2018) is particularly

relevant, as they study the task ordering decisions of radiologists and their impact on productivity. They find that radiologists prioritize tasks by similarity and the shortest processing time rule (SPT), yet both erode productivity. Similarly, KC et al. (2020) find that ED physicians prioritize easier tasks when faced with increasing workloads, which is detrimental to throughput and learning. Our work is similar in that we also study how decision makers prioritize tasks (patients). We identify a related but different mechanism than the SPT rule underlying the prioritization decisions (see Section 8). We note that two recent works use ED visit data to estimate the impact of low-acuity patients on the waiting time of high-acuity patients (Luo et al. 2017) and physician task switching cost (Duan et al. 2020) and thus are relevant. However, the objectives of these studies are different from ours.

Finally, our study is relevant to the queueing literature. Our modeling of the dynamics between physician assessment and testing is similar to the studies of de Véricourt and Jennings (2011), Dobson et al. (2013), Yom-Tov and Mandelbaum (2014), Huang et al. (2015), Campello et al. (2016), Carmen et al. (2018), and Çağlayan et al. (2019), which model the repetitive services provided to patients by either nurses or physicians with the aim of improving staffing or for performance evaluation. The model in van Leeuwaarden et al. (2016) also captures the dual resource constraints in healthcare systems with repeated services. Our work differs in that we focus on patient prioritization decisions, whereas van Leeuwaarden et al. (2016) focus on staffing. Models in Saghafian et al. (2012, 2014) are closely related to ours. Saghafian et al. (2012) study patient streaming decisions depending on predicted dispositions in a clearing model. Saghafian et al. (2014) study how physicians choose the next patient by an MDP model and patient sequencing at triage by a priority queue. They develop a complexity-augmented triage rule and demonstrate the validity and magnitude of its performance improvement by simulation. The constraints on physician capacity and the effect of ED blocking are discussed in their paper but not explicitly modeled (unlike our work).

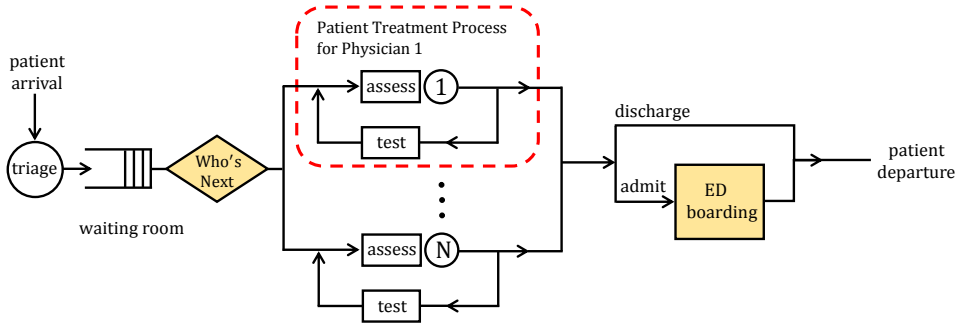
3. ED Patient Flow and Data

In this section, we describe the ED patient flow process and the data used in our empirical investigation. We note that different EDs may operate differently. Although our description is based on a Canadian model, we believe that the key characteristics are similar in most EDs. The general patient flow in the main ED area (the fast-track area is not included) is depicted in Figure 2.

3.1. Patient Flow

Patients arrive to the ED either by emergency medical services (EMS, such as ambulance) or by their own transportation. Upon arrival, patients are triaged into one of five categories. Following triage, patients wait in the waiting room to be seen by medical staff. When a physician finds herself able to see a new patient, a patient waiting to be seen is selected for initial assessment. We note that the selection is possibly a multi-step process involving more than one decision maker. In most EDs, the chief nurse decides which patient

Figure 2 General patient flow in the main area of an emergency department.



Note. The numbers in circles, i.e., 1 to N, represent the N physicians who work in parallel and perform assessments. The red dashed line encircles the flow of patients receiving treatment from Physician 1. The patient treatment process is similar for other physicians, and thus not signified in this figure.

to move to a treatment room, and a physician selects a patient for assessment from the roomed patients. Physicians occasionally select patients from the waiting room directly (Ding et al. 2019). For simplicity, we aggregate the selection process into a single step through which a decision maker chooses the next patient for assessment. We believe that this aggregated selection process can still reflect the patient prioritization decisions in EDs. It would be interesting to investigate the decisions of nurses and physicians separately if there was a sufficient level of detail in the data.

After the initial assessment, some patients may leave the ED, while others may undergo medical procedures, such as diagnostic tests (lab tests, imaging, etc.) or treatment by nurses. For simplicity, we refer to all procedures done by non-physician staff as *tests*. These patients join the test queue (see Figure 2), where they go through testing and wait for the results. The patient returns to the same physician for another round of assessment when the test results are ready. This assess-test process may be repeated; that is, a patient may see the same physician several times before a diagnostic decision is made. ED physicians are multi-tasking (KC 2013), i.e., at any time a physician is responsible for the care of several patients, some undergoing testing in the *test queue* and others waiting to see the physician in the *assess queue* (see Figure 2). According to our collaborating physicians, a physician knows her service capacity, i.e., the maximum number of patients that she can care for. This aligns with the descriptions in Saghafian et al. (2012) and Campello et al. (2016) and has important implications for our stochastic modeling in Section 5.

3.2. ED Blocking

An ED bed is assigned to a patient once she is selected for assessment. The bed can be an exam room or simply a moving stretcher. The assigned bed is held for the patient until she leaves ED, even when she is sent to the test center for testing. This is a common practice in many EDs (Saghafian et al. 2012, Armony et al. 2015). If a patient is admitted, an inpatient bed will be requested and the patient remains in the ED bed until

she is transferred. The total time from bed request to the time of transfer is called *boarding time* and patients awaiting transfer are called *boarding patients* or *boarders*. Boarding time is regarded as non-value adding since ED diagnosis and treatment are complete. Moreover, boarding time can be extremely long (Armony et al. 2015, Shi et al. 2015). When boarding patients occupy ED beds for prolonged periods of time, they block access to these spaces for newly arriving patients in the waiting room. We refer to this phenomenon as *ED blocking*, which is identified as the main cause of ED overcrowding in many countries (Olshaker and Rathlev 2006, Pines et al. 2011, Affleck et al. 2013). Note that all patients staying in ED beds need nursing care. Hence, it is impossible to solve the ED blocking problem by adding more beds without increasing the nurse staffing level.

3.3. Data Description and Cleaning

This study uses data from the ED electronic health record system of an urban hospital in Alberta, Canada. The dataset covers all patient visits from August 2013 to July 2015. The daily arrivals to this ED over the study period range from 131 to 257, and the median (mean) is 206 (207.4). Each observation includes a patient's triage information (age, gender, arrival mode, triage level, chief complaint), arrival time (time at which the patient is triaged), initial assessment time (time at which the patient is first selected by a physician for assessment), bed request time for admit patients (start time of boarding), last contact time (time at which the patient leaves the ED), and disposition. There are 145,162 observations after removing records with incomplete or wrong entries, such as a negative age or negative waiting time. All visit records are de-identified to protect the privacy of the patients and medical personnel.

In our study ED, a fast-track line is open daily for 14 hours from 10 am to midnight, which coincides with the high-load period of the ED during which there are on average at least 45 patients in the ED and at least 10 patients waiting to be seen. We note that the high-load period in our study ED overlaps that in Armony et al. (2015), which starts between 9 am and 12 pm and ends between 11 pm and 3 am (next day). We are interested in the patient prioritization behavior of ED decision makers during high-load hours in the main ED (the non-fast-track area) as that is the most congested area. Hence, fast-track patients (25,160 observations, 17.3% of the data) and patient visits outside of the high-load hours (30,927 observations, 21.3% of the data) are eliminated. Another motivation of removing the latter is that triage nurses occasionally defer the treatment of some patients waiting in the main ED so as to route them into fast-track later, when they know the fast-track line will be open soon, which complicates the prioritization process. We are interested in studying how a patient's priority is determined by decision makers, hence we drop patients of triage level 1 because most of them have life-threatening conditions and receive preemptive priority over all other patients (3,280 observations, 2.3% of the data). We also drop patients whose dispositions are *not* "admit" or "discharge" (including "left without being seen," "left against medical advice," etc.) as they are not our focus (2,581 observations, 1.8% of the data). We combine patients of triage levels 4 and 5 into one single low-acuity

Table 1 Summary statistics for variables of interest. ‡

Variables	CTAS2 (High-Acuity)	CTAS3 (Middle-Acuity)	CTAS4,5 (Low-Acuity)
Waiting time (mean and stdev in hours)			
Admit patients	1.39 (1.50)	2.11 (1.67)	2.20 (1.65)
Discharge patients	1.65 (1.55)	1.92 (1.57)	1.76 (1.45)
Time in ED beds (mean and Stdev in hours)			
Admit patients	10.02 (7.29)	10.24 (7.07)	9.79 (7.00)
Discharge patients	3.65 (3.13)	2.94 (2.78)	2.01 (2.22)
Arrival mode (ambulance arrival%)	38.7%	28.1%	17.5%
Gender (female%)	50.6%	57.8%	57.5%
Disposition (admit%)	34.6%	23.6%	11.1%
Age groups (occurrences)			
0 to 18 years	518	493	319
18 to 40 years	10,530	12,316	6,729
40 to 55 years	7,652	6,835	3,082
55 to 70 years	7,724	6,561	2,447
Over 70 years	7,689	7,801	2,621
Address †			
Region A	17,020	19,078	8,990
Region B	5,202	4,663	1,890
Region C	5,907	5,454	2,330
Region D	3,092	2,474	960
Others	2,892	2,337	1,028
Chief Complaint Codes (*Five codes with most occurrences*)			
Shortness of Breath	1,665	1,811	659
Chest Pain (Cardiac Features)	4,590	0	0
Headache	1,542	1,244	441
Depression / Suicidal	1,278	689	144
Vomiting And / Or Nausea	457	1,112	580
Observations	34,113	34,006	15,198

† We have the district level information for address. Region A (D) is the nearest (farthest).
‡ The correlation coefficients between variables can be found in Tables 5, 8, and 9 in the e-companion.

patient group to make its sample size more comparable to that of high and middle acuity patient groups, i.e., triage levels 2 and 3, respectively. Table 1 shows the summary statistics of the data after cleaning.

We apply a discrete-choice model to the dataset to investigate how decision makers choose new patients. The decision epochs correspond to the times when a decision maker decides to see a new patient. We describe the outcome and explanatory variables below.

3.3.1. Outcome Variable At time t, if a decision maker finds herself available and decides to see a new patient, she chooses one among all patients waiting to be seen at the ED, which composes the choice set at t, denoted by $J(t)$. Thus, the outcome variable in our study is whether patient j is chosen at decision epoch t, $j \in J(t)$. At any decision epoch, only one decision maker makes a choice, and one and only one patient is selected. Note that $J(t)$ is dynamic and time-dependent. Consider two consecutive decision epochs t_1 and t_2 ($t_1 < t_2$), then $J(t_2)$ contains all patients in $J(t_1)$ and patients that arrive between t_1 and t_2, less the patient selected at t_1 and patients who become absent from ED between t_1 and t_2 (because of leaving without being

seen, being transferred, etc.).

3.3.2. Explanatory Variables Over the study period, the medical personnel's compensation is in the form of shift-based salary. To our knowledge, there is no financial incentive for decision makers to select patients based on the complexity or medical expense. This concurs with observations by Ding et al. (2019) from Canadian EDs and Song et al. (2015) based on US EDs. When deciding which patient to select, we believe that ED decision makers' objective is to provide timely care to the patients who need it most urgently. Hence, we focus on clinical and operational factors that are potentially related to that objective. Motivated by the intriguing observation in Figure 1, one key variable of interest is a patient's disposition. Since the actual dispositions of patients in $J(t)$ are unknown at decision epoch t and are only revealed after the completion of ED treatment, we use the predicted disposition as a proxy (see more details in Section 4.2). We study how disposition affects patient prioritization by controlling the ED blocking level, which measures the extent to which the ED's ability to treat new patients is compromised by the ED beds being occupied by boarding patients at t. We control for ED crowding by including the number of patients waiting to be seen in our model. We use triage levels to control the discrepancy in patient prioritization across different triage levels. We also control the heterogeneity of patients within the same triage level by the *chief complaint codes*, such as "Chest Pain (Cardiac Features)," "Abdominal Pain," "Headache," etc. Other control variables include age group, gender, arrival mode, and patient waiting time thus far, all of which are categorical variables except the last one. A physician cannot observe when the last patient was selected if the patient was selected by other physicians. Hence, we believe that the time duration between decision epochs is not considered in patient prioritization decisions.

4. Empirical Investigation of Patient Prioritization

In this section, we empirically examine ED decision makers' prioritization behavior in our study ED. We first state the following assumption before presenting our empirical model.

ASSUMPTION 1. When selecting the next patient for initial assessment, a decision maker (i) can predict the disposition of a patient and (ii) is aware of the ED blocking level.

The choice of which patient to see next is made in the hospital's information system. Through a terminal, decision makers can access real-time information of all patients waiting to be seen, including their comprehensive triage data (which contains more details than our data, such as vital signs, whether they are revisit patients), waiting time thus far, etc. Previous studies show that ED physicians and nurses can predict a patient's disposition fairly accurately using triage information (see, e.g., Holdgate et al. 2007 and Vaghasiya et al. 2014). The number of boarding patients is available to nurses and physicians in real time.

4.1. Econometric Models

We investigate the relationship between a decision maker's choice of the next patient and the characteristics of the alternatives (patients) under resource constraints (e.g., ED blocking) by a discrete-choice model. We believe that variations in patients' characteristics and system resource constraints are the main drivers of decision makers' selection behavior. Hence, we choose a conditional logit model for our investigation.

Conditional logit models belong to the family of random utility models, in which a decision maker chooses the alternative that maximizes her perceived utility. More specifically, at decision epoch t, let Y_t represent a choice (patient) in the choice set $J(t)$, and U_{it} be the utility of choosing patient i from $J(t)$. We treat U_{it} as independent random variables with a systematic component V_{it} and a random component ε_{it}, i.e., $U_{it} = V_{it} + \varepsilon_{it}$. At any decision epoch t, the decision maker evaluates the utility of each patient in $J(t)$ and selects the one that maximizes her utility. Hence, the probability of choosing patient i from $J(t)$ is:

$$\Pr\{Y(t) = i\} = \Pr\left\{U_{it} = \max_{j \in J(t)} U_{jt}\right\} = \frac{\exp(V_{it})}{\sum_{j \in J(t)} \exp(V_{jt})}, \tag{1}$$

where the last equality holds if the error terms ε_{it} are independently and identically distributed with the standard Type I extreme value distributions (Train 2009). We then estimate the systematic term of the decision maker's utility in choosing patient i at decision epoch t, V_{it}, by maximizing the likelihood of choosing the observed choice of patient. The model is specified as follows:

$$\begin{aligned} V_{it} = {} & \beta_0 + \beta_1^T \mathbf{C}_i + \beta_2^T \mathit{CTAS}_i + \beta_{31}^T \mathit{CTAS}_i \times \mathit{WaitTime}_{it} + \beta_{32}^T \mathit{CTAS}_i \times \mathit{WaitTime}_{it}^2 + \beta_4^T \mathit{CTAS}_i \times \mathit{WaitRoomCensus}_t \\ & + \beta_5^T \mathit{CTAS}_i \times \mathit{BlockLevel}_t + \beta_6^T \mathit{CTAS}_i \times \mathit{Disposition}_i + \beta_7^T \mathit{CTAS}_i \times \mathit{Disposition}_i \times \mathit{BlockLevel}_t. \end{aligned} \tag{2}$$

The vector $\mathbf{C}_i$ contains the time-invariant characteristics of patient i, including age group, gender, arrival mode, and chief complaint, which are the clinical factors that decision makers can access and take into account during patient prioritization. Note that in our model, we use the categorized age groups, instead of the numerical values, to account for the possible nonlinear effect of age on decision makers' utility. The 3-by-1 vector CTAS_i is an indicator of the triage level of patient i. Specifically, $\mathit{CTAS}_i = (1,0,0)^T$ if patient i is of triage level 2, $\mathit{CTAS}_i = (0,1,0)^T$ if patient i is of triage level 3, and $\mathit{CTAS}_i = (0,0,1)^T$ if patient i is of triage level 4 or 5. The scalar $\mathit{WaitTime}_{it}$ represents the current waiting time of patient i, i.e., the time from patient i's arrival to the decision epoch t, and may have a nonlinear impact on patient priority (Ding et al. 2019, Ferrand et al. 2018). Moreover, the impact may vary across triage levels. Hence, we include the interaction terms of *CTAS* and *WaitTime* (both linear and quadratic terms) in Eq. (2). Similarly, we control the number of patients waiting to be seen in the ED at decision epoch t (normalized to be between 0 and 1), denoted by $\mathit{WaitRoomCensus}_t$, which is associated with ED decision making (Gorski et al. 2017). The ED blocking level at decision epoch t, denoted by $\mathit{BlockLevel}_t$, measures the extent to which the ED's ability to provide timely care is impaired.

One challenge in our study is the estimation of $BlockLevel_t$, which depends on the number of boarding patients and ED bed capacity at decision epoch t. Our data provide the number of boarding patients at any time. However, the ED bed capacity depends on both the numbers of physical beds and the time-varying nurse staffing levels, as ED beds can only be used to care for patients if enough nursing staff are on duty to ensure that care is safe and meets patients' needs. In our dataset, we do not observe nurse staffing levels. Hence, we use the 90th percentile of the distribution of the observed numbers of patients in ED beds at any given hour of the day over the 2-year horizon—rounded up to the nearest integer—as a proxy for ED bed capacity at any particular hour of the day.[3] We develop two measures for ED blocking level. The first measure is the ratio of the number of boarding patients over the number of "extra" beds (total bed capacity net of total physician capacity) at time t (referred to as *Measure 1* hereafter). Here, the total physician capacity is the product of the number of physicians on duty and an individual physician's capacity. The former is available in our dataset, and the latter is set to 7 following Saghafian et al. (2012). The second measure of the ED blocking level is the actual number of boarding patients at t normalized to be between 0 and 1 (referred to as *Measure 2* thereafter). Measure 1 seems to be a better measure because it captures the time-varying nature of both bed capacity and physician capacity. However, Measure 2 is simpler and readily available on the ED dashboard. It is unclear how exactly ED decision makers infer the blocking level. Hence, we test both measures and their variations in our empirical investigation.

The anticipated disposition of patient i by ED decision makers, denoted by $Disposition_i$, can be viewed as an exogenous treatment on patient i, which takes the value 1 if the anticipated disposition is *admit*; otherwise, it equals 0. Note that the anticipated disposition is unobservable in our data. Hence, we need to construct proxies (see Section 4.2 for details). To investigate how the treatment affects a patient's priority of being selected across triage levels, we add the interaction of *CTAS* and *Disposition*; we further add the three-way interaction of *CTAS*, *Disposition*, and *BlockLevel*, to show how the effect varies with the ED blocking level.

One may note that the model specification in Eq. (2) includes three-way interaction terms but the corresponding main effects and two-way interactions are not always included. The reasons is that the main effects of $WaitRoomCensus_t$ and $BlockLevel_t$ do not vary over choice alternatives and hence cannot be directly included in the model; otherwise, the model becomes un-identifiable. From a mathematical viewpoint, the definition of the choice probability in Eq. (1) implies that adding a common term to V_{it} for any patient $i \in J(t)$ does not change the choice probability. Hence, to control the effects of $WaitRoomCensus_t$ and $BlockLevel_t$, we need to specify them in ways that create differences in utility over alternatives. We choose to add their interactions with triage levels. The terms $WaitTime_{it}$ and $Disposition_i$ are alternative-specific constants and thus only their differences are relevant. Hence, their interactions with triage levels are included rather than

[3] The maximal bed capacity is rarely observed since hospitals may temporarily increase their capacity in extreme situations (Armony et al. 2015, Berry Jaeker and Tucker 2017). We follow Kim et al. (2014) and test the 95th percentile in the robustness checks. The estimated hourly bed capacity is provided in Table 7 of Appendix A in the e-companion.

their main effects. Interested readers are referred to Section 2.5 of Train (2009) for an excellent discussion that colloquially summarizes this phenomena as "A rising tide raises all boats."

4.2. Measure of Anticipated Disposition and Endogeneity Issues

One challenge in our investigation is that the anticipated disposition of patient i by a decision maker at any decision epoch, i.e., $Disposition_i$, is not observable in our data. One could use the actual disposition as a proxy. However, the actual disposition is assigned after the patient is selected for assessment. Hence, using the actual disposition as a proxy raises the issue of reverse causality and thus estimation bias. We choose to use the predicted disposition of patient i—predicted with the information collected at triage only—as a proxy for $Disposition_i$. We estimate a logit model to predict a patient's disposition. With only six basic patient characteristics as predictors (age group, gender, address, arrival mode, triage level, and chief complaint), a standard logistic regression model performs reasonably well with an out-of-sample c-statistic of 0.783 (see other measures including recall, precision, etc., in Table 6 in Appendix A in the e-companion). The estimation results are presented in the first column of Table 2. We also estimate a probit model for the purpose of robustness check (see details in Section 4.4) and present the estimation results in the second column of Table 2.

The effectiveness of regressing the ex-post disposition outcomes against ex-ante triage information to solve the endogeneity issue is contingent on the assumption that disposition decisions made by physicians are independent of the ED blocking level at the decision epoch. Otherwise, the ED blocking level will become a confounding variable impacting both the predicted disposition and the choice outcome, which would undermine the causal relation this paper tries to draw.

To validate this assumption, we add the ED blocking level to the disposition prediction model. The estimation results for Measures 1 and 2 of *BlockLevel* are shown in columns 3 and 4 of Table 2, respectively. The impact of *BlockLevel* on patient disposition is statistically insignificant under both measures at the 5% level. Our result aligns with Chen et al. (2019), which also concludes that boarder census has no impact on patient disposition using data from a US hospital. Gorski et al. (2017) finds a positive correlation between waiting room census and admission probability. Hence, we include waiting room census in the disposition prediction model as a robustness check. The estimation results (shown in the last column of Table 2) confirm that the effect of *BlockLevel* is not statistically significant but *WaitRoomCensus* is negatively correlated with admission probability.

We have also used half of the observations, i.e., when the blocking level is below (or above) its median, to train the disposition prediction model and then apply it to predict the disposition for all patients. The predicted disposition enters the discrete-choice model in Eq. (2) as a proxy for the anticipated disposition by physicians. The estimation results for both cases are highly consistent, which provides further evidence that the ED blocking level does not correlate with a patient's disposition. Hence, the predicted disposition is a valid proxy. See details of the estimation results and a further discussion in Appendix A of the e-companion.

Table 2 Estimation results for disposition prediction models. The first two columns are the estimation results of the two models (a logit model and a probit model) used in the choice model to predict *Disposition*. The last three columns are the estimation results of models that study the impact of the ED blocking level on patient disposition decisions.

Prediction Model Measure of *BlockLevel*	Logit	Probit	Logit Measure 1	Logit Measure 2	Logit Measure 1
Intercept	-1.203^{***}	-0.713^{***}	-1.210^{***}	-1.211^{***}	-1.119^{***}
BlockLevel			0.028	0.028	0.007
WaitRoomCensus					-0.262^{***}
Triage Level (Base=Level 2)					
Level 3	-0.658^{***}	-0.385^{***}	-0.658^{***}	-0.658^{***}	-0.658^{***}
Level 4,5	-1.380^{***}	-0.789^{***}	-1.379^{***}	-1.379^{***}	-1.379^{***}
Age Group (Base=18-40 years)					
0-18 years	0.250^{**}	0.136^{**}	0.250^{**}	0.250^{**}	0.253^{***}
40-55 years	0.518^{***}	0.289^{***}	0.518^{***}	0.518^{***}	0.518^{***}
55-70 years	0.988^{***}	0.564^{***}	0.988^{***}	0.988^{***}	0.989^{***}
>70 years	1.424^{***}	0.828^{***}	1.424^{***}	1.424^{***}	1.425^{***}
Gender (Male=1)	0.269^{***}	0.157^{***}	0.269^{***}	0.269^{***}	0.269^{***}
Arrival Mode (Ambulance=1)	0.842^{***}	0.498^{***}	0.842^{***}	0.842^{***}	0.839^{***}
Address (Base=Others)					
Region A	-0.264^{***}	-0.154^{***}	-0.264^{***}	-0.264^{***}	-0.261^{***}
Region B	0.017	0.011	0.017	0.017	0.020
Region C	-0.056	-0.031	-0.056	-0.056	-0.052
Region D	0.091^{*}	0.054^{*}	0.090^{*}	0.090^{*}	0.094^{*}
Chief Complaint †					
Shortness of Breath	0.386^{***}	0.237^{***}	0.386^{***}	0.386^{***}	0.388^{***}
Chest Pain (Cardiac Features)	-1.039^{***}	-0.614^{***}	-1.038^{***}	-1.038^{***}	-1.034^{***}
Headache	-1.148^{***}	-0.642^{***}	-1.148^{***}	-1.148^{***}	-1.145^{***}
Depression / Suicidal	0.049	0.034	0.049	0.049	0.049
Vomiting And / Or Nausea	0.200^{***}	0.112^{***}	0.201^{***}	0.201^{***}	0.202^{***}
McFadden pseudo R^2	0.180	0.181	0.180	0.180	0.181
(Equivalent linear model R^2)	(0.402)	(0.404)	(0.402)	(0.402)	(0.404)

† Base = *Abdominal Pain*. The remaining 164 chief complaint codes are not shown for the sake of space.
***p<0.001; **p<0.01; *p<0.05

4.3. Results and Discussions

To study decision makers' patient prioritization behavior, we estimate Eq. (2) by maximizing the likelihood of choosing the observed choice of patient. We account for the potential heteroscedasticity of ε_{it} by using the Huber-White Sandwich estimator. The variable $Disposition_i$ is the predicted probability of patient i being admitted. Hence, we adjust the standard errors in the estimation of the discrete-choice model. Next, we discuss the estimation results of Model 1 shown in Table 3, which serves as the baseline model for our robustness checks. The model goodness-of-fit is measured by the McFadden R^2. The empirically equivalent linear model R^2 are estimated from Figure 5.5 in Domencich and McFadden (1975).

Observation 1. *Across different triage levels, decision makers (i) apply an urgency-specific delay-dependent prioritization rule when selecting the next patient for initial assessment; (ii) further prioritize high acuity patients over middle-to-low acuity patients when more patients are waiting to be seen.*

The estimation results show that all patient characteristics, including age group, gender, arrival mode,

Table 3 Key determinants of patient prioritization decisions. Model 1 serves as the baseline model. Models 2–5 deviate from Model 1 by using Measure 2 of ED blocking level, by using the disposition predicted by a probit model, by controlling the quadratic term of *WaitRoomCensus*, and by removing patients with triage orders, respectively.

	Model 1	Model 2	Model 3	Model 4	Model 5
Triage Level = 2					
CTAS×*WaitTime*	0.661***	0.663***	0.661***	0.66***	0.372***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.013)
CTAS×*WaitTime*^{2}	−0.049***	−0.049***	−0.049***	−0.048***	−0.023***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
CTAS×*Disposition*	0.435***	0.507***	0.43***	0.436***	0.762***
	(0.098)	(0.100)	(0.124)	(0.098)	(0.128)
CTAS×*Disposition*×*BlockLevel*	−1.583***	−1.601***	−1.63***	−1.584***	−1.497***
	(0.224)	(0.200)	(0.236)	(0.224)	(0.306)
Triage Level = 3					
CTAS	−0.302***	−0.278***	−0.299***	−0.251***	−0.443***
	(0.040)	(0.041)	(0.041)	(0.053)	(0.053)
CTAS×*WaitTime*	1.035***	1.036***	1.036***	1.036***	0.942***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.016)
CTAS×*WaitTime*^{2}	−0.077***	−0.077***	−0.077***	−0.077***	−0.079***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)
CTAS×*WaitRoomCensus*	−1.034***	−1.041***	−1.034***	−1.354***	−1.002***
	(0.060)	(0.060)	(0.060)	(0.229)	(0.083)
CTAS×*Disposition*	−0.141	−0.018	−0.163	−0.14	−0.236
	(0.115)	(0.116)	(0.143)	(0.115)	(0.153)
CTAS×*Disposition*×*BlockLevel*	−2.791***	−2.802***	−2.827***	−2.792***	−2.235***
	(0.266)	(0.237)	(0.267)	(0.266)	(0.363)
Triage Level = 4, 5					
CTAS	−0.603***	−0.579***	−0.6***	−0.543***	−0.786***
	(0.047)	(0.048)	(0.049)	(0.064)	(0.059)
CTAS×*WaitTime*	1.088***	1.087***	1.089***	1.09***	1.042***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.021)
CTAS×*WaitTime*^{2}	−0.085***	−0.084***	−0.085***	−0.085***	−0.089***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)
CTAS×*WaitRoomCensus*	−0.577***	−0.585***	−0.577***	−0.952***	−0.65***
	(0.077)	(0.077)	(0.077)	(0.285)	(0.098)
CTAS×*Disposition*	−0.235	−0.012	−0.27	−0.236	−0.585*
	(0.214)	(0.218)	(0.234)	(0.214)	(0.287)
CTAS×*Disposition*×*BlockLevel*	−3.419***	−3.725***	−3.24***	−3.414***	−2.014*
	(0.653)	(0.583)	(0.622)	(0.653)	(0.831)
Observations	83189	83189	83189	83189	50885
McFadden pseudo R^2	0.069	0.069	0.069	0.069	0.067
(Equivalent linear model R^2)	(0.151)	(0.151)	(0.151)	(0.151)	(0.147)

Notes. This table reports the estimation results from the discrete-choice model. Robust standard errors are shown in the parentheses. ***p<0.001; **p<0.01; *p<0.05

and chief complaint code, are factors in decision makers' prioritization decisions. They are not included in Table 3 to save space. Interested readers are referred to Table 11 in Appendix A of the e-companion for the complete estimation results. The coefficients of *CTAS* for different triage levels in Model 1 imply that among patients with the same characteristics, serving a patient of higher urgency (smaller triage level) generates higher utility for decision makers. More precisely, selecting a level 2 patient results in 0.302 higher utility

than selecting a level 3 patient, and 0.603 higher utility than selecting a level 4 or 5 patient, when patient characteristics are the same and the effects of other factors are negligible. This observation is consistent with the principle of triage, i.e., classify and prioritize patients based on their urgency. By the CTAS protocol adopted in our study ED, patients classified as triage level 2 are more urgent than patients of level 3, who in turn are more urgent than patients of levels 4 and 5 (Beveridge et al. 1998). Hence, selecting a level 2 patient incurs higher utility than selecting patients of triage levels 3,4, or 5.

The interaction terms $CTAS \times WaitTime$ and $CTAS \times WaitTime^2$ are significant for all triage levels. Their coefficients suggest that the utility is concave down in *WaitTime*. Interestingly, the utility for all triage levels reaches its maximum when $WaitTime \approx 6.5$ hours, which is the 99th percentile of the waiting times in our dataset. Hence, the utility is increasing and concave in *WaitTime*, i.e., selecting a patient with longer waiting time always generates higher utility. However, the increment per unit of time slows down as patients wait longer. This observation explains why a less urgent patient could be selected by decision makers for treatment when there are urgent patients waiting. In other words, EDs do not operate like a multi-class queueing system where triage levels indicate strict priority. Rather, the accumulating priority queues studied in Li et al. (2017) are potentially more appropriate models. From a clinical perspective, patients who have waited longer face a higher risk of adverse outcomes (Sun et al. 2013). Hence, it is a rational decision to prioritize them. We also note that the increase in utility per unit of waiting time varies across triage levels: less urgent patients have greater marginal utility increase in waiting time. In summary, our results imply that decision makers apply an urgency-specific delay-dependent prioritization rule, which is consistent with the literature (Ferrand et al. 2018, Ding et al. 2019).

The coefficients of $CTAS \times WaitRoomCensus$ are significant and negative for middle-to-low acuity patients, i.e., triage levels 3, 4, and 5, which implies that when all other factors are the same, high acuity patients (triage level 2) are more likely to be selected when there are more patients in the waiting room. The intuition is that when the waiting room becomes more crowded, it is critical to attend to high acuity patients first, as their conditions are urgent and more likely to deteriorate, whereas middle-to-low acuity patients can wait longer with *relatively lower risk*. This aligns with the insight in Sun et al. (2018) that prioritization creates greater value when more patients are waiting to be served.

Observation 2. *Within the same triage level, decision makers (i) prioritize admit patients over discharge patients for high acuity patients, and do not consider disposition in the prioritization of middle-to-low acuity patients when the ED blocking level is sufficiently low; (ii) prioritize discharge patients when the ED blocking level is sufficiently high.*

Observation 1 explains how patients are prioritized across triage levels. However, it remains unclear why patients of the same urgency (i.e., the same triage level) are not seen in an FCFS manner. In particular, why and how does a patient's disposition affect her priority? The two interaction terms, $CTAS \times Disposition$ and

$CTAS \times Disposition \times BlockLevel$, help answer this question. From the estimates of Model 1 in Table 3, we make the following observations. When the ED blocking level is sufficiently low, only $CTAS \times Disposition$ is relevant. The coefficient of $CTAS \times Disposition$ for triage level 2 is significant, suggesting that selecting an admit patient generates 0.435 higher utility than selecting a discharge patient of the same triage level when all other factors are the same. Hence, admit patients are prioritized within triage level 2. The coefficients are not statistically significant for triage levels 3, 4, and 5, suggesting that their disposition does not affect their order of being seen.

The coefficients of $CTAS \times Disposition \times BlockLevel$ are negative for all triage levels, implying that as the ED blocking level increases, the utility of choosing admit patients decreases for all triage levels, relative to the utility of choosing discharge patients. This suggests that within the same triage level, a discharge patient may be seen earlier than an admit patient with similar characteristics and waiting times when the blocking level is sufficiently high. We note that the coefficient of $CTAS \times Disposition \times BlockLevel$ for triage levels 4 and 5 (-3.419) is larger in magnitude than that for triage level 3 (-2.791), which is significantly larger than that of triage level 2 (-1.583), suggesting that the ED blocking level has a greater impact on middle-to-low acuity patients.

In summary, our empirical investigation shows that choosing the next patient for treatment is a complex decision that depends on both clinical and operational factors. When the ED blocking level is sufficiently low, clinical factors such as patients' triage levels and waiting times dominate. When the ED blocking level increases, the operational factor kicks in, and the chance that discharge patients get prioritized increases. Our empirical results provide evidence that ED decision makers factor resource constraints in their patient prioritization decisions, contrary to the common perception that patient priority assignment is a clinical decision made during triage. In addition, physicians' behavior may vary due to the lack of operational guidelines. See Appendix B in the e-companion for more results and discussions.

4.4. Robustness Checks

To show the robustness of our findings, we examine several model specifications that deviate from the baseline model. In Model 1, the baseline model, the disposition of a patient is predicted by a logit model using six basic patient characteristics. As a robustness check, we replace the logit model with a probit model to ensure that our findings do not depend on any specific classifier.

The measure of ED blocking is critical to our study. We have constructed two different measures in Section 3.3.2, namely, Measure 1 and Measure 2. We propose a third measure, referred to as *Measure 3*, which deviates from Measure 1 by using the 95th percentile of the observed number of patients in ED beds as a proxy for the maximal ED bed capacity. We also deviate from Measure 1 by setting the capacity of an individual physician to 6.5; we refer to this measure as *Measure 4*. Inspired by Berry Jaeker and Tucker (2017), we propose a fifth measure, which deviates from Measure 2 by using the monthly maximal boarder

census for the normalization, rather than using the maximum over the 2-year study period. This measure is referred to as *Measure 5*. The summary statistics of the five measures are shown in Table 5 in Appendix A of the e-companion. We use all five measures to test the robustness of our results.

ED nurses may request tests for patients whose conditions satisfy pre-set protocols at triage. This practice is called *triage nurse ordering*, and its impact has been studied in the literature (e.g., Batt and Terwiesch 2016). In our study ED, nurses are allowed to order lab tests but not imaging tests. This may affect decision makers' behavior in choosing the next patient, since a decision maker might prioritize a patient whose test results are ready, or defer the assessment of a patient awaiting test results. In our dataset, we cannot observe whether a patient's test results are available when the patient is selected for assessment. Hence, we remove all visit records with triage orders and repeat our study to check the robustness of our results.

We also deviate from Model 1 and control the nonlinear effects of *WaitRoomCensus* by adding its quadratic term to the model. As a result, we estimated a total of 10 models, and their complete estimation results are provided in Tables 11 and 12 in Appendix A in the e-companion. Models 2–5 in Table 3 deviate from Model 1 by using Measure 2 for the ED blocking level, by using the disposition predicted by a probit model, by controlling the quadratic term of *WaitRoomCensus*, and by removing patients with triage orders, respectively.

We found that *the results of all the 10 models are qualitatively similar except for Model 5*. Specifically, the coefficient of $CTAS \times Disposition$ for low-acuity patients (triage levels 4 and 5) in Model 5 becomes significant at the 5% level, which might stem from the bias created by removing patients with triage orders since triage orders only apply to patients whose conditions fit certain criteria. Moreover, the early testing may reduce patients' acuity levels and allows them to wait longer. Physicians may also prioritize or hold up a patient depending on the availability of the test results. In addition, the term $CTAS \times Disposition \times BlockLevel$ for low-acuity patients becomes less significant than in the other models, which may be attributed to the smaller sample size (about 39% of the observations are removed).

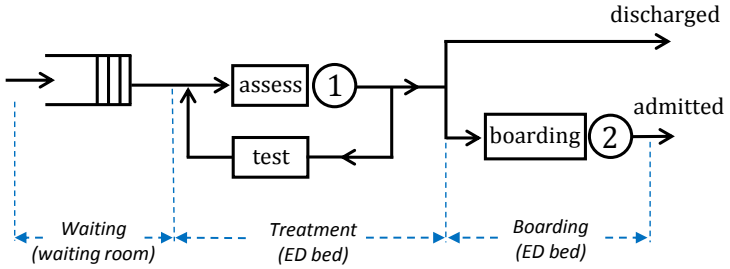
5. Rationale Behind Prioritization Behavior

In Section 4, using patient visit record data, we examine ED decision makers' behavior on patient prioritization and find that discharge patients are prioritized when the ED blocking level is high. In this section, we build a stylized MDP model to further understand the rationale behind such behavior. The aim is to understand how disposition affects a patient's priority under ED blocking, especially for patients of the same triage level, rather than explicitly modeling the complex ED operations in detail. Hence, we believe that this stylized model can capture the key characteristics of interest.

5.1. Model for ED Patient Flow and ED Blocking

We model the ED patient flow process during a high-load period using a two-station tandem queue with feedback at Station 1. A schematic depiction of patient flow is given in Figure 3. After triage, patients

Figure 3 Patient flow in emergency departments.



Note. The ED patient flow process is modeled as a two-station tandem queue with feedback at Station 1 (marked 1 and circled), which represents the physician assessment in the ED treatment phase. Station 2 (marked 2 and circled) represents the boarding process.

wait in the waiting room until chosen for service at Station 1, where multiple physicians work to provide emergency care. After physician assessment, patients join the test queue before returning to the assess queue with a constant probability (Yom-Tov and Mandelbaum 2014); if the patients do not join the test queue, the treatment process is completed, and then patients are either discharged home with probability $1-p$ or admitted with probability p $(0<p<1)$, in which case they will join the boarding queue waiting for transfer to inpatient beds. The service time at Station 2 corresponds to the boarding time of a patient. For the sake of tractability, we assume that the boarding time follows an exponential distribution with a constant rate μ_2, while being aware that the boarding process is highly time-dependent in reality (Shi et al. 2015). This will be relaxed in our simulation study in the next section. We assume that there are always patients waiting to be seen in the waiting room since we focus on decisions during a high-load period. We also assume that patients are only different in their dispositions and are identical in other characteristics including demographics, waiting time, etc., as this model focuses on how dispositions affect the order of patients being seen.

It is well known that ED physicians are multitasking. However, they generally do not take on more patients than their total service capacity (Campello et al. 2016) as this not only raises safety concerns but also can be counter-productive (KC 2013). Let C denote the total physician capacity of all of the physicians on duty. Saghafian et al. (2012) observe that an individual physician's capacity is generally no more than seven in their study ED. KC (2013) further suggests that the optimal capacity is five patients for an average physician to maximize throughput. Let B be the ED bed capacity, and x be the number of boarding patients. Since boarding patients need to stay in ED beds, we have $0 \le x \le B$. Because physicians are responsible for the care of all unfinished patients in the ED, patients in both the assess and test queues count as physicians' workload. Let K denote the total number of patients in the assess and test queues. We have $K \le \min\{C, B-x\}$, which implies that both physicians and beds can be the bottleneck resource in an ED. ED blocking occurs when $K = B-x < C$, i.e., lack of ED beds starves the physician resource.

We further assume that there is no unforced idling for physicians at Station 1. Then, whenever a patient's treatment in the ED is complete, the patient will leave, and a new patient will be picked from the waiting

room. Hence, we have $K = \min\{C, B - x\}$. We model the patient flow dynamics in the treatment phase, i.e., the interaction between the assess and test queues, as a closed network with population K—the total number of patients in the treatment phase. There are multiple parallel servers providing services at both queues. We assume that servers in the assess queue are pooled in that any of the physicians can serve any patient. This deviates from normal practice in our study ED, as patients usually return to the same physician for assessment after testing, so as to reduce *handoff* errors (Batt et al. 2019). We made this assumption to simplify our model and analysis. Let $\mu_1(\cdot)$ denote the throughput rate from this closed network, and it represents the rate at which patients complete their diagnosis and treatment in the ED. (Note that $\mu_1(\cdot)$ is *not* the rate at which physicians assess patients.) We have the following result.

PROPOSITION 1. *Assume that (i) there are multiple parallel servers in both the assess and test queues; (ii) the assess times and testing times are respectively i.i.d. exponential random variables and are independent of each other. Then, the throughput rate $\mu_1(K)$ is increasing and concave in K. Specifically, we have $\mu_1(0) = 0$.*

Proposition 1 implies that the rate at which a patient completes ED treatment and exits the ED increases with the number of patients whose diagnoses and treatments are in progress. This is the outcome of server pooling, i.e., the more patients in the cyclic network (the treatment phase), the less likely the servers will be idle. However, the marginal increment decreases as the internal delays become longer as physicians and test centers have more patients to serve, and the reduction in server idleness decreases with more patients in the treatment phase.

5.2. The MDP Formulation and Results

Next, we model the decision problem on prioritizing admit or discharge patients by an MDP formulation. For the sake of tractability, we aggregate the treatment phase, i.e., the assess/test interactions in Figure 3, into a single station whose service represents the diagnosis and treatment process at ED. We assume the service times at this station are exponentially distributed with rate $\mu_1(K)$. This approximation through state aggregation shares similarity with the $\mathcal{T}$ approximation in Campello et al. (2016). Since $K = \min\{C, B - x\}$, it is straightforward to show that $\mu_1(K)$ is decreasing and concave in x. To emphasize the dependence on x, we rewrite $\mu_1(K)$ as $\mu_1(x)$ in the rest of the paper. By Proposition 1, we have $\mu_1(B) = 0$, i.e., if all beds are occupied by boarding patients, then the rate of treating non-boarding patients drops to zero.

The decision epochs correspond to the times that physicians become available to serve a new patient. Denote the system state at time t by x, representing the number of boarding patients in ED at t. Hence, the state space is $\mathcal{S} = \{0, 1, 2, ..., B\}$. Whenever a physician becomes available to see a new patient, she can choose one from the waiting room based on the patient disposition, or she can choose a patient randomly; e.g., simply choose the first one in line. Hence, the action space is $\mathcal{A} = \{\textit{Choose Discharge}, \textit{Choose Admit}, \textit{Choose First in Line}\}$.

Assume that serving a discharge (admit) patient generates a utility of R_1 (R_2) for the decision maker. The utility of decision makers can be interpreted as the social benefit gained by serving a patient. The social benefit is greater when taking care of a more urgent patient. We assume that an admit patient's condition is no less urgent than a discharge patient, i.e., $R_2 \geq R_1 > 0$. We also assume that there is a negative utility associated with the action of selecting a patient: $-c_1$ ($-c_2$) for discharge (admit) patients where $c_i \geq 0,\ i = 1, 2$. The negative utility can be interpreted as the extra effort it takes to search for a specific type of patient from the dashboard of the hospital's electronic patient track system. The negative utility also corresponds to a cost for social injustice. Note that there is no searching/fairness cost for serving the first patient in line. To avoid triviality we also assume that $R_i - c_i > 0,\ i = 1, 2$. The decision maker's objective is to find a control policy to maximize the expected long-run average net social benefits over an infinite time horizon.

We next let $g(\pi, x) \equiv \liminf_{t\to\infty} V_t(\pi, x)/t,\ \forall x \in \mathcal{S}$, be the expected long-run average net social benefits, where $V_t(\pi, x)$ is the total expected net social benefits up to time t starting from state x under policy π. Then, the optimal long-run average net social benefits is $g^*(x) \equiv \sup_\pi g(\pi, x),\ \forall x \in \mathcal{S}$. We apply *uniformization* with the uniformization constant $\Lambda \equiv \mu_1(0) + \mu_2$ (Lippman 1975). Without loss of generality, we can redefine the time unit so that $\Lambda = 1$, and then $\mu_1(x)$ and μ_2 become, respectively, the probability that the next uniformized transition is a service completion at Stations 1 and 2. Let $v(x)$ be the bias function defined as the difference between the total expected net social benefits starting from state x and a reference state. The long-run average net social benefits optimality equations can be written as $v(x) + g = Tv(x),\ \forall x \in \mathcal{S}$, where g is the optimal average net social benefits per period of time after uniformization, and the format of the operator T is the same as the one defined in Appendix C. Next, we present Proposition 2, which establishes the existence and structural properties of the optimal policy.

Proposition 2. *Let x be the number of boarding patients. There exists an optimal stationary deterministic policy that takes the following form: choose admit patients if $x < x_a$ and choose discharge patients if $x > x_d$; otherwise, choose the first patient in line, where $x_a \equiv \max\{x : x \in \mathcal{S},\ v(x) - v(x+1) \leq R_2 - R_1 - (1-p)^{-1}c_2\}$ and $x_d \equiv \min\{x : x \in \mathcal{S},\ v(x) - v(x+1) \geq R_2 - R_1 + p^{-1}c_1\}$.*

Proposition 2 states that the optimal policy is of threshold-type, characterized by x_a and x_d. Note that x/B is the percentage of ED beds occupied by boarding patients, which is an indicator of the ED blocking level. Our result can be interpreted as follows: When the ED blocking level is relatively low ($x < x_a$), it is optimal to prioritize admit patients, i.e., clinical factor dominates when treatment capacity is not of concern; when the blocking level is high ($x > x_d$), it is optimal to prioritize discharge patients, i.e., operational factor dominates when the pressure on bed resource is high; when the blocking level is intermediate, it is optimal to follow FCFS. The proof is provided in Appendix C in the e-companion.

5.3. Connection to Empirical Findings

Next, we discuss the connection between Proposition 2 and our empirical findings for each triage level, and explain the mechanism behind ED decision makers' prioritization behavior.

The medical conditions for patients of triage level 2 are generally urgent. For example, the five most frequent complaint codes for triage level 2 patients are *Chest Pain (Cardiac Features), Abdominal Pain, Major Trauma (Blunt), Cardiac Type Pain*, and *Shortness of Breath.* Their conditions can deteriorate quickly. Among them, we believe that admit patients are more urgent than discharge patients ($R_2 > R_1$). Hence, serving admit patients generates higher social benefits than serving discharge patients, and admit patients are prioritized when ED beds are not the bottleneck resource ($x < x_a$). Patients of triage levels 3, 4, and 5 are less urgent and can wait some time without significantly compromising their care. In particular, admit and discharge patients are not significantly different with respect to their urgency, i.e., $R_1 \approx R_2$, and thus $x_a \leq 0$. Hence, within triage level 3 or levels 4–5, patients' order of being seen is independent of their disposition when the ED blocking level is sufficiently low ($x < x_d$).

When the ED blocking level is sufficiently high, e.g., when over $100x_d/B$ percent of the ED beds are occupied by boarding patients, it may become difficult for decision makers to find available beds to treat new patients. Thus, it is wise to take the bed capacity into consideration when selecting the next patient to see. Intuitively, it is better to start prioritizing discharge patients over admit patients. Otherwise, another bed may be occupied for a prolonged period, which only further reduces ED treatment capacity and aggravates ED blocking. We observe this behavior in the empirical results for all triage levels. The reduction in treatment capacity is reflected in our model by that $\mu_1(x)$ decreases with x. More importantly, our MDP model and the results derived from it explicitly explain the trade-off faced by ED decision makers in patient prioritization, i.e., gaining a greater social benefit in the short term (prioritizing admit patients) versus preserving a higher rate of gaining social benefits in the long run (prioritizing discharge patients).

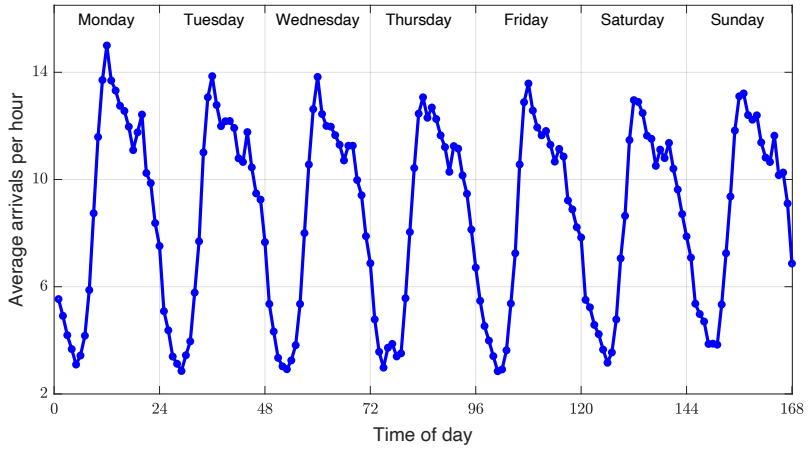
6. Improving ED Operations Through Patient Prioritization

In previous sections, we examine ED decision markers' prioritization behavior and the rationale behind it. However, the impact of such behavior on ED operational performance remains unclear. More importantly, how can we leverage these insights into patient prioritization to improve ED operations? To the best of our knowledge, there are no guidelines for patient prioritization. As a result, we expect decision makers to exhibit heterogeneous behavior (see Appendix B in the e-companion), which makes it difficult to examine the impact through observational data. In this section, we develop a discrete-event simulation model to compare prioritization policies that consider ED blocking with a policy that does not, and quantify the impact on average waiting time and length of stay (LOS). The simulation model assumes that decision makers follow an explicit prioritization rule, which deviates from reality. However, we believe that the results can demonstrate the benefits of priority rules that take into account the availability of ED resources, at least qualitatively.

6.1. Simulation Design

We simulate the ED patient flow process as a two-station tandem queue, as described in Section 5.1. We relax the assumption on the arrival process, and assume that patients arrive at the ED according to a non-stationary Poisson process, which has been shown to be a reasonable assumption (Kim and Whitt 2014). Instead of considering all five triage levels, we aggregate patients into two classes. Class 1 corresponds to patients of CTAS 1 (and part of CTAS2) who have the highest priority and almost always receive treatment immediately. Class 2 corresponds to less-urgent patients (CTAS 2–5) who need to wait for treatment if all physicians are busy. Two types of resources are required to treat patients: ED beds and physicians. A patient cannot be treated unless both resources have available capacity. After treatment, a patient may leave the system (be discharged), or be admitted and join another queue waiting for an inpatient bed (boarding) while occupying an ED bed. We assume that decision makers know the dispositions of patients waiting for treatment and may use such information for patient prioritization.

Figure 4 Hourly patient arrivals to our study ED by time of day and day of the week (data from August 2014 to July 2015).



The input analysis of the simulation is based on data collected between August 2014 and July 2015. Our descriptive statistical analysis finds that the hourly arrival rate depends on both the time of day and the day of the week (see Figure 4). Mondays are especially busy, a phenomenon that is referred to as *Monday effect* in emergency medicine. We also observe that the shift schedules of medical staff repeat every 24 hours. Hence, we choose one week as a cycle and treat the data of each week from August 2014 to July 2015 as a realization of the stochastic process that governs the true arrival process. To generate patient arrivals to the ED, we first calculate the inter-arrival times, dependent on the time of day and the day of the week, as the input parameters for a non-stationary Poisson process. Then, once the event of generating a new patient is triggered, we randomly sample a patient from the set of patients in the data who arrived at the ED during that hour on the particular day of the week (bootstrap), and assign this patient's profile (triage level,

disposition, etc.) to the newly generated patient. We generate service times (e.g., treatment/boarding times) for this patient in a similar manner: first, we group the treatment times and boarding times from the data by time of day and day of the week, triage level, and disposition; then, we randomly sample service times from the set that corresponds to the new patient's profile.

The number of physicians on duty at any time of the day is known. We assume that physicians are multitasking and have the same service capacity. We set the capacity to seven patients. It is however challenging to estimate the capacity of ED beds, as discussed in Section 3.3.2. We follow our empirical study setting and use the 90th percentile of the number of patients staying in ED beds during each hour of the day over the 2-year horizon to approximate the bed capacity at this particular hour of the day. Using the 90th percentile instead of the maximum avoids outliers created by data collection errors or temporary increases in ED capacity. Since both physician capacity and bed capacity are time-varying, we assume that an exhaustive discipline (Ingolfsson et al. 2007) is applied whenever the capacity decreases.

6.2. Prioritization Policies

Next, we use a simulation to compare patient prioritization policies that do or do not consider ED blocking. In the simulation, we use Measure 1 for the ED blocking level (see Section 3.3.2), i.e., the ratio of the number of boarding patients over the number of "extra" beds (total bed capacity net of total physician capacity). We start by defining the policies of interest.

Urgency-Based Priority Policy (UP): A non-idling policy, under which Class 1 patients receive priority over all other patients; FCFS is followed within each class.

Congestion-Urgency-Based Priority Policy (CUP): A non-idling policy, under which Class 1 patients receive priority over all other patients; FCFS is followed within Class 1. For Class 2, prioritize admit patients if blocking level is less than η_1 (≥ 0), prioritize discharge patients if the ED blocking level is greater than η_2 ($\geq \eta_1$), and follow FCFS otherwise.

CUP-Prioritize-Admit Policy (CUP-A): The same policy as Policy CUP except that within Class 2, discharge patients are never prioritized. Rather, prioritize admit patients if the ED blocking level is less than η_A (≥ 0) and follow FCFS otherwise, i.e., $\eta_1 = \eta_A,\ \eta_2 = \infty$.

CUP-Prioritize-Discharge Policy (CUP-D): The same policy as Policy CUP except that within Class 2 admit patients are never prioritized. Rather, prioritize discharge patients if the ED blocking level is greater than η_D (≥ 0) and follow FCFS otherwise, i.e., $\eta_1 = 0,\ \eta_2 = \eta_D$.

CUP-Prioritize-Admit-Discharge Policy (CUP-AD): The same policy as Policy CUP except that within Class 2, prioritize discharge patients if the ED blocking level is greater than η_{AD} and prioritize admit patients otherwise, i.e., $\eta_1 = \eta_2 = \eta_{AD}$.

Under Policy UP, patients are seen mainly based on their clinical urgency. Patients of the same urgency are seen in their order of arrival, so that fairness is exercised. Policy CUP is inspired by the optimal policy

in Section 5, which either prioritizes patients based on their dispositions or follows FCFS depending on ED blocking level. We compare UP and CUP by varying the two thresholds η_1 and η_2 in a wide range. In addition, we consider three policies that are special forms of Policy CUP, namely, CUP-A, CUP-D and CUP-AD. We note that Policy CUP-AD is similar to the prioritization of CTAS 2 patients in our empirical study, and Policy CUP-D aligns with the one used for CTAS 3, 4, and 5 patients. All three policies are simpler and easier to implement than Policy CUP. We compare them with Policy UP by varying their corresponding parameters over a broad range and quantify their impact on patient average waiting time and LOS.

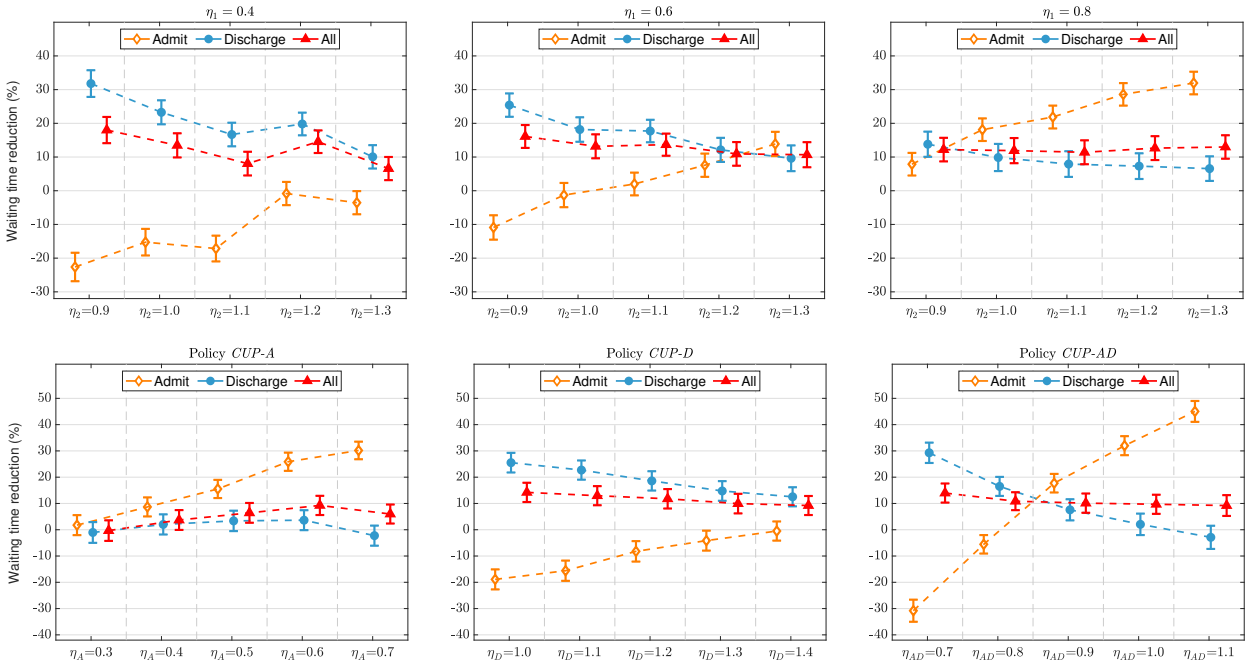
6.3. Comparison Results and Discussions

To make the simulation statistically efficient, we apply variance reduction techniques by using common random numbers when creating patient arrival processes under different policies and perform the output analysis using the replicated batch means method (Law and Kelton 2000, Argon and Andradóttir 2006). Specifically, starting from an empty ED, we simulate the ED operations under each policy for 10 replications with replication length 907 days. For each replication, we identify the first 7 days as the warm-up period by Welch's method (Law and Kelton 2000), thus they are removed from the output. We then take 7 out of every 9 days of the remaining 900 days as a "batch" and calculate the average waiting time and LOS using all of the patients who arrive and leave during the "batch" period. (The 2 days in every 9 days are removed to reduce auto-correlation among batches of the same replication.) Hence, we have 1,000 pairs of average waiting time and LOS for each policy.

The three plots in the top row of Figure 5 show the 95% confidence interval for the percentage reduction in the average waiting time by using CUP over UP for 15 combinations of η_1 and η_2, where $\eta_1 \in \{0.4, 0.6, 0.8\}$ and $\eta_2 \in \{0.9, 1.0, 1.1, 1.2, 1.3\}$. Our first observation is that prioritizing patients based on disposition and ED blocking level can reduce the average waiting time of all patients as much as 15%. It is intuitive that the more we prioritize discharge patients (smaller η_2), the greater the reduction in average waiting time for discharge patients. However, it comes at the cost of longer waiting for admit patients. As can be seen, the average waiting time of admit patients increases more (negative reduction) when a larger percentage of discharge patients are prioritized. Moreover, we observe that discharge patients have a greater impact on overall performance than admit patients as discharge patients are four times as many as admit patients.

There are scenarios in which *the average waiting times for patients of both dispositions decrease*, e.g., $(\eta_1, \eta_2) = (0.6, 1.3)$ and $\eta_1 = 0.8$ for all η_2, which might seem counter-intuitive at first glance. However, this is entirely possible if critical ED resources are rationed appropriately in a highly time-varying supply and demand environment, in which both physicians and beds can be bottlenecks for patient flow. When physicians are the bottleneck, i.e., there are empty beds and hence the ED blocking level is relatively low, prioritizing admit patients starts boarding earlier and increases bed utilization. When beds are the bottleneck, it is wise to prioritize discharge patients; otherwise, treating admit patients when few beds are available aggravates the

Figure 5 The 95% confidence interval for the percentage reduction in long-run average waiting time by using Policy *CUP* over Policy *UP* (the three figures in the top row) and by using Policy *CUP-A, CUP-D, CUP-AD* over Policy *UP* (the three figures in the bottom row), respectively.



level of ED blocking and impairs ED treatment capacity. Hence, a careful allocation of critical ED resources through patient prioritization can benefit all patients.

A closer look at the case $(\eta_1, \eta_2) = (0.6, 1.3)$ finds that the average waiting time for all patients decrease by 10%, which is about 10 minutes (the average is 106 minutes without prioritization). Woodworth and Holmes (2020) conclude that prolonging a patient's waiting time by 10 minutes increases the cost of care per ED visit by an average of 3% to 6% for moderately and most severe patients, respectively. Dawson and Zinck (2009) find that the average hospital cost (*not* including the costs of non-salaried physicians and lab tests) per ED visit in 2005–2006 was CA$148 in Ontario, Canada. With 150,000 ED visits per year, our calculation shows that the cost savings by using Policy CUP over UP range from CA$666,000 to CA$1,332,000. These numbers could easily be doubled if we add physician/test costs and inflation into the calculation. Note that this is only a rough estimate as the results in Woodworth and Holmes (2020) are based on a US ED and Dawson and Zinck (2009) use data from a different Canadian province more than 10 years ago. However, we believe that these numbers provide some insights into the value of patient prioritization.

The results of comparisons between Policy UP and the three simpler priority policies are shown in the bottom row of Figure 5, where $\eta_A \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, $\eta_D \in \{1.0, 1.1, 1.2, 1.3, 1.4\}$, and $\eta_{AD} \in \{0.7, 0.8, 0.9, 1.0, 1.1\}$. The results are similar to that of Policy CUP; however, the improvements of all three policies are less. We also note that Policy CUP-AD outperforms Policies CUP-A or CUP-D. That is, greater

improvement can be achieved if management is willing to deviate further from FCFS by prioritizing more patients. We observe similar patterns in LOS for all policies but with a smaller magnitude in reduction, as prioritization affects waiting time but not the treatment time. See more details in Figure 6 in the e-companion.

7. Impact on Waiting Time Prediction

In this section, we demonstrate how to incorporate decision makers' patient prioritization behavior to improve the accuracy of ED waiting time prediction. A growing number of hospitals (our study hospital included) have started posting predicted ED waiting time on their websites, smartphone apps, and screens in hospitals. An accurate prediction of waiting times can reduce patient waiting through a better coordination in hospital networks (Dong et al. 2019). Without such information, patients might be more prone to leave without being seen because they may wrongly infer waiting time (Batt and Terwiesch 2015). Therefore, providing an accurate waiting time prediction has attracted attention from both medical and operations communities. Nevertheless, the accuracy of prediction models has raised concern. The American College of Emergency Physicians warned that existing prediction methods provide misleading results and called for improving the prediction accuracy (ACEP 2012).

Two recent studies on waiting time prediction are Sun et al. (2012) and Ang et al. (2015). The first work developed a quantile regression model (QR) and its predictors include the number of patients of each triage level waiting to be seen by physicians, the number of patients of each triage level whose treatment started within the past hour, triage level, time of day, and day of the week. Ang et al. (2015) developed a Q-Lasso model based on a combination of queueing and statistical learning estimators. In addition to the predictors used in Sun et al. (2012), the Q-Lasso includes other variables such as local weather information, flu trend, the number of providers, and the number of nurses. We apply the two methods on our dataset and focus on low-acuity patients (triage levels 3, 4, and 5) following Ang et al. (2015). Some predictors used in Ang et al. (2015) that are not available in our dataset are omitted. Note that we have tried several other off-the-shelf prediction models, including neural network, XGBoost, etc. However, *none produces significant improvement over Q-Lasso in terms of reducing the root mean square error (RMSE)*. We then add the last three interaction terms in Eq. (2) into the models in Sun et al. (2012) and Ang et al. (2015), and refer to the new prediction models as *Priority QR* and *Priority Q-Lasso*, respectively.

We apply the four prediction models to our data. The predictor *BlockLevel* is by Measure 2, and *Disposition* is predicted by the logit model in the first column of Table 2. We choose 80% of our data for model training and the remaining 20% for testing. The absolute values and the percentage reductions of the RMSE are shown in Table 4. When we explicitly account for decision makers' prioritization behavior in the prediction models, the reductions in RMSE are statistically significant for both the QR and Q-Lasso (1.73% and 1.66%, respectively). We also test the performance on four subsets of the data, which contain patients who arrive and find the ED blocking level below the 50th percentile, and above the 50th, 75th, and 90th percentiles.

Table 4 RMSE (root mean square error) for waiting time prediction algorithms. Lower RMSE values are better.

	All patients	*BlockLevel* <50th percentile	*BlockLevel* >50th percentile	*BlockLevel* >75th percentile	*BlockLevel* >90th percentile
QR	81.95	83.66	83.04	79.66	84.56
Priority QR	80.55	82.66	81.44	77.21	80.64
Reduction in RMSE (%)	1.73	1.21	1.97	3.17	4.86
Q-Lasso	80.45	82.30	81.31	76.29	80.09
Priority Q-Lasso	79.12	81.55	79.77	73.35	75.67
Reduction in RMSE (%)	1.66	0.92	1.89	3.85	5.52

The results show that when the ED blocking level is relatively low (*BlockLevel* < 50th percentile), prediction accuracy is slightly improved by incorporating patient prioritization into QR or Q-Lasso. When the blocking level is high, the RMSE can be reduced by up to 4.86% and 5.52% for QR and Q-Lasso, respectively. We would expect greater improvement if decision makers behaved homogeneously. Note that our proposed predictions rely on the estimation of a patient's disposition. Such predictions are practically useful for (i) routing ambulances within a hospital network; (ii) providing waiting time estimate to patients at the end of their triage.

8. A Discussion on the SPT Rule

In this paper, we examine decision makers' patient prioritization behavior and find that decision makers may prioritize discharge patients over admit patients in response to ED blocking. However, the treatment variable *Disposition* also acts as an effective indicator of a patient's service duration. This motivates an alternative explanation: discharge patients are prioritized because decision makers are following the SPT rule, which has been observed in other healthcare settings (Ibanez et al. 2018). Under the SPT rule, the job with the shortest (expected) processing time of all jobs currently in the wait line should be prioritized so that the average waiting time is minimized (Section 7.10 of Cachon and Terwiesch 2008). The SPT rule seems to be a compelling alternative to our proposed mechanism. After all, the average time that admit patients stay in ED beds is about three times longer than that of discharge patients for patients of all acuity levels (see Table 1). We argue that SPT is relevant but may not serve as a substitute for our mechanism. We elaborate below.

The objective of SPT is to minimize the average waiting time. Hence, both long patient waiting time (*WaitTime*) and large number of patients waiting to be seen (*WaitRoomCensus*) could trigger changes in decision makers' prioritization behavior. In contrast, our mechanism states that the primary motivation for prioritizing discharge patients is to manage high levels of ED blocking. Hence, *BlockLevel* drives the changes in patient prioritization decisions. In other words, the SPT rule focuses on the waiting room census, whereas our mechanism looks at the back room census, more specifically, boarding patients. In the empirical models, we control both *WaitTime* and *WaitRoomCensus* by including their interactions with triage levels (see Section 4.1 for more details). The estimation results strongly support our mechanism as the variables of interest are statistically significant.

It is interesting that the Pearson correlation coefficients between *WaitRoomCensus* and any of our five measures for *BlockLevel* are negative with magnitudes smaller than 0.1 (see Table 9 in Appendix A in the e-companion), suggesting that a high blocking level does not imply a concurrent congested waiting room. This may be due to some time lag effect, in which case, one could argue that ED decision makers apply the SPT rule in anticipation of a crowded waiting room in the near future caused by the high blocking level in the present. Then, the SPT rule is a valid alternative for our mechanism, in which the processing time refers to the length of time a patient stays in an ED bed. However, from a modeling perspective, our mechanism (and the MDP inspired by it) explains the interactions between the two resources that determine the ED's capacity in treating patients, i.e., physicians and beds. As a result, our mechanism provides further insights. More specifically, when physician is the bottleneck resource and there are sufficient beds, operationally it is better to prioritize admit patients to start boarding earlier and increases bed utilization. Our simulation results confirm this insight by showing that policies that prioritize admit (discharge) patients when the blocking level is sufficiently low (high) can reduce the average waiting times of both admit and discharge patients.

As a last note, to understand decision makers' cognitive process in patient prioritization, we should be aware of medical practitioners' position on the root cause of ED overcrowding. Exemplary research on this topic includes Pines et al. (2011), in which physicians in 15 countries identified the boarding of admitted patients as the main cause of ED overcrowding. Similarly, the professional association of emergency physicians in Canada (CAEP) pointed out the following (Affleck et al. 2013): *"when inpatients occupy ED stretchers for prolonged periods of time they block access to these care spaces by ill and injured patients in the waiting room and increase waiting times for newly arriving patients ... the inability of admit patients to access in-patient beds from the ED is the most significant factor causing ED overcrowding in Canadian hospitals."* (Here, *inpatients* refer to *boarding patients* in our paper.) With this context, it seems an intuitive and sensible decision for decision makers to prioritize discharge patients in response to ED blocking. Our discussions with ED physicians confirmed this insight.

In conclusion, SPT is very relevant but may not explain our empirical findings on its own. However, there is not enough evidence to rule out SPT either, because the estimation results show that both *WaitTime* and *WaitRoomCensus* affect the prioritization decisions. It is possible that both our mechanism and SPT are at work. Especially, ED nurses might apply SPT since they are the ones that interact with and care for the patients in the waiting room.

9. Conclusions and Future Research

Motivated by an intriguing observation from comparing the average waiting times of admit and discharge patients by triage level, we study how ED decision makers choose the next patient for treatment. Using data from a large urban teaching hospital in Canada, we find that decision makers apply urgency-specific delay-dependent prioritization. Moreover, decision makers start to prioritize discharge patients to prevent

the ED from being further blocked when the blocking level is sufficiently high. We then draw insights from a stylized MDP formulation to explain the rationale behind such prioritization behavior. To the best of our knowledge, it has not been documented that medical workers consider ED blocking in their patient prioritization decisions. Our work fills this gap by providing empirical evidence and explaining the rationale behind it. Our work also contributes to the queueing literature by explicitly modeling the interactions between two bottleneck resources in a two-station multi-class service network. Our simulation study shows that priority policies—derived from our empirical findings and insights from the MDP model—can improve patient flow by reducing average waiting time and LOS. We also show how to leverage our findings to improve waiting time prediction algorithms.

9.1. Managerial Insights

Our study offers useful managerial insights. Our findings suggest that EDs may not operate like a multi-class queueing system in which triage levels indicate strict priority. Rather, ED decision makers use discretion and consider both clinical and operational factors when prioritizing patients. This is not surprising, as management teams at the ED and hospital levels are often practitioners themselves. However, our study provides additional insights into how to leverage such prioritization behavior to improve ED operations. Specifically, we devise prioritization policies under which the average waiting times and LOS for both admit and discharge patients are reduced. As a result, hospitals can save millions of dollars annually due to improved quality of care. Our queueing model explains the interaction between the two bottleneck resources at EDs, namely, physicians and beds (determined by nurses), which also provides insights into ED staffing. For example, during a surge in ED demand, bringing in a nurse is probably more effective than bringing in a physician when the ED blocking level is high. Our recommendations for improving ED waiting time forecasting is practically useful, particularly as an increasing number of hospitals have started posting their ED waiting time online and accurate prediction algorithms are urgently needed (ACEP 2012).

9.2. Relevance to Non-Healthcare Settings

Our empirical findings and MDP model reveal that in a dual-resource constrained system (physicians and beds), when one resource becomes the bottleneck (ED blocking), re-ordering jobs (prioritizing patients) can achieve a better match between capacity and demand. This distinguishing feature is relevant to many *case-manager type systems*, borrowing the term from Campello et al. (2016). For example, in call centers, a customer can be served only when there are agents and trunk lines available, both of which have limited capacity. In a CONWIP production system, a finite number of kanban cards may be introduced to control inventory among a set of workstations. Hence, both kanban cards and workstations need to have available capacity to start a new job. A warehouse has finite pallets and human (or robot) servers—both are necessary—for order fulfillment. In all of these examples, a job, be it a customer, a work, or an order, cannot be

processed if there are no available trunk lines, kanban cards, or pallets. Hence, our findings on patient prioritization are also relevant to service and manufacturing systems. The triage levels in our study represent the job types, e.g., regular and VIP customers in call centers, urgent and non-urgent orders in production lines. The predicted dispositions stand for an early prediction of job demands on certain resources. Our findings suggest that it is beneficial for operations managers to adjust the processing order of jobs in response to the pressure on bottleneck resources; specifically, to prioritize jobs that are less demanding on the bottleneck resources so as to ease the system pressure and preserve the system's capability.

9.3. Future Research Directions

Our study serves as a first step toward understanding the patient prioritization behavior of ED decision makers. However, there are many questions to be explored for the broader implementation of patient prioritization rules. For example, whether decision makers react to ED blocking too little, too much, or just right. Controlled experiments could provide more direct answers to this question. Our analysis is based on data from one hospital. Therefore, the findings may not extend to hospitals of different sizes or to hospitals where ED beds are not the bottleneck resource. Hence, it would be valuable to conduct analysis using data from other hospitals. Our mechanism on patient prioritization may be an aggregation of several lower-level mechanisms involving multiple decision makers. Hence, it would be of interest to examine the behavior of different stakeholders, e.g., nurses and physicians, separately. Ibanez et al. (2018) find that doctors exercise more discretion in task ordering as they accumulate experiences. Therefore, it would be valuable to include decision makers' characteristics in the choice model. Two other issues that are left out of this paper are the quality of care and ethical concerns due to the prioritization (and de-prioritization) of a certain group of patients. It would be of interest to infer the effects of disposition-dependent prioritization on quality of care. For example, the increased waiting for admit patients due to the prioritization of discharge patients may lead to higher risk of adverse health outcomes (Sun et al. 2013, Richardson and Bryant 2004). All of these issues would benefit from further investigation.

References

ACEP. Publishing wait times for emergency department care: an information paper. *Report, American College of Emergency Physicians, Baltimore*, 2012.

A. Affleck, P. Parks, A. Drummond, B. H. Rowe, and H. J. Ovens. Emergency department overcrowding and access block. *Canadian Journal of Emergency Medicine*, 15(6):359–370, 2013.

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

E. Ang, S. Kwasnick, M. Bayati, E. L. Plambeck, and M. Aratow. Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management*, 18(1):141–156, 2015.

N. T. Argon and S. Andradóttir. Replicated batch means for steady-state simulations. *Naval Research Logistics*, 53(6): 508–524, 2006.

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

R. J. Batt and C. Terwiesch. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science*, 61(1):39–59, 2015.

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

R. J. Batt, D. KC, B. R. Staats, and B. W. Patterson. The effects of discrete work shifts on a nonterminating service system. *Production and Operations Management*, 28(6):1528–1544, 2019.

J. A. Berry Jaeker and A. L. Tucker. Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science*, 63(4):1042–1062, 2017.

R. Beveridge, B. Clarke, L. Janes, N. Savage, J. Thompson, G. Dodd, M. Murray, C. N. Jordan, D. Warren, and A. Vadeboncoeur. Implementation guidelines for the canadian emergency department triage & acuity scale (CTAS). *Canadian Association of Emergency Physicians*, pages 1–32, 1998.

G. Cachon and C. Terwiesch. *Matching supply with demand*. McGraw-Hill Publishing, 2008.

Ç. Çağlayan, Y. Liu, T. Ayer, K. Pasupathy, D. Nestler, et al. Physician staffing in emergency rooms (ERs): Opening the black-box of ER care via a multi-class multi-stage network. *Available at SSRN 3400900*, 2019.

F. Campello, A. Ingolfsson, and R. A. Shumsky. Queueing models of case managers. *Management Science*, 63(3): 882–900, 2016.

Canadian Institute for Health Information. Commonwealth Fund Survey: Infographic, 2016. URL `https://www.cihi.ca/en/commonwealth-fund-survey-2016-infographic`. Accessed on May 18, 2020.

R. Carmen, I. Van Nieuwenhuyse, and B. Van Houdt. Inpatient boarding in emergency departments: Impact on patient delays and system capacity. *European Journal of Operational Research*, 271(3):953–967, 2018.

C. W. Chan, J. Dong, and L. V. Green. Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Operations Research*, 65(2):469–495, 2016.

W. Chen, B. Linthicum, N. T. Argon, T. Bohrmann, K. Lopiano, A. Mehrotra, D. Travers, and S. Ziya. The effects of emergency department crowding on triage and hospital admission decisions. *The American Journal of Emergency Medicine*, 2019.

T. Dai and S. Tayur. OM forum—healthcare operations management: A snapshot of emerging research. *Manufacturing & Service Operations Management*, 22(5):869–887, 2020.

H. Dawson and G. Zinck. ED spending in canada: A focus on the cost of patients waiting for access to an in-patient bed in ontario. *Healthcare Quarterly*, 12(1):25–28, 2009.

F. de Véricourt and O. B. Jennings. Nurse staffing in medical units: A queueing perspective. *Operations Research*, 59 (6):1320–1331, 2011.

S. Deo and I. Gurvich. Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science*, 57(7):1300–1319, 2011.

Y. Ding, E. Park, M. Nagarajan, and E. Grafstein. 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, 2019.

G. Dobson, T. Tezcan, and V. Tilson. Optimal workflow decisions for investigators in systems with interruptions. *Management Science*, 59(5):1125–1141, 2013.

T. A. Domencich and D. McFadden. *Urban travel demand: A behavioral analysis*. North-Holland Publishing Company, Oxford, England, 1975.

J. Dong, E. Yom-Tov, and G. B. Yom-Tov. The impact of delay announcements on hospital network coordination and waiting times. *Management Science*, 65(5):1969–1994, 2019.

Y. Duan, Y. Jin, Y. Ding, M. Nagarajan, and G. Hunte. The cost of task switching: Evidence from the emergency department. *Available at SSRN 3756677*, 2020.

Y. B. Ferrand, M. J. Magazine, U. S. Rao, and T. F. Glass. Managing responsiveness in the emergency department: Comparing dynamic priority queue with fast track. *Journal of Operations Management*, 58:15–26, 2018.

M. Freeman, N. Savva, and S. Scholtes. Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science*, 63(10):3147–3167, 2016.

J. K. Gorski, R. J. Batt, E. Otles, M. N. Shah, A. G. Hamedani, and B. W. Patterson. The impact of emergency department census on the decision to admit. *Academic Emergency Medicine*, 24(1):13–21, 2017.

A. Holdgate, J. Morris, M. Fry, and M. Zecevic. Accuracy of triage nurses in predicting patient disposition. *Emergency Medicine Australasia*, 19(4):341–345, 2007.

J. Huang, B. Carmeli, and A. Mandelbaum. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4):892–908, 2015.

M. R. Ibanez, J. R. Clark, R. S. Huckman, and B. R. Staats. Discretionary task ordering: Queue management in radiological services. *Management Science*, 64(9):4389–4407, 2018.

R. Ibrahim and W. Whitt. Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research*, 59(5):1106–1118, 2011.

A. Ingolfsson, E. Akhmetshina, S. Budge, Y. Li, and X. Wu. A survey and experimental comparison of service-level-approximation methods for nonstationary M(t)/M/s(t) queueing systems with exhaustive discipline. *INFORMS Journal on Computing*, 19(2):201–214, 2007.

D. KC. Does multitasking improve performance? Evidence from the emergency department. *Manufacturing & Service Operations Management*, 16(2):168–183, 2013.

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

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

D. KC, B. R. Staats, M. Kouchaki, and F. Gino. Task selection and workload: A focus on completing easy tasks hurts performance. *Management Science*, 66(10):4397–4416, 2020.

S.-H. Kim and W. Whitt. Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management*, 16(3):464–480, 2014.

S.-H. Kim, C. W. Chan, M. Olivares, and G. Escobar. ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science*, 61(1):19–38, 2014.

S.-H. Kim, J. Tong, and C. Peden. Admission control biases in hospital unit capacity management: How occupancy information hurdles and decision noise impact utilization. *Management Science*, 66(11):5151–5170, 2020.

L. Kuntz and S. Sülz. Treatment speed and high load in the emergency department—does staff quality matter? *Health Care Management Science*, 16(4):366–376, 2013.

A. M. Law and W. D. Kelton. *Simulation modeling and analysis*. McGraw-Hill New York, 3rd edition, 2000.

N. Li, D. A. Stanford, P. Taylor, and I. Ziedins. Nonlinear accumulating priority queues with equivalent linear proxies. *Operations Research*, 65(6):1712–1721, 2017.

S. A. Lippman. Applying a new device in the optimization of exponential queuing systems. *Operations Research*, 23 (4):687–710, 1975.

D. Luo, M. Bayati, E. L. Plambeck, and M. Aratow. Low-acuity patients delay high-acuity patients in the emergency department. *Available at SSRN 3095039*, 2017.

M. Murray, M. Bullard, E. Grafstein, et al. Revisions to the canadian emergency department triage and acuity scale implementation guidelines. *Canadian Journal of Emergency Medicine*, 6(6):421–427, 2004.

J. S. Olshaker and N. K. Rathlev. Emergency department overcrowding and ambulance diversion: The impact and potential solutions of extended boarding of admitted patients in the emergency department. *The Journal of Emergency Medicine*, 30(3):351–356, 2006.

J. M. Pines, J. A. Hilton, E. J. Weber, A. J. Alkemade, H. Al Shabanah, P. D. Anderson, M. Bernhard, A. Bertini, A. Gries, S. Ferrandiz, et al. International perspectives on emergency department crowding. *Academic Emergency Medicine*, 18(12):1358–1370, 2011.

A. Powell, S. Savin, and N. Savva. Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing & Service Operations Management*, 14(4):512–528, 2012.

D. B. Richardson and M. Bryant. Confirmation of association between overcrowding and adverse events in patients who do not wait to be seen. *Academic Emergency Medicine*, 11(5):462, 2004.

S. Saghafian, W. J. Hopp, M. P. Van Oyen, J. S. Desmond, and S. L. Kronick. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research*, 60(5):1080–1097, 2012.

S. Saghafian, W. J. Hopp, M. P. Van Oyen, J. S. Desmond, and S. L. Kronick. Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management*, 16 (3):329–345, 2014.

S. Saghafian, G. Austin, and S. J. Traub. Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering*, 5(2):101–123, 2015.

P. Shi, M. C. Chou, J. Dai, D. Ding, and J. Sim. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science*, 62(1):1–28, 2015.

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

B. C. Sun, R. Y. Hsia, R. E. Weiss, D. Zingmond, L. Liang, W. Han, H. McCreath, and S. M. Asch. Effect of emergency department crowding on outcomes of admitted patients. *Annals of Emergency Medicine*, 61(6):605–611, 2013.

Y. Sun, K. L. Teow, B. H. Heng, C. K. Ooi, and S. Y. Tay. Real-time prediction of waiting time in the emergency department, using quantile regression. *Annals of Emergency Medicine*, 60(3):299–308, 2012.

Z. Sun, N. T. Argon, and S. Ziya. Patient triage and prioritization under austere conditions. *Management Science*, 64 (10):4471–4489, 2018.

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

K. E. Train. *Discrete choice methods with simulation*. Cambridge university press, Cambridge, UK, 2009.

United States Government Accountability Office. Hospital emergency departments: Crowding continues to occur, and some patients wait longer than recommended time frames. *GAO Report (GAO-09-347)*, 2009.

M. R. Vaghasiya, M. Murphy, D. O'Flynn, and A. Shetty. The emergency department prediction of disposition (EPOD) study. *Australasian Emergency Nursing Journal*, 17(4):161–166, 2014.

J. S. van Leeuwaarden, B. W. Mathijsen, F. Sloothaak, and G. B. Yom-Tov. The restricted Erlang-R queue: Finite-size effects in service systems with returning customers. *arXiv preprint arXiv:1612.07088*, 2016.

S. J. Weiss, A. A. Ernst, R. Derlet, R. King, A. Bair, and T. G. Nick. Relationship between the national ED overcrowding scale and the number of patients who leave without being seen in an academic ED. *The American Journal of Emergency Medicine*, 23(3):288–294, 2005.

L. Woodworth and J. F. Holmes. Just a minute: The effect of emergency department wait time on the cost of care. *Economic Inquiry*, 58(2):698–716, 2020.

G. B. Yom-Tov and A. Mandelbaum. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014.