# 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
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 physician initial assessment) of discharge patients is shorter than that of admit patients for middle-to-low acuity patients, suggesting that the order of patients being served deviates from FCFS (first-come-first-served), and to certain extent, discharge patients are prioritized over admit patients. This observation is intriguing as among patients of the same triage level, admit patients—who need further care in the hospital—should be deemed no less urgent than discharge 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, when ED blocking level is sufficiently low, admit patients are prioritized over discharge patients for high acuity patients, and FCFS is followed for middle-to-low acuity patients. When the risk of ED being blocked becomes sufficiently high, decision makers start to prioritize patients who are less likely to be admitted after treatment at ED, in an effort to avoid further blocking the ED. We then analyze a stylized model to explain the rationale behind decision makers' prioritization behavior when the ED faces increasing risk of being blocked. We also investigate the impact of such prioritization behavior on ED operational performances and show how to leverage our findings to improve ED waiting time prediction. By testing and highlighting the central role of decision makers' patient prioritization behaviors, this paper advances our understanding of ED operations and patient flow.

*Key words*: Patient Prioritization, ED Blocking, Discrete Choice Model, MDP

## 1. Introduction

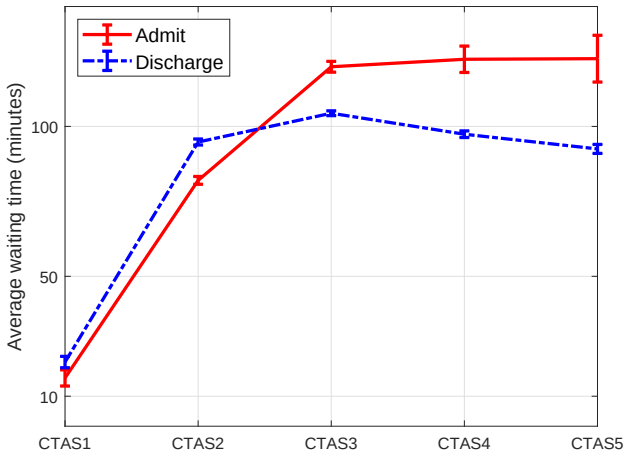
Emergency department (ED) waiting time—the total time from triage to physician initial assessment—is a well-established metric of the timeliness of emergency care. Unfortunately, long waiting time is extremely common in many countries around the world. In the United States, 1 out of 4 patients waits more than 2 hours to see a physician in 2006 (GAO. 2009). In 2015–2016, 1 out of 3 Canadian patients reports waiting

4 or more hours the last time they sought care at EDs (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), increased rate of readmission (Richardson and Bryant 2004), among others. Hence, it is critical to understand the drivers of ED waiting time. One key determinant is the order according to which patients are seen by ED healthcare professionals. Upon arrival at ED, patients are first triaged by a nurse, and a score is assigned based on their acuity and resource needs (in some triage protocols). In Canada, the most prevalent triage protocol is the Canadian Triage and Acuity Scale (CTAS), a five-point scoring scale (1 to 5) with smaller number indicating 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 upon arrival), however, as stated clearly in the implementation guidelines, these time responses are not established care standards. Without an explicit guideline, decision makers use discretion to select the next patient and the sequence 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 is also evidenced by Ding et al. (2019) using data from EDs in metro Vancouver. However, 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, is unclear.

Despite the importance of understanding how patients are prioritized at ED, the literature on this topic is limited. A very 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 rule of first-come-first-served (FCFS) 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 it below.

After the treatment at ED, patients who need further care at inpatient units will be admitted. We label them *admit patients* and those who do *not* need further care at the hospital *discharge patients*. The destination of patients after their treatment at ED is called *disposition*. Using patient-level data from an urban hospital in 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 more than discharge patients in average (significant at the 5% level), suggesting that discharge patients are prioritized over admit patients to certain extent (see Figure 1). This observation is intriguing since for patients of the same triage level, admit patients—who need further care after ED treatment—should be deemed no less urgent than discharge

**Figure 1** **Comparisons between the average waiting times of admit patients and discharge patients by triage level in an urban teaching hospital in Alberta, Canada. The error bars represent 95% confidence intervals.**




patients. Thus, one may wonder why ED decision makers prioritize discharge patients over admit patients within the same triage level, which motivates our investigation.[1]

The discrepancy in waiting times for patients within the same triage level is a reflection of the patient prioritization behaviors of ED healthcare professionals, and to some extent, the resource allocation within the same triage level. To the best of our knowledge, this has not been studied in the literature. Our objective is to shed some light on ED decision makers' prioritization behaviors. Particularly, we focus on three research questions:

1. How does disposition affect the prioritization of patients from the same triage level?
2. What is the rationale behind ED healthcare professionals' prioritization behavior?
3. What is the impact of such behaviors on ED operational outcomes?

To answer the first question, we empirically examine patient prioritization decisions by a discrete choice model using three years' 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 the ED's ability of treating new patients is compromised due to many ED beds being occupied by *boarding patients*—admit patients waiting in ED beds to be transferred. We first apply binary classification models to predict the likelihood of the disposition of a patient only using information from triage. The predicted disposition enters the discrete choice model as a proxy of the estimated disposition by a decision maker. We address the second research question by developing a Markov decision process (MDP)

[1] 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 at 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.

model to study the optimal prioritization policy and connect it to the empirical findings. At last, we discuss the impact and implications of our findings. Our study makes the following contribution to the literature:

**1. Empirical findings.** Using a discrete choice model, we find that ED decision makers apply urgency-specific delay-dependent prioritization in general. That is, higher urgency and longer waiting time both lead to higher priority. In addition, we find that (i) When ED blocking level is sufficiently low, admit patients are prioritized over discharge patients for high acuity patients (triage level 2), and FCFS is followed for middle-to-low acuity patients (triage levels 3 and 4) within the same triage level; (ii) As the blocking level increases, decision makers prioritize patients who are less likely to be admitted after ED treatment. To our knowledge, this is among the first work that studies how ED decision makers prioritize patients in response to the risk of ED blocking.

**2. Rationale behind the prioritization behavior.** We build an MDP formulation to study how to select patients from the same triage level under different levels of ED blocking with the objective of maximizing social benefits. By characterizing the structure of the optimal policy, we show that it is optimal for decision makers to dynamically prioritize patients, especially, prioritize discharge patients when ED faces high risk of being blocked. The structure of the optimal policy aligns with our empirical findings, and provides insights on the rationale of decision makers behind their dynamic patient prioritization behaviors.

**3. Implications of our findings.** Prioritizing discharge patients when facing high risk of ED blocking is the main takeaway of this paper. Through a simulation study, we demonstrate the potential benefits of policies devised from such prioritization behaviors on ED operational performances, such as the expected long-run average waiting time and length of stay. We also show how to leverage our findings to improve algorithms on ED waiting time prediction. Our results provide support for broader implementation and standardization of such practice. However, discretion should be exercised as there are factors that our study does not fully capture. Controlled experiments are called to provide more direct evidence on the benefits from such prioritization behaviors.

The rest of this paper is organized as follows. We conclude this section with a review of related literature. We introduce the study setting and describe the patient-level data for our empirical investigation in Section 2. We empirically study how decision makers choose the next patient to see in Section 3. In Section 4, we develop an MDP formulation and characterize the optimal policy. We discuss the connections between the optimal policy and our empirical findings and explain the rationale behind decision makers' prioritization behavior. In Sections 5 and 6, we examine the impact of our finding on ED operational outcomes and waiting time prediction, respectively. Section 7 concludes this work and points to future directions. All proofs and additional results are given in the appendix.

### 1.1. Literature Review

In recent years, operations research/management tools have been applied widely to improve patients' access to emergency care (see Saghafian et al. (2015) and Dai and Tayur (2019) for overview). Studies on healthcare professionals' behavior are particularly relevant to ours. It is known that healthcare professionals respond to system workload by making different prioritization and capacity rationing decisions (KC and Terwiesch 2009, 2012, Powell et al. 2012, Kuntz and Sülz 2013, Kim et al. 2014, Batt and Terwiesch 2016, Berry Jaeker and Tucker 2016, Freeman et al. 2016, Ding et al. 2019, Kim et al. 2019). These studies suggest that the decision making of healthcare practitioners is not purely driven by clinical factors, and identify various mechanisms through which system workload impacts their behavior. Our work studies the patient prioritization behaviors of ED nurses and physicians thus are 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 critically 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. Our work differs in that we focus on the impact of capacity constraints in addition to other factors such as triage levels and waiting times. The discrepancy is driven potentially by that "physicians are the bottleneck resources in the treatment process" in their study hospitals, while both can be bottlenecks in our study ED. We also study a decision model to explain the mechanism behind decision makers' prioritization behavior.

Another stream of relevant papers concerns the phenomenon *ED blocking* or *bed block*, referring to situations in which admit patients wait in ED beds to be transferred and lead to insufficient ED beds to treat new patients. Despite that it is *the most significant factor* causing ED overcrowding (Affleck et al. 2013), research on this topic has been relatively limited (Saghafian et al. 2015). Two recent works (Shi et al. 2015, Chan et al. 2016) study this problem by controlling the discharge (inspection) timing at inpatient wards. They model the inpatient flow dynamics through time-varying queues and proposed 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 arriving to inpatient wards so as to alleviate ED blocking.

A number of articles have investigated various aspects of EDs, and thus are relevant. The research topics that have been studied include waiting time prediction (Ibrahim and Whitt 2011, Ang et al. 2015), the impact of delay announcement (Dong et al. 2019), ambulance diversion decisions and impact (Allon et al. 2013, Xu and Chan 2016), among others. 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), Tan and Staats (2020), Ibanez et al. (2018), Ferrand et al. (2018). Among them, Ibanez et al. (2018) and Tan and Staats (2020) are particularly

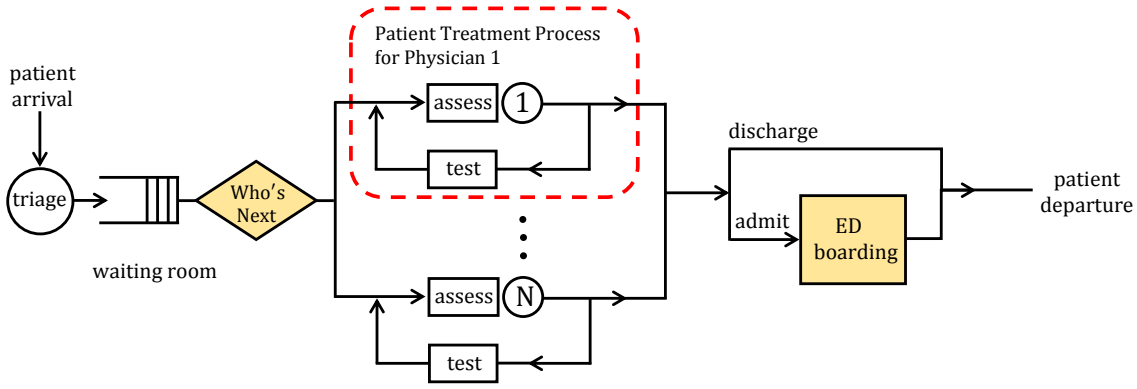
relevant, as the former studies task ordering decision of a radiologist and its impact on productivity and the latter studies how the customer routing decisions depends on workload, service speed, and sales skill. Our work is similar in that we also study how decision makers use discretion to order tasks (prioritize patients), but our study focuses on the prioritization decisions with respect to the change in system state.

Finally, part of our empirical results implies that decision makers apply time-dependent prioritization, which makes our work also relevant to the stochastic scheduling literature, in particular, to studies on time-dependent priority rules (see, e.g., Kleinrock 1964, Sharif et al. 2014, Li et al. 2017). However, the prioritization decision in our work also depends on system congestion level which differs from the aforementioned studies. Our focus and objectives also differ from existing works significantly.

## 2. ED Patient Flow and Data

In this section, we describe the ED patient flow process and the data for our empirical investigation. We note that different EDs may operate differently. Though our description is based on a Canadian model, we believe that the key characteristics are similar in most EDs. We represent the general patient flow in Figure 2, and will come back to it whenever it helps readers understand the depiction.

**Figure 2 The general patient flow in an emergency department.**




### 2.1. Patient Flow

The arrival modes of patients to the ED are either by emergency medical services (EMS, such as ambulance) or by their own transportation. Upon arrival, patients are triaged into 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 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 the study of such an aggregated selection process can still reflect the patient prioritization decisions at EDs. It would be interesting to investigate the decision makings by nurses and physicians separately if there were sufficiently details in the data.

After the initial assessment, some patients may leave the ED, while others may undergo some medical procedure, 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*. A patient will join the test queue (Figure 2) to go through testing and wait for the results if tests are needed. The patient returns to the same physician for another round of assessment when test results are ready. This assess-test process can repeat itself, 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). We learned from our collaborating physicians that every physician knows her service capacity, i.e., the maximum number of patients she can care. 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 4.

### 2.2. 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 patient holds the assigned bed, even when she is sent to the test center for testing, until she leaves ED. 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 continues staying in the ED bed until she is transferred out. The total time from bed request to 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 diagnosis and treatment at ED are completed. 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 care spaces by newly arriving patients in the waiting room. We refer this phenomenon as *ED blocking*, which is known as the main contributor to ED overcrowding (Affleck et al. 2013). Note that all patients staying in ED beds need nursing care, regardless of boarding patients or patients whose diagnosis are in process. Thus, it is impossible to solve the ED blocking problem by adding more beds without increasing the nurse staffing level.

### 2.3. Data Description and Cleaning

This study uses data from the ED electronic health record system of an urban hospitals in Alberta, Canada. The dataset covers all patient visits from August 2013 to July 2015, approximately 151,000 observations.[2]

[2] We have another dataset containing only the triage information and disposition of patients who visited this ED between August 2015 and August 2016. This dataset is used for the training of binary classification models. See more details in Section 3.2.1.

The daily arrivals to this ED over the study period ranges from 131 to 257, and the median (mean) is 206 (207.4). The dataset contains patient-level visit records. 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 ED), and disposition. All visit records are de-identified to protect the privacy of patients and medical personnel.

**Table 1 Summary Statistics**

| Variables | CTAS2 | CTAS3 | CTAS4 |
|---|---|---|---|
| Waiting time (Mean and Stdev in minutes) | | | |
| All patients | 96.3 (93.2) | 118.9 (96.2) | 111.6 (90.4) |
| Admit patients | 86.5 (91.5) | 128.5 (101.7) | 135.0 (100.0) |
| Discharge patients | 101.1 (93.7) | 115.7 (94.4) | 108.4 (88.4) |
| Age (Mean and Stdev in years) | 52.6 (20.5) | 51.6 (21.5) | 47.7 (21.3) |
| Gender (female%) | 51.5% | 58.2% | 59.1% |
| Disposition (admit%) | 32.7% | 23.1% | 11.6% |
| Arrival mode (ambulance arrival%) | 38.1% | 28.4% | 20.0% |
| Observations | 29582 | 29471 | 9288 |

In our study ED, a fast-track line is open daily from 10:00 to 24:00, which concurs with the period when the ED is critically loaded. We are interested in prioritization decisions of non-fast-track patients during peak load hours, thus, visit records outside of the peak load hours and that of fast-track patients are eliminated (approximately 36.4% of available data). Physicians may cherry-pick patients near end of shift to avoid undesirable outcomes such as excessive patient handoffs or overtime (Batt et al. 2019, Chan 2018). This end-of-shift effect is not our focus, thus, we drop the last new patient selected in the shift (approximately 5.5% of available data). Since we are interested in studying how a patient's priority is determined by decision makers, we drop patients of triage level 1 because they receive preemptive priority over all the other patients, and their dispositions will unlikely affect their priorities. In addition, more than 90% of triage level 5 patients go to fast-track and are discharged home after ED treatment. Hence, this part of the data is highly biased and thus are eliminated (approximately 4.5% of available data). We also drop patients whose dispositions are not "admit' or "discharge" as they are not our focus (about 0.27% of available data). Visit records with rare chief complaint codes (frequency less than 200) and postcode (frequency less than 150) are eliminated (approximately 6.1% of available data). Table 1 is the summary statistics table of the data after cleaning.

We use the dataset to investigate how decision makers choose new patients by a discrete choice model. 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.

**2.3.1. 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 arrived between $t_1$ and $t_2$, less the patient selected at $t_1$ and patients who become absent from ED between $t_1$ and $t_2$ (due to leave without being seen or being transferred, etc.).

**2.3.2. Explanatory Variables** Over the study period, the compensation to medical personnel is shift-based salary. To our knowledge, there is no financial incentive for decision makers to select patients based on their complexity or medical expense. This concurs with the observation in Ding et al. (2019) in which the authors also focus on Canadian hospitals. When deciding which patient to select, we believe that ED decision makers' objective is to provide timely care to patients who need it most urgently. Hence, we focus on clinical and operational factors that are potentially related to the 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 3.1). We study how disposition affects patient prioritization by controlling ED blocking level, which measures the extent to which the ED's ability of treating new patients is compromised due to ED beds being occupied by boarding patients at $t$.

One challenge in our study is the estimate of ED blocking level, which depends on both the number of boarding patients and the capacity of ED beds. From our data, the number of boarding patients at any time can be counted, however, the capacity of ED beds depends on the nurse staffing levels as ED beds are suitable for caring patients only 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 observed numbers of patients in ED beds at any given hour of the day over the two-year horizon—round it to the nearest integer—as a proxy of ED bed capacity at this particular hour of the day. We develop two measures for ED blocking level at any time $t$. The first one is the actual number of boarding patients at $t$ normalized to be between 0 and 1. The second one is the ratio of the number of boarding patients over the number of "extra" beds (total bed capacity net of physician capacity) at time $t$. The latter seems to be a better measure of blocking level as both bed capacity and physician capacity are time-dependent (they depend on the shift schedules of nurses and physicians). However, the former is simpler and readily available on ED's dashboard. It is unclear how exactly ED decision makers infer the blocking level. Hence, we test both measures in our empirical investigation.

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. Chief complaint codes are entered at triage by a nurse through a drop-down menu. There are 170 codes observed from our datasets. We drop the codes with less than 200 observations and the corresponding visit records (approximately 3.6% of available data). 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.

## 3. Empirical Investigation of Patient Prioritization

In this section, we empirically examine ED decision makers' prioritization behaviors in our study ED. We first make 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 select is made in the hospital's medical information system. Through a terminal, decision makers can access real-time information of all patients waiting to be seen, including comprehensive triage data (includes more details to our dataset, such as vital signs, revisit patient or not), waiting time thus far, etc. Existing studies have shown that ED physicians and nurses can predict a patient's disposition fairly accurately using triage information, see Holdgate et al. (2007) and Vaghasiya et al. (2014) among others. The number of boarding patients is an important measure of ED crowding, available to nurses/physicians in real time. Hence, when selecting a new patient, decision makers can infer the ED blocking level based on the given information.

### 3.1. Econometric Models

We investigate the relationship between a decision maker's choice of the next patient to see and the characteristics of the alternatives (patients) under resource constraints (e.g., ED blocking) by a discrete choice framework. We believe the variations in patients' characteristics and system resource constraints are the main drivers of decision makers' behavior in selecting the next patient, not decision makers' own attributes (such as specialty, experiences, etc.), hence, we choose a mixed logit model for our investigation. Mixed logit model is highly flexible and can approximate any random utility model (Train 2009). It differs from conditional logit models in that it allows random "taste" variation to account for heterogeneity in decision makers and correlation in unobserved factors over time (Train 2009). They both belong to the family of random utility models, in which a decision maker chooses the alternative that maximizes her perceived utility.

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 $\varepsilon_{it}$, 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 perceived 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 $\varepsilon_{it}$ 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:

$$\begin{aligned} V_{it} = \beta_0 + \beta_1^T \mathbf{C}_i + \beta_2^T CTAS_i \times WaitTime_{it} + \beta_3^T CTAS_i \times Disposition_i \\ + \beta_4^T CTAS_i \times Disposition_i \times BlockLevel_t. \end{aligned} \tag{2}$$

In (2), the model parameters $\beta_j's$ are assumed to follow a Normal probability distribution to capture the variation in decision makers' prioritization behaviors. The vector $\mathbf{C}_i$ contains the time-invariant characteristics of patient $i$, including age group, gender, arrival mode, triage level, 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 vector $CTAS_i$ is an indicator of the triage level of patient $i$. The element of $CTAS_i$ is 1 if it corresponds to patient $i$'s triage level and 0 otherwise. The scalar $WaitTime_{it}$ represents the current waiting time of patient $i$, i.e., the time duration from patient $i$'s arrival to the decision epoch $t$. Waiting time is an important determinant of patient priority, and the effect of *WaitTime* on the utility of patients may vary across triage levels, thus, we include the interaction term of *CTAS* and *WaitTime* in Equation (2). Furthermore, waiting time may affect a patient's utility in a nonlinear way (Ding et al. 2019, Ferrand et al. 2018). We will test the robustness of our results by including different function forms of *WaitTime*, particularly nonlinear forms. The ED blocking level, denoted by *BlockLevel*, measures the extent to which the ED's ability of providing timely care is impaired. The anticipated disposition of patient $i$ by ED decision makers, denoted by $Disposition_i$, can be viewed as an exogenous treatment on patient $i$, $i \in J(t)$. To investigate how the treatment affects a patient's priority of being selected across triage levels, we add the interaction term of *CTAS* and *Disposition*; we further add the interaction term of *CTAS*, *Disposition*, and *BlockLevel*, to show how the effect varies with the ED blocking level.

One challenge in our investigation is that the anticipated disposition of patient $i$ by a decision maker at any decision epoch, $Disposition_i$, is not available in our dataset. 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 concern of reverse causality and ignoring it leads to potential bias. To solve this issue, we follow the idea of endogenous treatment regression models (Heckman 1977, Ibanez et al. 2018) and use the predicted disposition of patient $i$ as a proxy for $Disposition_i$. The prediction is done via binary classification models, and the predictors are basic patient characteristics collected at triage. We note that there are more than one binary classification models, such as logistic regression, probit model, CART (classification and regression tree), etc. Though disposition prediction is not the focus of this paper, it is important to make sure that our findings hold regardless of the choice of the classifier. We tried the aforementioned three methods and found that the results are similar. Therefore, we only present the results from the logistic regression model in the paper. The results for other classifiers are deferred to Appendix A.

### 3.2. Results and Discussions

**3.2.1. Disposition Prediction** To avoid potential endogeneity issue, we request another dataset containing only the triage information and dispositions of patients who visited this ED between August 2015 and August 2016, approximately 81,000 records. We use this dataset to train prediction models, and apply the trained model to our first dataset. With only six basic patient characteristics as predictors (age group, gender, postcode, arrival mode, triage level, and chief complaint), a standard logistic regression model performs reasonably well with an out-of-sample c-statistic close to 0.79.

**3.2.2. Patient Prioritization** To study decision makers' patient prioritization behaviors, we estimate Equation (2) by maximizing the likelihood of choosing the observed choice of patient. The random component of a patient's utility $\varepsilon_{it}$ may have heteroscedasticity, which could make the maximum likelihood estimates of the parameters inconsistent and biased. We account for the potential heteroscedasticity by using the Huber-White Sandwich estimator. Next, we discuss the results of Model 1 in Table 2. Model 1 serves as the baseline model for our robustness check later on, whose specification is shown in Equation (2). The variable $Disposition_i$ is the likelihood of patient $i$ being admitted estimated by a logit model. Note that the model goodness-of-fit is measured by the McFadden pseudo $R^2$. The empirically equivalent $R^2$ in linear models are estimated from Figure 5.5 in Domencich and McFadden (1975).

**Observation 1.** *Decision makers apply urgency-specific delay-dependent prioritization rule when selecting the next patient for initial assessment.*

The estimation results show that patient characteristics, including age group, gender, arrival mode, and chief complaint code, all are factors that affect decision makers' prioritization decisions; they are not included

**Table 2 Key determinants of patient prioritization decisions for triage levels 2, 3, and 4. Model 1 serves as the base model, and Models 2-5 are part of the robustness tests.**

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| *Triage Level* = 2 | | | | | |
| *CTAS* × *WaitTime* | 0.005*** | 0.004*** | 0.005*** | 0.004*** | 0.004*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *CTAS* × *Disposition* | 0.556*** | 0.37*** | 0.858*** | 0.923*** | 0.603*** |
| | (0.048) | (0.030) | (0.062) | (0.095) | (0.058) |
| *CTAS* × *Disposition*×*BlockLevel* | −0.693*** | −0.482*** | −2.069*** | −2.063** | −0.496*** |
| | (0.117) | (0.081) | (0.086) | (0.209) | (0.043) |
| *Triage Level* = 3 | | | | | |
| *CTAS* | −0.531*** | −0.52*** | −0.396*** | −0.398*** | −0.39*** |
| | (0.024) | (0.018) | (0.023) | (0.027) | (0.021) |
| *CTAS* × *WaitTime* | 0.009*** | 0.008*** | 0.009*** | 0.008*** | 0.008*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *CTAS* × *Disposition* | 0.117* | 0.041 | 0.088 | −0.274* | −0.46*** |
| | (0.053) | (0.035) | (0.104) | (0.118) | (0.074) |
| *CTAS* × *Disposition*×*BlockLevel* | −0.666*** | −0.569*** | −2.529*** | −2.225*** | −0.462*** |
| | (0.129) | (0.094) | (0.206) | (0.284) | (0.057) |
| *Triage Level* = 4 | | | | | |
| *CTAS* | −0.707*** | −0.55*** | −0.498*** | −0.557*** | −0.458*** |
| | (0.034) | (0.025) | (0.031) | (0.035) | (0.028) |
| *CTAS* × *WaitTime* | 0.01*** | 0.007*** | 0.008*** | 0.009*** | 0.007*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *CTAS* × *Disposition* | 0.109 | −0.038 | 0.167 | −0.641* | −0.996*** |
| | (0.134) | (0.083) | (0.235) | (0.302) | (0.190) |
| *CTAS* × *Disposition*×*BlockLevel* | −1.598*** | −0.947*** | −4.078*** | −3.227*** | −0.489** |
| | (0.444) | (0.248) | (0.633) | (0.894) | (0.165) |
| Observations | 68341 | 68341 | 68341 | 40012 | 68341 |
| McFadden pseudo $R^2$ | 0.047 | 0.046 | 0.058 | 0.051 | 0.047 |
| (Equivalent linear model $R^2$) | (0.095) | (0.093) | (0.124) | (0.105) | (0.095) |

*Notes.* This table reports the results from mixed choice model. Robust standard errors are shown in the parentheses. Controls not shown are age group, gender, arrival mode, and chief complaint code.
***p<0.001; **p<0.01; *p<0.05

in Table 2 as they are not the focus on this paper. As seen from Model 1 of Table 2, among patients with the same characteristics, serving a patient of higher urgency (smaller triage level) generates higher utility to decision makers. This is consistent with the principle of triage, i.e., classify and prioritize patients according to their clinical urgency. By the triage protocol, patients classified as level 2 are more urgent than that of levels 3 and 4, thus, selecting a level 2 patient incurs higher utility to decision makers than a level 3 (or 4) patient. The interaction term, *CTAS* × *WaitTime*, is significant and positive for all triage levels, implying that selecting a patient with longer waiting time generates higher utility. This implies that a less urgent patient who has waited longer could be selected by ED decision makers for treatment when there are more urgent patients waiting. In other words, ED does not operate like a multi-class queueing system where triage levels indicate strict priority. Rather, the accumulating priority queues studied in Sharif et al. (2014) and Li et al. (2017) are potentially more appropriate models. From a clinical perspective, patients who have waited longer

face higher risk of adverse outcome. Moreover, their conditions may deteriorate faster as they wait longer (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 (in minute) varies across triage levels: less urgent patients have a greater marginal utility increase in waiting time, thus, selecting a less urgent patient with longer waiting time may result in higher utility. In summary, our results imply that decision makers apply urgency-specific delay-dependent prioritization rule, which aligns with results in existing literature (Ding et al. 2019, Ferrand et al. 2018).

**Observation 2.** *Decision makers (i) prioritize admit patients or follow FCFS when ED blocking level is sufficiently low; (ii) prioritize discharge patients when ED blocking level is sufficiently high.*

Observation 1 explains why less urgent patients may get seen earlier. However, it remains unclear why patients of the same urgency are not seen in an FCFS manner. Especially, why and how does a patient's disposition affect her priority? The two interaction terms in Equation (2), $CTAS \times Disposition$ and $CTAS \times Disposition \times BlockLevel$, help explain it. From the estimates of Model 1 in Table 2, we make the following observations: When the risk of ED blocking is sufficiently small, then only the first interaction term is relevant. For patients of triage level 2, selecting admit patients generates higher utilities than selecting discharge patients. Thus, admit patients are prioritized within triage level 2. However, the difference in utility between admitted and discharge patients is insignificant for triage levels 3 and 4. Thus, FCFS is followed. As ED blocking level increases, the utility of choosing admit patients decreases for all triage levels, compared with the utility of choosing discharge patients. This result suggests that within the same triage level, a discharge patient may get seen earlier than an admit patient when the blocking level is sufficiently high, given they have similar characteristics and waiting times. We note that the coefficient of $CTAS \times Disposition \times BlockLevel$ for triage level 4 is significantly larger in magnitude than that for triage levels 2 and 3, meaning that ED blocking level has a greater impact on triage level 4 patients.

**Observation 3.** *ED physicians exhibit heterogeneous patient prioritization behaviors.*

Prioritizing discharge patients when facing high risk of ED blocking is not an explicit policy in the study ED. We believe this is a spontaneous reaction of practitioners who face ED blocking on a daily basis and understand the causes and consequences of ED blocking profoundly. Such decisions may depend on physicians' training, past experience, and risk attitude. Hence, it is plausible that decision makers may behave differently on whether to prioritize discharge patients and, if so, under what circumstances. To gain more insights, we perform a similar analysis using the baseline model (Model 1) on individual physicians with the same dataset. Our data include 130 physicians with 68,341 patient visit records, on average 525.7 visits per physician. Despite that we choose the five physicians who have treated most patients in our dataset, the sample size is not sufficient to estimate the mixed logit model. We then resort to estimate a conditional logit

model, which often leads to same conclusions as mixed logit models (Maddala 1983, Dahlberg and Eklöf 2003).

The estimation results are shown in Table 3. It appears that all five physicians take patients' estimated dispositions into their prioritization decisions to certain extent, however, their behaviors are different from each other and vary across triage levels, both qualitatively and quantitatively. The estimation results are more consistent across the five physicians for triage levels 2 & 3, which resemble Observations 1 & 2. On the other hand, the estimated coefficients for terms of triage level 4 are mostly insignificant. A deeper look at the data reveals that there are only a few patients of triage level 4 being admitted (between 20 to 31 patients for the five physicians over the two-year study period), and the number of discharge patients dominates. This might explain why *CTAS*×*Disposition*×*BlockLevel* of triage level 4 has bigger robust standard errors than that of triage levels 2 & 3 for all five physicians thus leads to insignificant results.

**Table 3 Key determinants of patient prioritization decisions for five physicians who treated most patients in our data.**

| | MD 1 | MD 2 | MD 3 | MD 4 | MD 5 |
|---|---|---|---|---|---|
| *Triage Level* = 2 | | | | | |
| *CTAS*×*WaitTime* | 0.003*** | 0.003*** | 0.002*** | 0.008*** | 0.003*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *CTAS*×*Disposition* | 1.463*** | 0.939* | 1.601*** | 0.931* | 0.311 |
| | (0.315) | (0.458) | (0.462) | (0.453) | (0.443) |
| *CTAS*×*Disposition*×*BlockLevel* | −2.800*** | −1.799 | −3.573** | −3.444** | −2.232* |
| | (0.834) | (1.127) | (1.117) | (1.242) | (1.108) |
| *Triage Level* = 3 | | | | | |
| *CTAS* | −0.407*** | −0.401*** | −0.473*** | −0.593*** | −0.894*** |
| | (0.096) | (0.135) | (0.138) | (0.144) | (0.148) |
| *CTAS*×*WaitTime* | 0.007*** | 0.008*** | 0.006*** | 0.012*** | 0.009*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *CTAS*×*Disposition* | 0.431 | 0.949 | 1.128* | 0.298 | 0.702 |
| | (0.392) | (0.555) | (0.568) | (0.529) | (0.521) |
| *CTAS*×*Disposition*×*BlockLevel* | −2.400* | −4.918*** | −4.623** | −3.767* | −4.337** |
| | (1.013) | (1.430) | (1.415) | (1.507) | (1.376) |
| *Triage Level* = 4 | | | | | |
| *CTAS* | −0.284* | −0.243 | −0.390* | −1.27*** | −0.794*** |
| | (0.128) | (0.187) | (0.172) | (0.234) | (0.189) |
| *CTAS*×*WaitTime* | 0.004*** | 0.007*** | 0.003*** | 0.015*** | 0.005*** |
| | (0.001) | (0.000) | (0.001) | (0.001) | (0.000) |
| *CTAS*×*Disposition* | 0.643 | 0.934 | 0.624 | −1.095 | 0.971 |
| | (0.922) | (1.579) | (1.434) | (1.774) | (1.004) |
| *CTAS*×*Disposition*×*BlockLevel* | −3.006 | −4.103 | −1.274 | −0.663 | −2.401 |
| | (2.837) | (4.327) | (3.983) | (5.006) | (2.981) |
| Observations | 2250 | 1172 | 1141 | 1116 | 1042 |
| McFadden pseudo $R^2$ | 0.038 | 0.038 | 0.036 | 0.095 | 0.044 |
| (Equivalent linear model $R^2$) | (0.076) | (0.076) | (0.073) | (0.209) | (0.088) |

*Notes.* Robust standard errors are shown in the parentheses. Controls not shown are age group, gender, arrival mode, and chief complaint code. ***p<0.001; **p<0.01; *p<0.05

In summary, our empirical investigation shows that choosing the next patient is a complex decision that depends on both clinical and operational factors at EDs. When the risk of ED blocking is sufficiently low, clinical factors such as patients' triage levels and waiting times dominate. When the risk of ED blocking increases, the operational factor kicks in, and the chance that discharge patients get high priority increases. Our empirical results provide evidence that it is possible that ED decision makers take resource constraints into their patient prioritization decisions, in contrary to the common perception that patient priority assignment is a clinical decision made during triage. In addition, physicians may behave differently due to the lack of operational guidelines.

### 3.3. Robustness Check

To show the robustness of our findings, we estimate model specifications that deviate from the baseline model, i.e., Model 1. In the base model, the disposition of a patient is predicted by a logit model using six basic patient characteristics. In reality, decision makers can infer a patient's disposition based on more detailed patient information and their experiences, both of which are not accounted in our logit model. As a robustness check, we use the actual disposition as a proxy for the estimated disposition while being aware of the potential endogeneity issue discussed in Section 3.1. We also tested two cases, replacing the logit model by two other binary classifiers, namely, the probit model and CART, to make sure that our findings do not depend on any specific classifier.

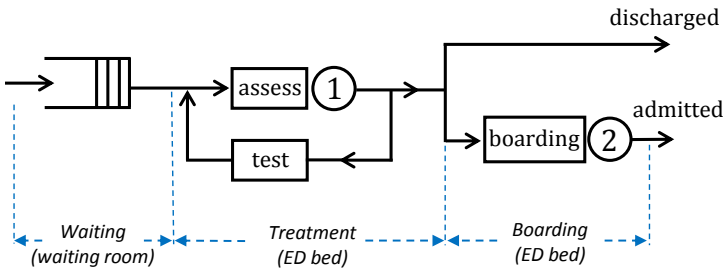
At EDs, triage nurses can request certain tests for patients whose conditions satisfy pre-set protocols. This practice is called *triage nurse ordering*, and its impact has been studied in the literature (e.g., Batt and Terwiesch 2016). In our study ED, nurses are allowed to order lab tests but not imaging 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 do not observe whether a patient's test results are available or not when the patient is selected for assessment. Hence, we remove all visit records with triage orders and repeat our study to check the robustness our results.

We also deviate from Model 1 in the following three aspects: (i) control another layer of patient characteristics, namely, the chief complaint codes; (ii) consider the nonlinear effects of *WaitTime* on the utility of choosing a patient (Ding et al. 2019) by adding a quadratic term to the model, and (iii) consider two different ways to proxy the ED blocking level (see discussion in Section 2.3.2). As a result, we estimated a total of 19 models. Columns labeled Model 2, 3, 4, and 5 in Table 2 show the results for models deviating from Model 1 by using the true disposition, by controlling patient's chief complaint code, by removing patients with triage orders, and by using the second proxy for ED blocking level, respectively. The results for other model specifications are qualitatively similar to those reported in this paper and are deferred to Appendix A.

## 4. The Rationale Behind the Prioritization Behaviors

In Section 3, using patient visit records data, we examined ED decision makers' behavior on patient prioritization and found that discharge patients are prioritized when ED blocking level is high. In this section, we build a stylized MDP model to further understand the rationale behind such behavior. The focus of our model is to understand how disposition affects a patient's priority under the pressure of ED blocking, especially for patients of the same triage level, rather than to explicitly model the complex ED operations in details. Hence, we believe that this stylized model can capture the key characteristics of interest.

**Figure 3 A diagram of patient flow in emergency departments.**




### 4.1. Model of ED Patient Flow and ED Blocking

We model the ED patient flow process during peak load hours using a two-station tandem queue with feedback at station 1. A schematic depiction of patient flow is as follows (Figure 3): After triage, patients wait in the waiting room until chosen for service at station 1 by physicians. The service at station 1 corresponds to physician assessment. After physician assessment, patients join the test queue before returning to the assess queue with a constant probability; otherwise, the treatment process is completed, and then patients are 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, while being aware that the boarding process is highly time-dependent in reality (Shi et al. 2015). We relax it in the simulation. We assume that during peak load hours, there are always patients waiting to be seen in the waiting room. We also assume that patients are only different in their dispositions and identical in other characteristics such as triage level, demographics, waiting time, complaint, etc., due to that this model focuses on how dispositions will affect the order of patients being seen. However, we will relax all these assumptions and take time-dependent arrival process, different triage levels and demographics into account in our simulation model in Section 5.

It is well known that ED physicians are multitasking, however, they generally do not take on more patients than their maximum service capacity (Campello et al. 2016) due to safety concern and that it can be counter-productive (KC 2013). Let $C$ denote the maximum service capacity. Although it is hard to imagine that

physicians keep a fixed number in mind as their capacity, Saghafian et al. (2012) observe that $C$ is generally no more than 7 in their study ED. Let $B$ be the ED bed capacity, and $x$ be the number of boarding patients. Since boarding patients at station 2 need to stay in ED beds, we have $0 \le x \le B$. Under the assumption that physicians do not idle, the total number of patients in the assess and test queues[3], denoted by $K$, is $K \equiv \min\{C, B - x\}$. The expression of $K$ implies that both physicians and ED beds can be the bottleneck resource of ED. The phenomena of ED blocking refers to the situation that $C > B - x$, i.e., lack of ED beds starves the physician resource.

Assume that physicians do not idle during peak load hours unless their maximum service capacity is reached. Hence, whenever a patient's treatment at ED is completed, the patient will leave and a new patient will be selected for treatment from the waiting room. We model the patient flow dynamics of the treatment phase, i.e., the interaction between the assess and test queues, as a two-node cyclic network with population $K$—the total number of patients in the treatment phase. 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 at ED. We have the following result.

Proposition 1. *Suppose that there are multiple parallel servers providing services at both the assess and test queues, and the service times are assumed to be exponentially distributed. Then, the throughput rate $\mu_1(K)$ is increasing and concave in $K$. Specially, we have $\mu_1(0) = 0$.*

Proposition 1 implies that the rate at which a patient completes his treatment and exits ED increases with the number of patients whose diagnosis and treatment 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 servers will stay idle. However, the marginal increment decreases as the internal delays become longer due to that physicians and test centers have more patients to serve, and the reduction in server idleness decreases with more patients in the treatment phase.

### 4.2. 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 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 $\mu_1(K)$. Since $K = \min\{C, B - x\}$ where $x$ is the number of patients in the boarding queue and $0 \le x \le B$, we know from Proposition 1 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 and thus we have $\mu_1(B) = 0$.

[3] Patients from both queues count as physicians' workload as physicians are responsible for the care of all the patients whose diagnosis and treatment at ED are not yet finished.

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 estimated disposition of the patient, or she can choose patients in a FCFS manner, i.e., 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, $c_i \geq 0$, $i = 1, 2$. 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. As discussed in Assumption 1, decision makers can access real-time information of all patients waiting to be seen through the electronic system, including comprehensive triage data, however, they have to login to the system to access the data in order to assess patients' disposition. The negative utility also corresponds to a penalty for violating the rule of fairness. 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

$$g(\pi, x) \equiv \liminf_{t \to \infty} \frac{V_t(\pi, x)}{t}, \ \forall x \in \mathcal{S},$$

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 $\pi$, which is a sequence of decision rules that map from $\mathcal{S}$ to $\mathcal{A}$ and that specify the action taken at any state and time. Then, the optimal long-run average net social benefits is defined as

$$g^*(x) \equiv \sup_{\pi} \ g(\pi, x), \ \forall x \in \mathcal{S}.$$

Next, 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 $\mu_2$ become, respectively, the probability that the next uniformized transition is a service completion at station 1 and 2. With probability $(\mu_1(0) - \mu_1(x))/\Lambda$, there is an artificial transition and the system state remains unchanged. 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

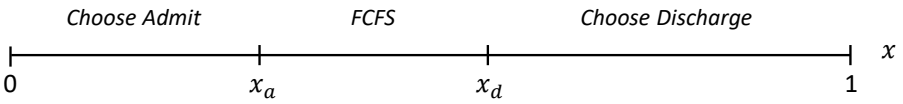
as $v(x) + g = Tv(x),\ \forall x \in \mathcal{S}$, where $g$ is the optimal average net social benefits per period of time after uniformization, and the operator $T$ is defined as:

$$\begin{aligned} Tv(x) = \mu_1(x) \max\left\{R_1 + v(x) - c_1, R_2 + v(x+1) - c_2, R_0 + (1-p)v(x) + pv(x+1)\right\} \\ + \mu_2 v(x-1) + [\mu_1(0) - \mu_1(x)]\, v(x),\ \text{if } x \in \mathcal{S} \setminus \{0\}, \\ Tv(0) = \mu_1(0) \max\left\{R_1 + v(0) - c_1, R_2 + v(1) - c_2, R_0 + (1-p)v(0) + pv(1)\right\} + \mu_2 v(0), \end{aligned}$$

where $v(\cdot) : \mathcal{S} \to \mathbb{R}$. Each term in the maximization is the cost-to-go if the corresponding action in $\mathcal{A}$ is taken. We assume that $v(x) = -\infty$ for $x \notin \mathcal{S}$ for simplicity. We are now ready to present Proposition 2, which establishes the existence and structural properties of the optimal policy. (The proof is provided in Appendix B.)

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 $x_a \equiv \max\{x : x \in \mathcal{S},\ v(x) - v(x+1) \le R_2 - R_1 - (1-p)^{-1}c_2\}$ and $x_d \equiv \min\{x : x \in \mathcal{S},\ v(x) - v(x+1) \ge R_2 - R_1 + p^{-1}c_1\}$.*

**Figure 4 Visual depiction of the optimal policy.**




Proposition 2 states that the optimal prioritization policy is of threshold-type, characterized by two points $x_a$ and $x_d$. Note that $\frac{x}{B}$ is the percentage of ED beds occupied by boarding patients, which is an indicator of the level of ED blocking. A bigger $x$ represents a higher 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 operational factor 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 clinical factor when ED is facing the pressure of being blocked; when the blocking level is intermediate, it is optimal to choose the first patient in line, i.e., follow the FCFS rule. Figure 4 demonstrates this threshold structure by a numerical example.

**Connecting to the Empirical Findings: A Discussion**

The optimal policy of our MDP model helps explain the rationale behind ED decision makers' prioritization behavior. We connect Proposition 2 to the empirical finding for each triage level as follows.

Triage level 2 patients are in general very urgent and their conditions can potentially deteriorate fast. Among them, we believe admit patients are more urgent than discharge patients. Hence, serving admit patients generates higher social benefits than serving discharge patients, i.e., $R_1 > R_2$. Thus, admit patients are prioritized when $x < x_a$, i.e., when ED bed is not the bottleneck resource. This is when the clinical factor drives the prioritization decision.

Triage levels 3 and 4 patients are less urgent and can wait for some time without significantly compromising their care. Especially, the difference in urgency between admit and discharge patients is not very significant, i.e., $R_1 \simeq R_2$, and $x_a \leq 0$. Hence, within triage level 3 or 4, patients are seen in a FCFS manner regardless of their disposition when ED blocking level is sufficiently low ($x < x_d$).

When ED blocking level is sufficiently high, i.e., when over $\frac{x_d}{B} \times 100$ percent of ED treatment beds are already occupied by boarding patients, ED decision makers may find it difficult to find an available bed to treat new patients. That is when the ED bed capacity is taken into consideration in the decision of which patient to see next. Intuitively, it is better to start prioritizing discharge patients over admit patients. Otherwise, another ED bed may be occupied for a prolonged period of time by a boarding patient, which further reduces ED's treatment capacity and makes patients in the waiting room wait longer. We observe this behavior for all three triage levels studied.

In short, when ED bed is not the resource bottleneck, decision makers prioritize admit patients within CTAS 2 and follow FCFS within CTAS 3&4; on the other hand, when ED bed becomes the bottleneck resource, i.e., ED blocking level is sufficiently high, decision makers start to prioritize discharge patients over admit patients across all three triage levels, to keep ED from further being blocked by boarding patients.

## 5. Impact of Patient Prioritization on Operational Performances

In previous sections, we have examined ED decision markers' prioritization behavior and the rationale behind it. What remains unclear is the impact of such behavior on ED operational performances. To our best knowledge, no guideline exists for ED patient prioritization. Hence, it is expected that decision makers may exhibit heterogeneous behaviors, which makes it difficult to examine the impact through observational data. In this section, we develop a discrete-event simulation model calibrated using real data to evaluate the impact. Using the simulation model, we compare a prioritization policy that considers ED blocking, inspired by ED decision markers' prioritization behavior, with a policy that does not, and quantify the impact on average patient 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 ED resource availability into account, at least qualitatively.

### 5.1. Simulation Design

We simulate the ED patient flow process as a two-station tandem queue as described in Section 4.1. We relax the assumption on the arrival process, and assume that patients arrive to the ED according to a non-stationary Poisson process, which has been shown to be a reasonable assumption (Kim and Whitt 2014a,b). 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 (discharged), or get 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 priority assignment.

The input analysis of the simulation is based on our data collected between August 2014 and July 2015. We observe that the arrival process shows a daily pattern. Moreover, medical staff's shift schedules (especially physicians) repeat itself every 24 hours and remain unchanged during this period. Hence, we choose 24 hours as a cycle and treat the data of each day from August 2014 to July 2015 as a realization of an underlying stochastic process. To generate patient arrivals to the ED, we first estimate the inter-arrival times dependent on time of day 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 to the ED during this hour (bootstrap), and assign this patient's profile (triage level, disposition, etc.) to the new generated patient. We generate service times (e.g., treatment/boarding times) for this patient in a similar manner: first, group the treatment times and boarding times from the data by hour of the day, 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 hour of the day is known to us. We assume that physicians are multitasking and have the same service capacity. We set the capacity to be 7 patients and perturb it to test the robustness of the simulation results. It is however challenging to estimate the capacity of ED beds, as it depends on both the numbers of physical beds and nursing staff, which makes it time-dependent. We use the 90th percentile of the number of patients staying in ED beds at each hour of the day over the two-year horizon—round it to the nearest integer—to approximate the bed capacity at this particular hour of the day. Using the 90th percentile instead of the maximum avoids outliers due to data collection errors or temporary increase of ED capacity.

### 5.2. Prioritization Policies

Next, we compare patient prioritization policies that consider (or not) ED blocking by simulation. In the simulation, we adopt the second measure of ED blocking level (see Section 2.3.2), i.e., the ratio of the number of boarding patients over the number of "extra" beds (total bed capacity net of physician capacity). We start by defining policies of interest.

***Urgency-Based Priority Policy (UP):*** A non-idling policy, under which class 1 patients receive priority over all 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 patients; FCFS is followed within class 1. For class 2, prioritize admit patients if blocking level is less than $\eta_1$ ($\geq 0$), prioritize discharge patients if ED blocking level is greater than $\eta_2$ ($\geq \eta_1$), 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 ED blocking level is less than $\eta_A$ ($\geq 0$) 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 ED blocking level is greater than $\eta_D$ ($\geq 0$) 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 ED blocking level is greater than $\eta_{AD}$ and prioritize admit patients otherwise, i.e., $\eta_1 = \eta_2 = \eta_{AD}$.

The Urgency-Based Priority Policy (Policy UP) is clinical driven, under which patients are seen mainly based on their urgency. Among patients of the same urgency, they are seen in their order of arrival, so that fairness is exercised (to certain extent). Policy CUP is inspired by the insights from the structure of the optimal policy of the MDP model in Section 4, which prioritizes patients of different dispositions or follows FCFS depending on the congestion level of ED resource. We compare it with Policy UP by varying the two thresholds $\eta_1$ and $\eta_2$ in a wide range. In addition, we consider three policies that are of special forms of Policy CUP, namely, Policies CUP-A, CUP-D and CUP-AD, as defined above. We note that Policy CUP-AD shares similarity with the empirical policy used for the prioritization within CTAS 2 patients in our study ED, while Policy CUP-D aligns with the one for CTAS 3 & 4 patients. All three policies are simpler and easier to implement than Policy CUP. We compare them with Policy UP by varying the corresponding parameters in a broad range and quantify their impact on patient average waiting time and LOS.

### 5.3. Comparison Results

We simulate each policy for 10 replications, each with 907 days. For each replication, the first 7 days serve as the warm-up period thus are removed. We then take the first 7 days in every 9 days of the remaining 900 days as a "batch" and calculate the average waiting time and LOS using all patients who arrive and leave during the "batch" period. Hence, we have total 1000 pairs of average waiting time and LOS for each policy. We compare the performance of Policy UP with each of the four policies that have taken patient disposition into the prioritization decision, and report the percentage reduction in both average waiting time and LOS.

**Figure 5 The 95% confidence interval for the percentage reduction in long-run average waiting time and LOS by using Policy *CUP* over Policy *UP* for admit patients, discharge patients, and all patients, respectively, when $\eta_1 \in \{0.4, 0.6, 0.8\}$ and $\eta_2 \in \{0.9, 1.0, 1.1, 1.2, 1.3\}$.**

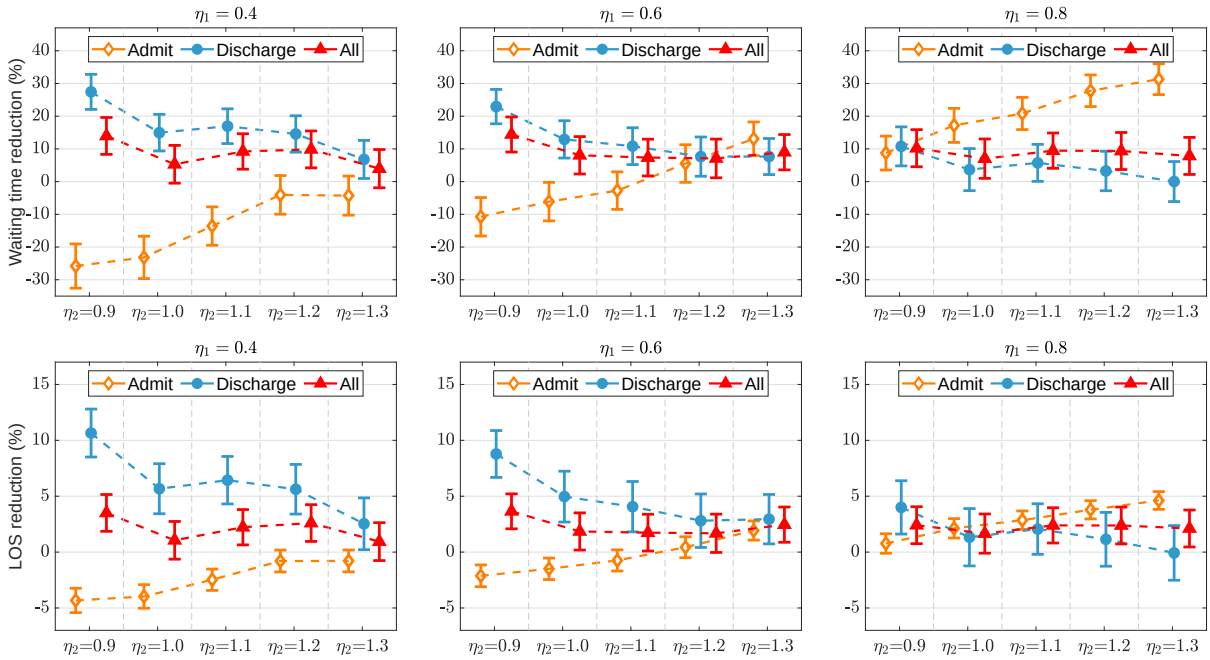



Figure 5 shows the 95% confidence interval for the percentage reduction in the expected long-run average waiting time and LOS by using Policy CUP over Policy UP for 15 combinations of $\eta_1$ and $\eta_2$, where $\eta_1 \in \{0.4, 0.6, 0.8\}$ and $\eta_2 \in \{0.9, 1.0, 1.1, 1.2, 1.3\}$. Our first observation is that prioritizing patients based on disposition and ED blocking level can reduce long-run average waiting time of all patients, by as much as 15% in some scenarios. It is intuitive that the more we prioritize discharge patients (smaller $\eta_2$), the greater the percentage reduction in average waiting time for discharge patients. However, it comes at the cost of more waiting for admit patients. As can be seen, the average waiting time of for admit patients increases (negative

reduction) more when a larger percentage of discharge patients are prioritized. Moreover, we observe that discharge patients have a bigger impact on the overall performance than admit patients due to that the former are four times as many as the latter. We observe similar patterns for the impact on average patient LOS with smaller magnitude in reduction. That is due to that the priority policy only impacts patient waiting time but not their treatment time, hence, the reduction in average patient LOS comes from the reduction in patient waiting time.

There are scenarios that *the average waiting times for patients of both dispositions decrease*. See, e.g., $\eta_1 = 0.6$ and $\eta_2 = 1.3$, $\eta_1 = 0.8$ and $\eta_2 = 0.9$, although these are not the scenarios with the greatest percentage reduction. It might seem counter-intuitive at first glance. We argue that this is entirely possible should critical ED resources be appropriately rationed in a highly time-varying supply and demand environment, where both physicians and ED beds can be the bottleneck for patient flow. When physician is the bottleneck, i.e., there are empty beds hence ED blocking level is relatively low, prioritizing admit patients gets patients into boarding and ready for admission earlier. It increases the utilization of ED bed while not compromising physician's treatment capability. Otherwise if they were treated when few beds are available, they would stay in beds for boarding which only aggravates the level of ED blocking. On the other hand, when ED bed becomes the bottleneck, discharge patients are prioritized so that physicians will not be forced to stay idle due to the lack of beds. Hence, a sensible allocation of critical ED resources through patient prioritization can achieve significant improvement for all patients.

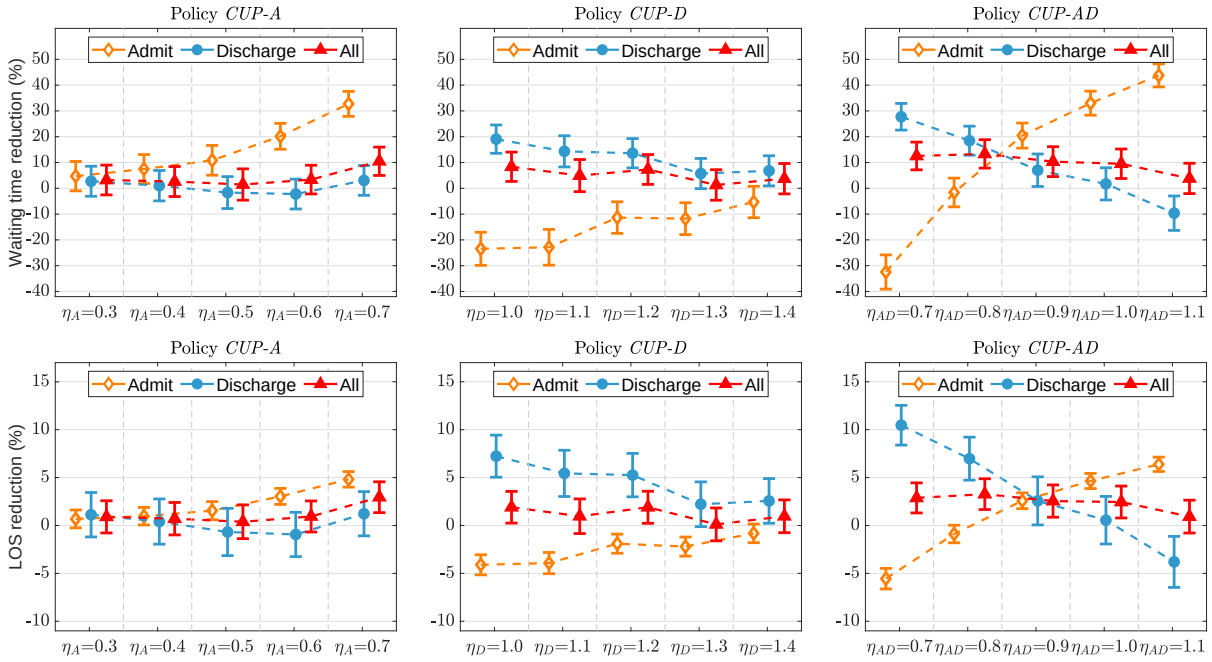
The results of comparisons between Policy UP and the three simpler priority policies are shown in Figure 6, where the threshold for each policy is chosen from $\eta_A \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, $\eta_D \in \{1.0, 1.1, 1.2, 1.3, 1.4\}$ and $\eta_{AD} \in \{0.7, 0.8, 0.9, 1.0, 1.1\}$. The results are similar to that of Policy CUP in that the average waiting time of all patients is reduced, however, the improvements of all three policies are less compared to Policy CUP. 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.

## 6. Impact on Waiting Time Prediction

In this section, we demonstrate how our findings can help improve the accuracy of existing prediction methods, based on a better understanding of ED decision makers' patient prioritization behaviors.

A growing number of hospitals (our study hospital included) have started posting their predicted ED waiting time through their websites, smartphone apps, billboards, and screens within hospitals. An accurate prediction of waiting time can increase the coordination in hospital networks thereby reduce overall ED waiting (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 research/management

**Figure 6** **The 95% confidence interval for the percentage reduction in long-run average waiting time and LOS by using Policy *CUP-A, CUP-D, CUP-AD* over Policy *UP* for admit patients, discharge patients, and all patients, respectively, when $\eta_A \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, $\eta_D \in \{1.0, 1.1, 1.2, 1.3, 1.4\}$ and $\eta_{AD} \in \{0.7, 0.8, 0.9, 1.0, 1.1\}$.**




communities. Nevertheless, the accuracy of prediction models has drawn 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 progress on waiting time prediction are documented in Sun et al. (2012) and Ang et al. (2015). The first work developed a quantile regression model to predict the ED waiting time. The 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), Q-Lasso also includes other variables such as local weather information, flu trend, the number of providers and 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) but 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 mean squared errors (MSE)*. We then add the three interaction terms in Equation (2) into the prediction models in Sun et al. (2012) and Ang

et al. (2015), compare with models without them, and report the percentage reduction in MSE. The predictor *Disposition* is estimated from a logit model.

**Table 4 Percentage Reduction in Mean Squared Errors over Two Existing Waiting Time Prediction Models**

| | All patients | *BlockLevel* <50th percentile | *BlockLevel* >50th percentile | *BlockLevel* >75th percentile | *BlockLevel* >90th percentile |
|---|---|---|---|---|---|
| Quantile Regression | 3.48% | 1.07% | 5.05% | 5.79% | 6.72% |
| Q-Lasso | 2.90% | 1.83% | 3.73% | 4.69% | 7.20% |

We choose 80% of our data for model training and remaining 20% for testing. The percentage reductions in MSE by adding the three interaction terms in Equation (2) are shown in Table 4. When we explicitly take decision makers' prioritization behavior into the prediction models and apply them on all patients in our dataset, the reductions in MSE are statistically significant for both the quantile regression model and Q-Lasso (3.48% and 2.90%, respectively). We also test the performance on four subsets of the data, which contain patients who arrive at ED and find ED blocking level below the 50th percentile, above the 50th percentile, above the 75th percentile, and above 90th percentile, respectively. The results show that when the ED blocking level is relatively low (*BlockLevel* < 50th percentile), the prediction accuracy improves little by incorporating patient prioritization to quantile regression or Q-Lasso. On the other hand, when the ED block level is high, the MSE can be reduced by up to 6.72% and 7.20% for quantile regression and Q-Lasso, respectively. Note that we would expect greater improvement if ED decision makers behave homogeneously.

## 7. Conclusions

Motivated by an intriguing observation from comparing the average waiting times of admit and discharge patients by triage levels, we study ED decision makers' prioritization behavior in choosing the next patient for treatment. Using data from a large teaching hospital in Canada, we find that decision makers apply urgency-specific delay-dependent prioritization. Moreover, decision makers start to prioritize discharge patients to avoid ED being further blocked when ED blocking level is sufficiently high. We then draw insights from a stylized MDP formulation to explain the rationale behind such prioritization behaviors. To the best of our knowledge, it has not been documented that medical workers take ED crowding and blocking level into patient prioritization decisions. Our work fills in the gap by providing empirical evidence and explain the rationale behind. We then perform a simulation to study the impact of such behaviors on ED operational performances, and show that priority policies—derived from our empirical findings and insights from our 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 in emergency departments.

In fact, the medical society has long understood the root cause of ED overcrowding—the prolonged occupation of ED beds by boarding patients. The Canadian Association of Emergency Physicians (CAEP) pointed out the following in their position statement (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 the *boarding patients* in our paper.) With this context, it seems an intuitive and sensible decision for decision makers (not necessarily all) to prioritize discharge patients; otherwise, treating an admit patient leads to another bed being occupied for a prolonged period of time, which only aggravates ED overcrowding. Our discussion with ED physicians confirmed this intuition.

Our analysis and results are based on patient data at one hospital. Therefore, the findings may not extend to hospitals of different sizes or hospitals where ED beds are not the bottleneck resource. The insights from the MDP model give us more confidence that if both ED physicians and beds can be the bottleneck resource, which is the case in many hospitals (Affleck et al. 2013), it is plausible that decision makers may prioritize discharge patients when ED blocking level is high. Nevertheless, it would be valuable to conduct similar analysis using data from other hospitals. Two other issues that are left out of this paper are patient safety and ethical concerns due to the prioritization (and de-prioritization) of a certain group of patients. They deviate from the focus of this paper, but they are certainly important questions that need to be answered for the results to be implementable. 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.

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. S. 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, 2016.

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.

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.

D. C. Chan. The efficiency of slacking off: Evidence from the emergency department. *Econometrica*, 86(3):997–1030, 2018.

M. Dahlberg and M. Eklöf. *Relaxing the IIA assumption in locational choice models: a comparison between conditional logit, mixed logit, and multinomial probit models*. Nationalekonomiska institutionen, 2003.

T. Dai and S. Tayur. Healthcare operations management: A snapshot of emerging research. *Manufacturing & Service Operations Management*, 2019.

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.

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

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

J. J. Heckman. Dummy endogenous variables in a simultaneous equation system, 1977.

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.

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.

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

D. S. 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. S. 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.

S.-H. Kim and W. Whitt. Choosing arrival process models for service systems: Tests of a nonhomogeneous poisson process. *Naval Research Logistics*, 61(1):66–90, 2014a.

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, 2014b.

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, to appear*, 2019.

L. Kleinrock. A delay dependent queue discipline. *Naval Research Logistics Quarterly*, 11(3-4):329–341, 1964.

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.

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.

G. S. Maddala. *Limited-dependent and qualitative variables in econometrics*. Cambridge university press, Cambridge, UK, 1983.

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.

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.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 2005.

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, 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.

J. G. Shanthikumar and D. D. Yao. Stochastic comparisons in closed jackson networks. In M. Shaked and J. G. Shanthikumar, editors, *Stochastic orders and their applications*, chapter 14, pages 433–460. Academic Press, New York, 1994.

A. B. Sharif, D. A. Stanford, P. Taylor, and I. Ziedins. A multi-class multi-server accumulating priority queue with application to health care. *Operations Research for Health Care*, 3(2):73–79, 2014.

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.

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, 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.

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.

K. Xu and C. W. Chan. Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management*, 18(3):314–331, 2016.

# Appendices

## Appendix A. Results for Robustness Check

We test the robustness of our findings by considering the following model specifications: (i) with or without visit records of patients with triage orders; (ii) with or without controlling another layer of patient characteristics, namely, the chief complaint codes; (iii) with or without a quadratic term of *WaitTime* to the model, and (iv) consider two different ways to proxy the ED blocking level; and (v) different proxies of the disposition of a patient, such as prediction by a logit model, using the actual disposition from data, etc. The specification and $R^2$ of each model are shown in Table 5. The results of Models 1–5 are provided in Table 2, and that of Models 5–19 in Tables 6 and 7. Controls not shown in the tables are age group, gender, arrival mode, chief complaint code, and the quadratic term of *WaitTime* (only applies to model specifications that contain it). For all models from different specifications, the regression results are remarkably consistent, which provides strong evidence that our findings are robust.

**Table 5 Specifications for each of the 19 robustness checks**

| Model ID | With Triage Orders | With Chief Complaint | With $WaitTime^2$ | Blocking Level | Classification Method | Number of Observations | McFadden pseudo-$R^2$ | Equivalent $R^2$ |
|---|---|---|---|---|---|---|---|---|
| 1 | yes | no | no | proxy 1 | logit | 68,341 | 0.047 | 0.095 |
| 2 | yes | no | no | proxy 1 | true | 68,341 | 0.046 | 0.093 |
| 3 | yes | yes | no | proxy 1 | logit | 68,341 | 0.058 | 0.124 |
| 4 | no | no | no | proxy 1 | logit | 40,012 | 0.051 | 0.105 |
| 5 | yes | no | no | proxy 2 | logit | 68,341 | 0.047 | 0.095 |
| 6 | yes | no | yes | proxy 1 | logit | 68,341 | 0.064 | 0.140 |
| 7 | yes | no | yes | proxy 2 | logit | 68,341 | 0.064 | 0.140 |
| 8 | yes | yes | no | proxy 2 | logit | 68,341 | 0.058 | 0.124 |
| 9 | yes | yes | yes | proxy 1 | logit | 68,341 | 0.077 | 0.171 |
| 10 | yes | yes | yes | proxy 2 | logit | 68,341 | 0.077 | 0.171 |
| 11 | no | no | no | proxy 2 | logit | 40,012 | 0.051 | 0.105 |
| 12 | no | no | yes | proxy 1 | logit | 40,012 | 0.043 | 0.088 |
| 13 | no | no | yes | proxy 2 | logit | 40,012 | 0.044 | 0.088 |
| 14 | no | yes | no | proxy 1 | logit | 40,012 | 0.068 | 0.149 |
| 15 | no | yes | no | proxy 2 | logit | 40,012 | 0.068 | 0.149 |
| 16 | no | yes | yes | proxy 1 | logit | 40,012 | 0.065 | 0.142 |
| 17 | no | yes | yes | proxy 2 | logit | 40,012 | 0.064 | 0.140 |
| 18 | yes | no | no | proxy 1 | CART | 68,341 | 0.047 | 0.095 |
| 19 | yes | no | no | proxy 1 | Probit | 68,341 | 0.047 | 0.095 |

## Appendix B. Proofs of Propositions 1 and 2.

**Proof of Proposition 1:** Denote the number of servers and number of patients at the assess and test queues by $n_i$ and $x_i$, respectively, where $i = 1$ corresponds to the assess queue, and $i = 2$ corresponds to the test queue. Denote the service rate of each server at the assess and test queues by $\lambda^1$ and $\lambda^2$, respectively. Since the assess

**Table 6 Estimation results for Models 6–15 in Table 5.**

| | Model 6 | Model 7 | Model 8 | Model 9 | Model 10 |
|---|---|---|---|---|---|
| *Triage Level* = 2 | | | | | |
| *CTAS*×*WaitTime* | 0.011*** | 0.011*** | 0.005*** | 0.013*** | 0.013*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *CTAS*×*Disposition* | 0.699*** | 0.588*** | 0.721*** | 0.872*** | 0.748*** |
| | (0.061) | (0.057) | (0.082) | (0.084) | (0.080) |
| *CTAS*×*Disposition*×*BlockLevel* | −2.016*** | −0.546*** | −0.532*** | −2.337*** | −0.632*** |
| | (0.150) | (0.043) | (0.044) | (0.157) | (0.046) |
| *Triage Level*=3 | | | | | |
| *CTAS* | −0.509*** | −0.512*** | −0.4*** | −0.508*** | −0.513*** |
| | (0.021) | (0.021) | (0.023) | (0.024) | (0.024) |
| *CTAS*×*WaitTime* | 0.017*** | 0.017*** | 0.009*** | 0.019*** | 0.019*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *CTAS*×*Disposition* | −0.073 | −0.38*** | −0.219* | 0.158 | −0.161 |
| | (0.075) | (0.069) | (0.099) | (0.098) | (0.094) |
| *CTAS*×*Disposition*×*BlockLevel* | −2.496*** | −0.504*** | −0.503*** | −2.71*** | −0.556*** |
| | (0.187) | (0.054) | (0.058) | (0.192) | (0.055) |
| *Triage Level*=4 | | | | | |
| *CTAS* | −0.603*** | −0.615*** | −0.507*** | −0.639*** | −0.653*** |
| | (0.029) | (0.029) | (0.031) | (0.032) | (0.032) |
| *CTAS*×*WaitTime* | 0.018*** | 0.018*** | 0.008*** | 0.02*** | 0.02*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *CTAS*×*Disposition* | −0.288 | −0.857*** | −0.517* | 0.217 | −0.377 |
| | (0.190) | (0.176) | (0.220) | (0.213) | (0.197) |
| *CTAS*×*Disposition*×*BlockLevel* | −3.366*** | −0.478** | −0.561*** | −3.753*** | −0.541*** |
| | (0.574) | (0.159) | (0.169) | (0.574) | (0.152) |
| | Model 11 | Model 12 | Model 13 | Model 14 | Model 15 |
| *Triage Level* = 2 | | | | | |
| *CTAS*×*WaitTime* | 0.004*** | 0.004*** | 0.004*** | 0.005*** | 0.005*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *CTAS*×*Disposition* | 0.692*** | 1.095*** | 0.858*** | 1.093*** | 0.829*** |
| | (0.085) | (0.101) | (0.090) | (0.128) | (0.120) |
| *CTAS*×*Disposition*×*BlockLevel* | −0.456*** | −2.298*** | −0.539*** | −2.37*** | −0.522*** |
| | (0.057) | (0.226) | (0.061) | (0.217) | (0.059) |
| *Triage Level*=3 | | | | | |
| *CTAS* | −0.404*** | −0.656*** | −0.675*** | −0.384*** | −0.39*** |
| | (0.027) | (0.031) | (0.031) | (0.030) | (0.030) |
| *CTAS*×*WaitTime* | 0.008*** | 0.016*** | 0.016*** | 0.01*** | 0.009*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *CTAS*×*Disposition* | −0.657*** | −0.51*** | −0.859*** | −0.176 | −0.564*** |
| | (0.108) | (0.133) | (0.119) | (0.156) | (0.147) |
| *CTAS*×*Disposition*×*BlockLevel* | −0.363*** | −2.795*** | −0.529*** | −2.403*** | −0.402*** |
| | (0.080) | (0.330) | (0.091) | (0.297) | (0.083) |
| *Triage Level*=4 | | | | | |
| *CTAS* | −0.571*** | −4.276*** | −3.992*** | −0.59*** | −0.602*** |
| | (0.035) | (0.124) | (0.115) | (0.040) | (0.040) |
| *CTAS*×*WaitTime* | 0.009*** | 0.038*** | 0.037*** | 0.01*** | 0.01*** |
| | (0.000) | (0.001) | (0.001) | (0.000) | (0.000) |
| *CTAS*×*Disposition* | −1.393*** | −1.497** | −0.074 | −0.08 | −0.87** |
| | (0.269) | (0.570) | (0.438) | (0.344) | (0.317) |
| *CTAS*×*Disposition*×*BlockLevel* | −0.216 | −13.669*** | −6.344*** | −3.871*** | −0.395 |
| | (0.238) | (1.484) | (0.534) | (0.921) | (0.250) |

***p<0.001; **p<0.01; *p<0.05

**Table 7 Estimation results for Models 16–19 in Table 5.**

| | Model 16 | Model 17 | Model 18 | Model 19 |
|---|---|---|---|---|
| *Triage Level* = 2 | | | | |
| *CTAS*×*WaitTime* | 0.007*** | 0.008*** | 0.004*** | 0.004*** |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| *CTAS*×*Disposition* | 1.216*** | 0.958*** | 1.235*** | 0.741*** |
| | (0.140) | (0.135) | (0.094) | (0.064) |
| *CTAS*×*Disposition*×*BlockLevel* | −2.655*** | −0.625*** | −2.628*** | −1.96*** |
| | (0.236) | (0.066) | (0.220) | (0.156) |
| *Triage Level*=3 | | | | |
| *CTAS* | −0.684*** | −0.683*** | −0.306*** | −0.384*** |
| | (0.035) | (0.036) | (0.023) | (0.021) |
| *CTAS*×*WaitTime* | 0.019*** | 0.019*** | 0.008*** | 0.008*** |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| *CTAS*×*Disposition* | −0.541** | −1.014*** | −0.064 | −0.168* |
| | (0.174) | (0.169) | (0.097) | (0.081) |
| *CTAS*×*Disposition*×*BlockLevel* | −2.794*** | −0.522*** | −2.727*** | −2.388*** |
| | (0.330) | (0.095) | (0.231) | (0.203) |
| *Triage Level*=4 | | | | |
| *CTAS* | −2.947*** | −3.073*** | −0.26*** | −0.454*** |
| | (0.087) | (0.090) | (0.034) | (0.028) |
| *CTAS*×*WaitTime* | 0.034*** | 0.035*** | 0.007*** | 0.007*** |
| | (0.001) | (0.001) | (0.000) | (0.000) |
| *CTAS*×*Disposition* | −0.19 | −2.47*** | −0.691*** | −0.317 |
| | (0.519) | (0.523) | (0.167) | (0.202) |
| *CTAS*×*Disposition*×*BlockLevel* | −11.727*** | −1.847*** | −2.266*** | −3.636*** |
| | (1.346) | (0.346) | (0.414) | (0.592) |

***p<0.001; **p<0.01; *p<0.05

and test queues form a cyclic network with $K$ patients, we have $x_1 + x_2 = K$ and $0 \le x_i \le K,\ i = 1,2$. Thus, the system state can be fully described by $x_2$. When $x_2 = i$, the service times at the assess and test queues are therefore exponentially distributed with rates $\lambda_i^1 = \min\{n_1, K-i\}\lambda^1$ and $\lambda_i^2 = \min\{n_2, i\}\lambda^2$, respectively. Note that both $\lambda_i^1$ and $\lambda_i^2$ are increasing and concave in the number of patients at the assess and test queues, respectively. Note also that this network structure holds for any positive integer number of servers at each queue, i.e., $1 \le n_i \le \infty,\ i = 1,2$. We can use a continuous-time Markov chain to model the dynamic of the cyclic network and the balance equations can be written as $p\lambda_i^1\pi_i = \lambda_{i+1}^2\pi_{i+1},\ i = 0,1,\cdots,K-1$. where $\{\pi_i,\ i = 0,1,\cdots,K\}$ is the stationary distribution. Together with $\sum_{i=0}^{K}\pi_i = 1$, we can solve the system of linear equations and get

$$\pi_0 = \frac{1}{1+\sum_{k=1}^{K}\prod_{j=0}^{k-1}\rho_j},\ \ \pi_i = \frac{\prod_{j=0}^{i-1}\rho_j}{1+\sum_{k=1}^{K}\prod_{j=0}^{k-1}\rho_j},\ i = 1,2,\cdots,K, \tag{3}$$

where $\rho_j \equiv p\lambda_j^1/\lambda_{j+1}^2,\ j = 0,1,\cdots,K-1$. We are interested in the rate that patients complete their treatment and exit ED (either discharged home or admitted to hospital wards), $\mu_1(K)$. It can be written as

$$\mu_1(K) = \sum_{i=0}^{K}(1-p)\lambda_i\pi_i = \frac{(1-p)\left[\lambda_0^1 + \sum_{k=1}^{K}\lambda_i^1\prod_{j=0}^{i-1}\rho_j\right]}{1+\sum_{k=1}^{K}\prod_{j=0}^{k-1}\rho_j}. \tag{4}$$

Denote the commutative number of service completions up to time $t$ from the assess queue by $D(K,t)$. We have

$$\lim_{t\to\infty}\frac{D(K,t)}{t} = \sum_{i=0}^{K} p\lambda_i\pi_i = \frac{1-p}{p}\mu_1(K) \tag{5}$$

From Theorem 14.D.1 of Shanthikumar and Yao (1994), we know that $D(K,t)$ is increasing and concave in $K$. Therefore, $\mu_1(K)$ is increasing and concave in $K$. It is obvious that $\mu_1(0) = 0$. □

Next, we prove Proposition 2 by the value iteration algorithm. We first study the discounted net social benefit version of the MDP. The uniformization constant $\Lambda = \mu_1(0) + \mu_2 + \alpha$ where $\alpha$ is the continuous discount factor. The optimality equations become $v = Tv$, where the format of the operator $T$ is the same as the one defined in the paper, and $v(x)$ is the total discounted net social benefits starting from state $x$. The fact that the state and action spaces are finite, one-period utility and costs are stationary and bounded, and that the discounting factor for the uniformized system $\alpha/(\alpha+\mu_1(0)) < 1$ implies that the maximum in the optimality equations $v(x) = Tv(x)$ is achieved and that there exists an optimal policy that is stationary and deterministic (see Chapter 6 in Puterman (2005)). This also implies that we may restrict our attention to this class of policies. We next define $Dv(x) \equiv v(x) - v(x+1),\ 0 \le x < B$, and then $Tv(x)$ simplifies into

$$\begin{aligned}
&Tv(B) = \mu_2 v(B-1) + \mu_1(0)v(B),\\
&Tv(x) = \mu_1(x)\max\{R_0 - pDv(x), R_1 - c_1, R_2 - c_2 - Dv(x)\} + \mu_2 v(x-1) + \mu_1(0)v(x),\ 0 < x < B,\\
&Tv(0) = \mu_1(0)\max\{R_0 - pDv(0), R_1 - c_1, R_2 - c_2 - Dv(0)\} + \mu_2 v(0) + \mu_1(0)v(0).
\end{aligned} \tag{6}$$

Let $\mathcal{F}$ be the set of functions defined on $S$ such that if $f(x) \in \mathcal{F}$, then (i) $Df(x) \ge 0$ for $0 \le x \le B-1$; (ii) $Df(x) \le Df(x+1)$ for $0 \le x \le B-2$. Then, we have the following result.

Lemma 1. *If $v \in \mathcal{F}$, then $Tv \in \mathcal{F}$.*

**Proof:** We first prove that $DTv(x) = Tv(x) - Tv(x+1) \ge 0$ for $0 \le x \le B-1$. Plug in Equation (6) into $DTv(x)$, we have

$$\begin{aligned}
DTv(x) = {}& \mu_1(x)\max\{R_0 - pDv(x), R_1 - c_1, R_2 - c_2 - Dv(x)\} - \mu_1(x+1)\max\{R_0 - pDv(x+1),\\
&R_1 - c_1, R_2 - c_2 - Dv(x+1)\} + \mu_2 Dv(x-1) + \mu_1(0)Dv(x),\ \text{if } 0 < x \le B-1,\\
DTv(0) = {}& \mu_1(0)\max\{R_0 - pDv(0), R_1 - c_1, R_2 - c_2 - Dv(0)\} - \mu_1(1)\max\{R_0 - pDv(1), R_1 - c_1,\\
&R_2 - c_2 - Dv(1)\} + \mu_1(0)Dv(0).
\end{aligned}$$

Since $Dv(x) \geq 0$ for any $0 \leq x \leq B-1$, we have

$$\begin{aligned} DTv(x) \geq \mu_1(x) \max\{R_0 - pDv(x), R_1 - c_1, R_2 - c_2 - Dv(x)\} + \mu_1(0)Dv(x) \\ - \mu_1(x+1) \max\{R_0 - pDv(x+1), R_1 - c_1, R_2 - c_2 - Dv(x+1)\}. \end{aligned} \tag{7}$$

We next consider six separate cases to show the right-hand side of (7) is non-negative for any $0 \leq x \leq B-1$.

*Case 1.* When $Dv(x) \leq Dv(x+1) \leq R_2 - R_1 - (1-p)^{-1}c_2$, we have

$$\begin{aligned} DTv(x) &\geq \mu_1(x)\left[R_2 - c_2 - Dv(x)\right] - \mu_1(x+1)\left[R_2 - c_2 - Dv(x+1)\right] + \mu_1(0)Dv(x) \\ &= D\mu_1(x)(R_2 - c_2) + \mu_1(x+1)Dv(x+1) + \left[\mu_1(0) - \mu_1(x)\right]Dv(x) \geq 0. \end{aligned}$$

The last inequality follows since $\mu_1(x)$ is concave and non-increasing in $x$ and $Dv(x) \geq 0$ for any $0 \leq x \leq B-1$.

*Case 2.* When $Dv(x) \leq R_2 - R_1 - (1-p)^{-1}c_2 \leq Dv(x+1) \leq R_2 - R_1 + p^{-1}c_1$, we have

$$\begin{aligned} DTv(x) &\geq \mu_1(x)\left[R_2 - c_2 - Dv(x)\right] - \mu_1(x+1)\left[R_0 - pDv(x+1)\right] + \mu_1(0)Dv(x) \\ &\geq (1-p)\mu_1(x)(R_2 - R_1 - (1-p)^{-1}c_2) + p\mu_1(x+1)(R_2 - R_1 - (1-p)^{-1}c_2) \geq \mu_1(x+1)Dv(x) \geq 0. \end{aligned}$$

The second and third inequalities follow from $\mu_1(x) \geq \mu_1(x+1)$, $Dv(x) \geq 0$ and $Dv(x+1) \geq R_2 - R_1 - (1-p)^{-1}c_2 \geq 0$; the last inequality follows from $\mu_1(x) \geq 0$ and $Dv(x) \geq 0$.

*Case 3.* When $Dv(x) \leq R_2 - R_1 - (1-p)^{-1}c_2 \leq R_2 - R_1 + p^{-1}c_1 \leq Dv(x+1)$, we have

$$\begin{aligned} DTv(x) &\geq \mu_1(x)\left[R_2 - c_2 - Dv(x)\right] - \mu_1(x+1)(R_1 - c_1) + \mu_1(0)Dv(x) \\ &\geq \mu_1(x+1)(R_2 - R_1 - c_2 + c_1) \geq \mu_1(x+1)[R_2 - R_1 - (1-p)^{-1}c_2 + c_1] \geq 0. \end{aligned}$$

The second inequality follows from the non-increasing property of $\mu_1(x)$ and the last inequality follows from $R_2 - R_1 - (1-p)^{-1}c_2 \geq Dv(x) \geq 0$.

*Case 4.* When $R_2 - R_1 - (1-p)^{-1}c_2 \leq Dv(x) \leq Dv(x+1) \leq R_2 - R_1 + p^{-1}c_1$, we have

$$\begin{aligned} DTv(x) &\geq \mu_1(x)\left[R_0 - pDv(x)\right] - \mu_1(x+1)\left[R_0 - pDv(x+1)\right] + \mu_1(0)Dv(x) \\ &= (\mu_1(x) - \mu_1(x+1))R_0 + \left[\mu_1(0) - p\mu_1(x)\right]Dv(x) + \mu_1(x+1)pDv(x+1) \geq 0. \end{aligned}$$

The last inequality follows from that $\mu_1(x)$ is non-increasing and $Dv(x) \geq 0$.

*Case 5.* When $R_2 - R_1 - (1-p)^{-1}c_2 \leq Dv(x) \leq R_2 - R_1 + p^{-1}c_1 \leq Dv(x+1)$, we have

$$\begin{aligned} DTv(x) &\geq \mu_1(x)\left[R_0 - pDv(x)\right] - \mu_1(x+1)\left(R_1 - c_1\right) + \mu_1(0)Dv(x) \\ &\geq \mu_1(x)R_0 - \mu_1(x+1)(R_1 - c_1) \geq \mu_1(x+1)p(R_2 - R_1 + p^{-1}c_1) \geq 0. \end{aligned}$$

The second and third inequalities follow from that $\mu_1(x)$ is non-increasing and $R_2 - R_1 + p^{-1}c_1 \geq 0$.

*Case 6.* When $R_2 - R_1 + p^{-1}c_1 \le Dv(x) \le Dv(x+1)$, we have

$$DTv(x) \ge \mu_1(x)(R_1 - c_1) - \mu_1(x+1)(R_1 - c_1) + \mu_1(0)Dv(x) \ge 0.$$

The last equality follows from that $\mu_1(x)$ is non-increasing. Hence, we conclude that $DTv(x) \ge 0$.

Next, we show that $DTv(x+1) \ge DTv(x)$ for $0 \le x \le B-2$. When $0 < x \le B-2$, we have

$$\begin{aligned}
&DTv(x+1) - DTv(x)\\
&= 2\mu_1(x+1)\max\{R_0 - pDv(x+1), R_1 - c_1, R_2 - c_2 - Dv(x+1)\}\\
&\quad - \mu_1(x)\max\{R_0 - pDv(x), R_1 - c_1, R_2 - c_2 - Dv(x)\} + \mu_1(0)[Dv(x+1) - Dv(x)]\\
&\quad - \mu_1(x+2)\max\{R_0 - pDv(x+2), R_1 - c_1, R_2 - c_2 - Dv(x+2)\} + \mu_2[Dv(x) - Dv(x-1)].
\end{aligned} \tag{8}$$

When $x = 0$, we have

$$\begin{aligned}
&DTv(x+1) - DTv(x) = 2\mu_1(1)\max\{R_0 - pDv(1), R_1 - c_1, R_2 - c_2 - Dv(1)\}\\
&\quad - \mu_1(0)\max\{R_0 - pDv(0), R_1 - c_1, R_2 - c_2 - Dv(0)\} + \mu_1(0)[Dv(1) - Dv(0)]\\
&\quad - \mu_1(2)\max\{R_0 - pDv(2), R_1 - c_1, R_2 - c_2 - Dv(2)\} + \mu_2 Dv(0).
\end{aligned} \tag{9}$$

Hence, combining Equations (8) and (9) and for any $0 \le x \le B-2$, we have

$$\begin{aligned}
&DTv(x+1) - DTv(x)\\
&\ge 2\mu_1(x+1)\max\{R_0 - pDv(x+1), R_1 - c_1, R_2 - c_2 - Dv(x+1)\} - \mu_1(x)\max\{R_0 - pDv(x), R_1 - c_1, R_2 - c_2\\
&\quad - Dv(x)\} - \mu_1(x+2)\max\{R_0 - pDv(x+2), R_1 - c_1, R_2 - c_2 - Dv(x+2)\} + \mu_1(0)[Dv(x+1) - Dv(x)]\\
&\ge [2\mu_1(x+1) - \mu_1(x+2)]\max\{R_0 - pDv(x+1), R_1 - c_1, R_2 - c_2 - Dv(x+1)\}\\
&\quad - \mu_1(x)\max\{R_0 - pDv(x), R_1 - c_1, R_2 - c_2 - Dv(x)\} + \mu_1(0)[Dv(x+1) - Dv(x)] \equiv G(x),
\end{aligned}$$

where the first inequality holds since $Dv(0) \ge 0$ and $Dv(x) - Dv(x-1) \ge 0$ for $0 < x \le B-2$, and the second inequality holds since $Dv(x+2) \ge Dv(x+1)$. We next show $G(x) \ge 0$ for any $0 \le x \le B-2$ by considering six separate cases.

*Case 1.* When $Dv(x) \le Dv(x+1) \le R_2 - R_1 - (1-p)^{-1}c_2$, we have

$$\begin{aligned}
G(x) &= [2\mu_1(x+1) - \mu_1(x+2)][R_2 - c_2 - Dv(x+1)] - \mu_1(x)[R_2 - c_2 - Dv(x)] + \mu_1(0)[Dv(x+1) - Dv(x)]\\
&= [2\mu_1(x+1) - \mu_1(x+2) - \mu_1(x)][R_2 - c_2 - Dv(x+1)] + [\mu_1(0) - \mu_1(x)][Dv(x+1) - Dv(x)] \ge 0,
\end{aligned}$$

where the inequality follows from that $\mu_1(x)$ is concave and non-increasing, $Dv(x) \le Dv(x+1)$, and $Dv(x+1) \le R_2 - R_1 - (1-p)^{-1}c_2 < R_2 - c_2$.

*Case 2.* When $Dv(x) \le R_2 - R_1 - (1-p)^{-1}c_2 \le Dv(x+1) \le R_2 - R_1 + p^{-1}c_1$, we have

$$G(x) = [2\mu_1(x+1) - \mu_1(x+2)][R_0 - pDv(x+1)] - \mu_1(x)[R_2 - c_2 - Dv(x)] + \mu_1(0)[Dv(x+1) - Dv(x)]$$

$$\geq [2\mu_1(x+1)-\mu_1(x+2)]R_0-\mu_1(x)(R_2-c_2)+[\mu_1(0)-2p\mu_1(x+1)+p\mu_1(x+2)]Dv(x+1)$$
$$+[\mu_1(x)-\mu_1(0)][R_2-R_1-(1-p)^{-1}c_2],$$

where the inequality follows since $\mu_1(x)\leq\mu_1(0)$ and $Dv(x)\leq R_2-R_1-(1-p)^{-1}c_2$. If $\mu_1(0)-2p\mu_1(x+1)+p\mu_1(x+2)\geq 0$, we have

$$G(x)\geq [2\mu_1(x+1)-\mu_1(x+2)]R_0-\mu_1(x)(R_2-c_2)+[\mu_1(0)-2p\mu_1(x+1)+p\mu_1(x+2)][R_2-R_1-(1-p)^{-1}c_2]$$
$$+[\mu_1(x)-\mu_1(0)][R_2-R_1-(1-p)^{-1}c_2]=[D\mu_1(x+1)-D\mu_1(x)][R_1+p(1-p)^{-1}c_2]\geq 0,$$

where the last inequality follows from that $\mu_1(x)$ is concave and using $Dv(x+1)\geq R_2-R_1-(1-p)^{-1}c_2$. If $\mu_1(0)-2p\mu_1(x+1)+p\mu_1(x+2)\leq 0$, using $Dv(x+1)\leq R_2-R_1+p^{-1}c_1$, we have

$$G(x)\geq [D\mu_1(x+1)-D\mu_1(x)]R_1+[\mu_1(0)-p\mu_1(x)](1-p)^{-1}c_2+[\mu_1(0)-2p\mu_1(x+1)+p\mu_1(x+2)]p^{-1}c_1$$
$$=[D\mu_1(x+1)-D\mu_1(x)](R_1-c_1)+[\mu_1(0)-p\mu_1(x)][p^{-1}c_1+(1-p)^{-1}c_2]\geq 0,$$

where the last inequality follows from the concavity of $\mu_1(x)$ and $R_1\geq c_1$.

*Case 3.* When $Dv(x)\leq R_2-R_1-(1-p)^{-1}c_2\leq R_2-R_1+p^{-1}c_1\leq Dv(x+1)$, we have

$$G(x)=[2\mu_1(x+1)-\mu_1(x+2)](R_1-c_1)-\mu_1(x)[R_2-c_2-Dv(x)]+\mu_1(0)[Dv(x+1)-Dv(x)]$$
$$\geq [2\mu_1(x+1)-\mu_1(x+2)](R_1-c_1)-\mu_1(x)(R_2-c_2)+\mu_1(0)R_2-R_1+p^{-1}c_1$$
$$+[\mu_1(x)-\mu_1(0)]R_2-R_1-(1-p)^{-1}c_2$$
$$=[D\mu_1(x+1)-D\mu_1(x)](R_1-c_1)+[p^{-1}\mu_1(0)-\mu_1(x)](1-p)^{-1}c_1\geq 0,$$

where the last inequality follows from the non-increasing and concave property of $\mu_1(x)$ and $R_1\geq c_1$.

*Case 4.* When $R_2-R_1-(1-p)^{-1}c_2\leq Dv(x)\leq Dv(x+1)\leq R_2-R_1+p^{-1}c_1$, we have

$$G(x)=[2\mu_1(x+1)-\mu_1(x+2)][R_0-pDv(x+1)]-\mu_1(x)[R_0-pDv(x)]+\mu_1(0)[Dv(x+1)-Dv(x)]$$
$$=[D\mu_1(x+1)-D\mu_1(x)][R_0-pDv(x+1)]+[\mu_1(0)-p\mu_1(x)][Dv(x+1)-Dv(x)]$$
$$\geq [D\mu_1(x+1)-D\mu_1(x)][R_0-p(R_2-R_1+p^{-1}c_1)]+[\mu_1(0)-p\mu_1(x)][Dv(x+1)-Dv(x)]\geq 0,$$

where the last inequality follows from the non-increasing and concave property of $\mu_1(x)$, $Dv(x+1)\geq Dv(x)$ and $R_1\geq c_1$.

*Case 5.* When $R_2-R_1-(1-p)^{-1}c_2\leq Dv(x)\leq R_2-R_1+p^{-1}c_1\leq Dv(x+1)$, we have

$$G(x)=[2\mu_1(x+1)-\mu_1(x+2)](R_1-c_1)-\mu_1(x)[R_0-pDv(x)]+\mu_1(0)[Dv(x+1)-Dv(x)]$$
$$\geq [2\mu_1(x+1)-\mu_1(x+2)](R_1-c_1)-\mu_1(x)R_0+[p\mu_1(x)-\mu_1(0)]Dv(x+1)+\mu_1(0)Dv(x+1)$$

$$\geq [2\mu_1(x+1) - \mu_1(x+2)](R_1 - c_1) - \mu_1(x)R_0 + p\mu_1(x)(R_2 - R_1 + p^{-1}c_1)$$
$$= [D\mu_1(x+1) - D\mu_1(x)](R_1 - c_1) \geq 0,$$

where the first inequality follows from that $p\mu_1(x) \leq \mu_1(0)$ and $Dv(x+1) \geq Dv(x)$; the second inequality follows from $Dv(x+1) \geq R_2 - R_1 + p^{-1}c_1$; and the last inequality follows from the concavity of $\mu_1(x)$ and $R_1 \geq c_1$.

*Case 6.* When $R_2 - R_1 + p^{-1}c_1 \leq Dv(x) \leq Dv(x+1)$, we have

$$G(x) = [D\mu_1(x+1) - D\mu_1(x)](R_1 - c_1) + \mu_1(0)[Dv(x+1) - Dv(x)] \geq 0.$$

where the inequality follows from the concavity of $\mu_1(x)$, $Dv(x+1) \geq Dv(x)$, and $R_1 \geq c_1$. Combining the results of the six cases, we conclude that $DTv(x+1) \geq DTv(x)$ for all $0 \leq x \leq B-2$. □

**Proof of Proposition 2:** We prove our results by verifying the conditions in Theorem 6.11.3 of Puterman (2005). It is obvious that the state space $\mathcal{S}$ is countable and Assumptions 6.10.1 and 6.10.2 of Puterman (2005) hold. Hence, we only need to show that conditions (a), (b), and (c) in Theorem 6.11.3 of Puterman (2005) hold. Condition (a) holds by Lemma 1. Next, consider a stationary policy $\pi$ that if decision makers choose an admit patient at $1 \leq x \leq B$, then they will choose admit patients at $x'$ where $x' \leq x$; if decision makers choose a discharge patient at $0 \leq x \leq B-1$, then they will choose discharge patients at $x'$ where $x' \geq x$; otherwise, choose the first patient in line. After that, $\pi$ follows the optimal policy. From the optimality equations, we find that $v \in \mathcal{F}$ implies that policy $\pi$ is an optimal policy. Hence, condition (b) holds. Finally, condition (c) holds, i.e, $\mathcal{F}$ is closed, because the limit of any convergent sequence of functions that satisfy the two conditions of $\mathcal{F}$ will satisfy them as well. This concludes the proof that there exists an optimal stationary policy whose value function belongs to $\mathcal{F}$ for the discount net social benefits version of the problem, and the same structural results can be extended to the average net social benefit version of the MDP by verifying three SEN conditions given in Section 7.2 of Sennott (1999) in a fairly straightforward manner.