Who to See Next? Guiding Physicians with Time-dependent Patient Selection Policies Under Shift Work

Mahdi Shakeri
Haskayne School of Business, University of Calgary, Calgary, AB T2N 1N4, Canada, mahdi.shakeri@ucalgary.ca
Marco Bijvank
Haskayne School of Business, University of Calgary, Calgary, AB T2N 1N4, Canada, marco.bijvank@haskayne.ucalgary.ca
Soroush Saghafian
Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA, soroush_saghafian@hks.harvard.edu

Patient selection in healthcare settings with walk-in patients is often a decision at a physician's discretion. In some overcrowded environments such as Emergency Departments (EDs), how patients are selected and prioritized can significantly impact operational performance as well as clinical outcomes. We propose a time-dependent policy to guide ED physicians in deciding who to serve next considering that they often have new patients waiting for an initial assessment, returning patients waiting for a reassessment, and might have to hand off their patients to a different physician at the end of their shift. To gain insights, we formulate a transient optimal control problem and demonstrate that a time-threshold-type policy is optimal. For special cases, we derive prioritization rules between patient classes that can be seen as extensions of the $c\mu$-rule when there are returning patients as well as a penalty cost for patient hand-offs. For more general settings, insights are obtained into when to prioritize certain patient classes as well as how this prioritization can switch during the physician's shift. Simulation analyses calibrated with detailed hospital data under realistic ED features such as multiple physicians and time-varying arrival rates indicate that our proposed policy can improve ED operational performance measures. In particular, our results show that, if the insights from our analysis are implemented, ED throughput can significantly increase, and the wait time, length of stay, number of patient hand-offs, and number of patients who leave without being seen by a physician can decrease compared to both the current practice and various other potential selection policies.

Key words: Patient selection; Priority queues; Time-dependent decision-making; Optimal control

1. Introduction

Emergency Department (ED) crowding has been a worldwide problem for more than two decades. This common phenomenon originates from demand for emergency services that is exceeding the ability of an ED to provide adequate care within an appropriate time frame. It results in prolonged ED wait times and lengths of stay (LOS) for patients, which contributes to adverse treatment times and patient outcomes (Sun et al. 2013, Carter et al. 2014, Rasouli et al. 2019, Abir et al. 2019, Jones et al. 2021) as well as more patients who leave the ED without being seen by a physician

(Janke et al. 2022). Even though increased ED wait times have been the subject of considerable examination in the literature (Morley et al. 2018, Pearce et al. 2023), there is no clear strategy to address ED crowding. Consequently, EDs must continue to provide care during periods with large patient volumes, and health care professionals (e.g., physicians and nurses) often face a heavy workload. ED Physicians have some discretion to make decisions based on their individual expertise, beliefs, and preferences. This might lead to processes and practices that are less than optimal from a resource utilization perspective and it can directly impact operational measures such as ED wait times and patient LOS. In this paper, we propose new guidelines for physicians on which patients to see next in a crowded ED to improve patient flow and ED performance.

The sequence in which a physician selects the next patient to be examined or treated for either an initial assessment or reassessment depends on the health care provider's own understanding of the status of the ED, the patients, and themselves. Beyond EDs, addressing the question "who should see the patient?" and whether physicians' discretions in choosing their next patient should be fully eliminated in hospitals so as to avoid deviations from preferred patient-provider assignments has been shown to be important in improving various performance metrics, including operational efficiency and quality of care (Atkinson and Saghafian 2023). In EDs, a common practice is to first assign patients an acuity score during triage—a process in which a nurse collects medical information, vital signs and the chief complaint of a patient. In the United States, the Emergency Severity Index (ESI) is widely used to assess a patient's acuity during triage (Wolf et al. 2023), while in Canada this assessment is based on the Canadian Triage and Acuity Scale (CTAS) (Bullard et al. 2017). Both ESI and CTAS are 5-level triage scores that can help identify which patients are more critical to receive treatment by classifying them into priority groups from most severe (level 1) to least severe (level 5). While EDs try to prioritize patients using triage levels, it is shown that ED physicians do not always stay with these triage levels as strict prioritization (Ding et al. 2019). Moreover, even among patients with the same triage level, various factors influence the physician's decision on which patient to see next, including the patient's waiting time, whether the patient requires an initial assessment or a reassessment (e.g., after obtaining test results) as well as how much time a physician has left in their shift. With regard to the last aspect, Batt et al. (2019) and Zaerpour et al. (2022) have shown with patient-level data that the likelihood for an ED physician to pick up a new patient drastically decreases near the end of a physician's shift. One of the main reasons for this is that the care for patients needs to be transferred to a different physician at the end of a shift, which can affect the continuity of care. This practice of passing work from one physician to another is called *hand-offs*. It is known that patients who are handed off are more likely to revisit an ED, which is an indication that they have received lower clinical care quality (Batt et al. 2019).

To systematically remove hand-offs, create a balanced workload among physicians, and also eliminate the negative consequences of physicians' cherry-picking behavior in selecting their patients, some EDs have made use of rotational-based patient-provider assignments, in which arriving patients are randomly assigned to each physician unless the physician is close to the end of their shift (Traub et al. 2016a,b). While showing some promising impacts, these approaches do not consider patients waiting times or other system characteristics, and simply pursue randomizing the assignment of an arriving patient to physicians. Although randomized rotational-based patient-provider assignments can improve performance, they are suboptimal. Moreover, they do not guide physicians in how they should prioritize their tasks in handling new patients versus those who are already in-process but need attention (e.g., for a reassessment). Our goal in this paper is to address these gaps. Specifically, we aim to address the following research questions: (i) When should a physician prioritize patients with a higher urgency over patients with a lower urgency? (ii) When should a physician prioritize new patients over patients who are already in-progress but need attention (e.g., for a reassessment or follow-up interaction after their test result becomes available)? and (iii) When should a physician stop selecting a new patient and only focus on in-progress patients to prevent hand-offs? In order to answer these questions, we will introduce a general queueing system that we analyze with fluid approximations and subsequently propose an optimal control problem.

Studying prioritization in queueing systems with multiple customer classes is not new. When the waiting time of customers is penalized linearly with time, Cox and Smith (1961) show that the well-known $c\mu$-rule is optimal for an $M/G/1$ queue with multiple priority classes. According to the $c\mu$-rule, customers with a larger $c_i\mu_i$ index are assigned a higher priority, where c_i is the waiting cost per unit time and μ_i is the service rate of class i customers. Van Mieghem (1995) proves that this rule is asymptotically optimal under heavy traffic when the waiting cost is non-decreasing convex in time. This result is extended to more general settings by Mandelbaum and Stolyar (2004). In our work, we extend the $c\mu$-rule in two directions: (1) we differentiate between customers (patients in our setting) who require service for the first time and those who return to the same server after some random delay, and (2) we assume servers (physicians in our setting) operate in shifts where customers still under service of the server at the end of the shift get penalized. Since we consider an ED setting, we refer to the customers who wait for their initial service as *new patients* and those who wait for a follow-up service as *in-progress patients*. This terminology is similar to prior work, including Saghafian et al. (2012) and Huang et al. (2015), who also focus on prioritization rules in an ED setting (see Section 2 for more details).

Our main theoretical contribution in this paper is that we do not propose a static policy, and instead, we derive a time-dependent policy in which the priority of a certain patient type (new or in-progress) or class (urgent or non-urgent) depends on the timing of a physician's shift. Furthermore,

the actual time thresholds on when to switch priorities in our setting depend on the status of the ED at the time when the physician starts their shift. We do this by introducing a queueing model with a waiting cost for new patients, a holding cost for in-progress patients, and a penalty cost for in-progress patients who are handed off at the end of the physician's shift. A similar approach is proposed by Ouyang et al. (2021), but they only determine the timing of when to stop accepting new patients where in-progress patients are always prioritized over new patients and no different patient classes are considered. Batt et al. (2019) study the impact of such a time threshold (or "cutoff policy" as they call it) on hand-offs and productivity via a simulation study. No previous work includes multiple patient classes for new and in-progress patients where physicians experience hand-offs at the end of their shift.

The remainder of this paper is organized as follows. In Section 2 ,we provide an overview of the literature that is related to our work. In Section 3, we introduce a general queueing model of the ED setting under study. The optimal control problem is formulated in Section 4, whereas our main theorems that derive an optimal policy are presented in Section 5. Our numerical experiments are presented in Section 6, and concluding remarks are provided in Section 7. Additional technical details, including all proofs, are presented in the appendix.

2. Related Literature

The problem considered in this paper is related to two streams of research in the literature. First, it is closely related to works that apply queueing modeling to study service systems where customers may re-enter the queue. Second, it is related to the extensive body of work on scheduling queues with multiple classes of customers. We will provide a brief overview of these two related literature streams.

2.1. Queueing Systems with Returning Customers

In this section, we only discuss studies of queueing systems where some customers return for another service after a positive delay time.[1] Most papers that study a multi-server queueing system with returning customers make the assumption that the set of servers is pooled when a customer returns. Such a fully pooled system implies that a customer is likely to be served by different servers. de Véricourt and Jennings (2011) study such a system for a nursing home where patients alternate between needing assistance and being content. The efficiency of nurse-to-patient ratio policies is examined using a multi-server $M/M/s/\cdot/n$ queueing system with a finite demand source. The closed queueing model aims at determining minimum staffing levels that are required to satisfy a given constraint on the probability that patients are delayed beyond a desired threshold. A similar nurse staffing problem is studied by Yankovic and Green (2011), who model demand to come from a homogeneous Poisson process as well as from patients who request care during their stay.

When the arrival process of new customers is a non-homogeneous Poisson process with a time-dependent arrival rate, Yom-Tov and Mandelbaum (2014) are among the first to study a time-varying $M/M/s$ queueing system where customers may return to the system. They use time-varying fluid and diffusion approximations to analyze the system, and propose a physician staffing algorithm based on a time-varying square-root staffing policy with the offered load as inputs. The approach is extended by Liu and Whitt (2017) to a setting where the return probability, service-time distribution and subsequent delay distribution (before returning for a new service) can vary between return visits, and these distributions are allowed to be non-exponential. More recently, Furman et al. (2021) evaluate the performance of a non-stationary multi-server service system where customers can decide to join a customer base after their initial service and these customers can seek service again at a later point in time. Their model assumes there are dedicated servers for new customers and returning customers. The authors derive a fluid approximation of the queueing system, which results in a system of ordinary linear differential equations (ODEs) that are used to assign a number of servers to the two customer classes. Chan et al. (2024) study a control problem within such a queueing model, where a decision-maker can influence the probability of a customer's return, with lower return probabilities incurring higher costs.

These models all assume that the servers for returning customers are pooled, and hence, a returning customer can be served by a variety of servers. A different approach is taken by Campello et al. (2017), where a customer is assigned to a specific server and the server repeatedly interacts with these customers. This better matches the reality of EDs, where each physician is responsible for their own set of patients (until the end of their shift). Furthermore, an upper limit is imposed on the number of customers simultaneously assigned to a single server. This results in two separate queues; customers awaiting assignment to a server in a pre-assignment queue (when all servers have reached their limit) and customers awaiting service in an internal queue once assigned to a specific server. Customers are assigned to servers based on a join-shortest-queue (JSQ) discipline, whereas customers in the internal queue are processed on a first-come-first-served (FCFS) policy. To analyze the system performance, the authors formulate four models to approximate the baseline model. The use of a case manager for different customer classes is studied by Kamalahmadi et al. (2023) in a healthcare setting, where the caseload and case-mix (i.e., number of patients per class) are decision variables that are optimized while minimizing the average length of stay for patients. In their work, servers are hospitalists who alternate between rounding and responding service modes (i.e., scheduled visits to each patient and requested services as need arises, respectively). During rounds, hospitalists make discharge decisions. Consequently, the optimal case-mix assigned to a physician should depend on the frequency by which patients need interventions. Therefore, in

addition to complex patients, a physician also requires some simple patients to reduce discharge delay.

Another stream of research investigates queueing systems where the service rate and the probability that a customer returns depend on the congestion level in the system. Chan et al. (2014) study such a system, where the service rate and return probability increase when the number of busy servers working on new customers is above a certain threshold. This happens when the server speeds up as the system becomes congested. The authors analyze their stochastic model via a deterministic fluid approximation that consists of a system of two ODE's. Ingolfsson et al. (2020) analyze a similar queueing system where the difference is the timing when the delay for the return of a customer happens. In all above-mentioned papers, the delay only happens for returning customers, with the exception of the model in Ingolfsson et al. (2020) where a delay occurs before it is decided whether the customer will return. Barjesteh and Abouee-Mehrizi (2021) extend the queueing system of Chan et al. (2014) with multiple customer classes where the service rate and return probability depend on the number of customers in the system in a more general fashion. The authors derive a fluid limit to analyze the dynamics and characterize the equilibria of the system.

2.2. Controlling Multiple Customer Classes and Types

For queueing systems with heterogeneous customer classes but no returning customers, the well-known $c\mu$-rule (Cox and Smith 1961) and $Gc\mu$-rule (Van Mieghem 1995, Mandelbaum and Stolyar 2004) have been proved to be optimal in various configurations (see the discussion in Section 1). Atar et al. (2010) show that for systems with customer abandonment, the $c\mu/\theta$-rule is optimal, where θ denotes the abandonment rate. In the context of EDs, more complex features need to be considered. For example, assigning patients from the ED to inpatient hospital wards can be modeled as a multi-class queueing system with heterogeneous and flexible servers. In this setting, policies proposed by Saghafian et al. (2024) and Shakeri et al. (2023) address these complexities and have been shown to perform well in improving ED performance. Nevertheless, as discussed in Section 2.1, almost all queueing studies that include returning customers focus on determining the required number of service providers (i.e., optimizing a staffing problem) under the assumption that new and returning customers are served on a FCFS basis or that the returning customers are always prioritized over new customers. A limited number of papers study the dynamic scheduling of multiple customer classes through a queueing system with returning customers. The patient flow in EDs and many other service systems that require knowledge-based decisions involve customers who return for another evaluation by the same or a different server (e.g., after being sent for a medical test). Various models are proposed in the literature for optimizing the ED patient flow and such knowledge-based service systems; see Saghafian et al. (2015) and Saghafian et al. (2018) for more detailed discussions.

Some studies such as Dobson et al. (2013) investigate how to prioritize new versus existing patients in the presence of interruptions. They consider settings in which patients who are in the system after the initial service generate additional work for the physician in addition to the return visit, such as filing paperwork or answering nurse questions. These interruptions are given priority by the physician, and a maximum number of in-progress patients can be in the system simultaneously. In the queueing system studied by Dobson et al. (2013), the delay for in-progress patients comes from a single server queue whereas many other studies in the literature make use of models in which the delay is due to an infinite server queue. When in-progress patients do not generate additional interruptions as they wait, using a Markov Decision Process (MDP), Dobson et al. (2013) show that it is throughput maximizing to prioritize new patients if the number of patients in the system is below a certain threshold value; otherwise, prioritizing in-progress patients becomes optimal. The exact optimal policy is complicated when interruptions occur. Using an asymptotic analysis, the authors show that the server should give priority to returning customers when maximizing the throughput of the system.

The optimal control of a single-server queueing system with multiple patient classes is studied by Saghafian et al. (2014). They extend the $c\mu$-rule to the following: a physician should prioritize patients in decreasing order of $p_i\mu_i$, where p_i is the probability that a patient of class i departs the ED system after service. The authors also include the probability of adverse events and misclassification for patients as well as patient medical complexity (besides urgency). To the best of our knowledge, prioritization rules between new and in-progress ED patients were first studied by Saghafian et al. (2012) and Huang et al. (2015). Saghafian et al. (2012) consider the fact that the ED service process is multi-stage with patients requiring reassessments from their physician. They study the problem of "who to see next" and derive prioritization rules on whether and when physicians should prioritize in-progress patients (a policy termed "Prioritize Old") or new patients (a policy termed "Prioritize New"). However, they do not consider end-of-shift concerns in their analysis, and instead focus on policies that can perform well in EDs when considering a combination of metrics such as mean time-to-first treatment and length of stay. Huang et al. (2015) minimize the holding cost of in-progress patients subject to a maximum wait time for new patients. The proposed threshold policy prioritizes in-progress patients unless the wait time of a new patient is within ϵ time units of the maximum wait time. If new patients are prioritized, the patient with the shortest time remaining until reaching the maximum wait time should be selected. And if in-progress patients are prioritized, the patient class with the largest value of index $C_i'(Q_i(t))\mu_i^e$ should be selected, where $C_i'(Q_i(t))$ is the derivative of a convex cost function of the number of in-progress patients in class i at time t, and μ_i^e is the effective service rate of in-progress patients in class i. Finally, when there are multiple physicians as well as delay targets for new and in-progress

patients (w.r.t. wait time and length of stay, respectively), He et al. (2019) propose a dynamic scheduling framework where a hybrid optimization problem is solved sequentially.

Other than these analytical studies that gain insights from stylized queueing models, there are also a number of empirical or simulation-based studies that focus on the order in which ED physicians select patients to see next among their available patients. Ferrand et al. (2018) compare the performance of different priority policies with simulation. One of their observations is that the average LOS decreases significantly for all patients when physicians first prioritize returning patients over new patients, and then use the triage level for prioritization (in comparison to strictly using triage levels for both new and returning patients). They also consider a system where patients accumulate waiting cost, where the cost c_i per time unit is higher for more urgent patients. Consequently, a patient with a lower acuity who has waited longer can be prioritized over a higher acuity patient who has waited less time. Similar cost functions are empirically studied with patient-level ED visit data by Ding et al. (2019), who conclude that the marginal waiting cost increases during the initial wait of a customer, but becomes constant if the customer has waited beyond a threshold value.

3. Modeling the Patient Flow During a Shift as a Queueing System

To address our research questions, we formulate the control problem in terms of patients who need service in an ED.[2] The arrival of patients to the ED occurs in a stochastic manner. After registration, they will be triaged by a nurse and assigned a triage level based on their acuity, vital signs, and some other characteristics. For ease of understanding, we start our analyses by analyzing a simplified model of the ED flow where we classify patients into two classes: class 1 and class 2, which can be viewed, for example, as urgent and non-urgent patients. However, in developing our analytical results, we do not impose this view and assume they are two generic classes of patients. In this model, class $i \in \{1, 2\}$ patients arrive according to a Poisson process with rate λ_i.[3] A physician can choose a triaged patient from the waiting area for an *initial assessment*. We assume that the duration of the initial assessment for class i patients is exponentially distributed with rate μ_i. After the initial assessment of a patient from priority class i, the physician requests medical examination or testing (e.g., laboratory tests and/or diagnostic imaging) with probability θ_i, or dispositions the patient with probability $1 - \theta_i$ during which the patient will be either discharged home or admitted to the hospital. For tractability, we model the delay due to such tests or procedures as i.i.d. exponential random variables. This is reasonable as test facilities are shared with other units outside the ED. Consequently, the wait times for these facilities can be assumed to be exponentially distributed (or nearly so) and that the delay can be approximated by $\cdot/M/\infty$ models from the perspective of the ED physician. Furthermore, we assume that the delay comes from generic testing with a service rate δ that is the same across different patient classes.

After the test results are obtained, the patient returns to the physician for a *reassessment* (or *follow-up interaction*). We assume that the duration of the reassessment for class i patients is exponentially distributed with rate μ_i'. Once the reassessment is completed, the physician sends the class i patient for further medical tests with probability θ_i' or dispositions the patient with probability $1-\theta_i'$. As such, a patient goes through the initial assessment only once but may receive multiple follow-up assessments. We refer to patients waiting for an initial assessment as *new patients*, and patients who are undergoing further medical tests or procedures or who are waiting to receive a follow-up assessment as *in-progress (IP) patients*. In Section 5.4, we extend our model to the case where new patients waiting for an initial assessment may leave the ED without being seen by a physician. Furthermore, we perform various robustness checks by relaxing some of our modeling assumptions. For example, we consider scenarios in which the probability for an additional assessment decreases in the number of previous assessments and scenarios in which the delay for a reassessment depends on the patient class (see Appendix EC.5.3).

As is the case in most EDs, we assume that a new patient can be selected by any available physician, whereas a returning in-progress patient must be seen by the physician who performed the initial assessment. Consequently, an individual physician may have several patients under their care. It is up to the physician's discretion to decide which available patient to see next. If the physician gives priority to new patients to reduce their wait time, the in-progress patients have to spend more time in the ED. Especially patients who need multiple rounds of reassessments will experience an increased LOS. In contrast, giving priority to in-progress patients can effectively shorten their LOS at the expense of prolonged wait times for new patients. We study this trade-off and shed light on suitable decision-making rules that can be followed by physicians. In practice, physicians make these decisions by also taking into account how much time is left in their shift. For example, they might be reluctant to take on a new patient when they are close to the end of their shift. Hence, we analyze our setting illustrated in Figure 1 by considering time-dependent shift effects, where a physician works a shift with a duration of T time units. In this setting, whenever a shift is ending, the physician stops working and hands over their remaining IP patients to another physician who (usually) just started their shift.

To further incorporate the reality of the problem faced by ED physicians, we also assume that the prioritization of patients is non-preemptive. This means that the assessment of a patient will not be interrupted due to the arrival of a patient with a higher priority. Moreover, because each physician has their own set of IP patients, and that the EDs are often overcrowded with many patients waiting to be served in their waiting area, the decision-making of each physician is to a great extent independent of that of other physicians. Therefore, we consider a single-server model to gain clear insights into our research questions. However, in our simulation study (calibrated with

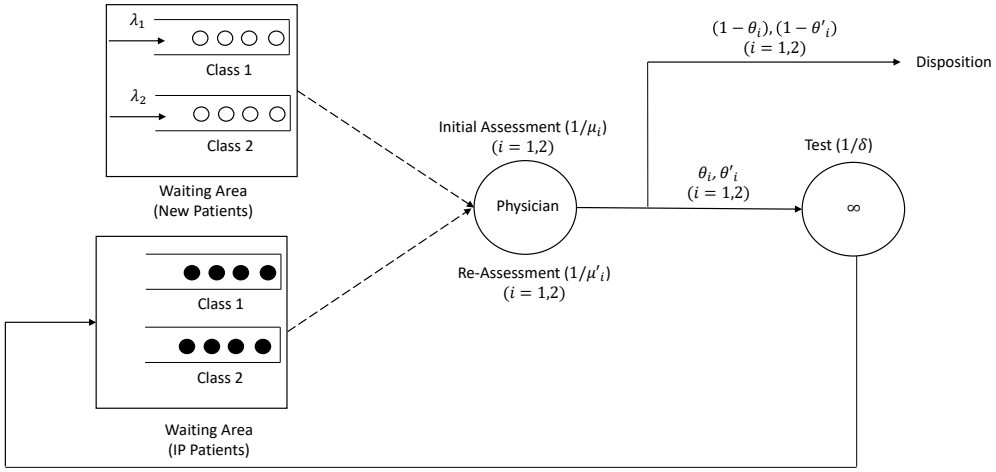


Figure 1 Patient flow in an ED: a multi-class queueing system with delayed returns

hospital data) in Section 6, we relax this assumption and test our proposed policy in a multi-server setting, where 16 physicians are scheduled to work over the course of one day.

To capture the state of the system depicted in Figure 1 at time t, we employ the stochastic process $\{Q(t), t \geq 0\}$, where $Q(t) = (Q_j(t), j \in \{1,2,3,4,5,6\})$. In this formulation, $Q_1(t)$ and $Q_2(t)$ represent the number of new patients in the waiting area at time t that are of class 1 and class 2, respectively. The states $Q_3(t)$ and $Q_4(t)$ denote the number of available IP patients at time t who are waiting for a reassessment by the physician and are of class 1 and class 2, respectively. Furthermore, $Q_5(t)$ and $Q_6(t)$ denote the number of IP patients at time t who experiencing delay (e.g., undergoing tests or waiting for test results) of class 1 and class 2, respectively. We use $U(t) = (U_j(t), j \in \{1,2,3,4\})$ to denote the control variables, where binary variable $U_j(t)$ represents whether or not at time t the physician is assessing a new patient of class $i = j$ for $j = 1,2$, or an IP patient of class $i = j-2$ for $j = 3,4$. These control variables need to be determined based on a patient selection policy Π from the set of admissible patient selection policies Ω that satisfies $\sum_j U_j(t) \leq 1$ at any time t. As the system state directly depends on the patient selection policy, we use $Q^{\Pi}(t)$ to highlight this dependency..

The system incurs two types of costs: an immediate cost and a terminal cost. The immediate cost reflects the waiting cost per unit of time associated with keeping different patient classes waiting, while the terminal cost captures the penalty cost associated with any patient hand-off at the end of the physician's shift. Specifically, we consider a holding cost h'_j, $j \in \{1,2\}$, for each unit of time that a new patient of class $i = j$ waits for an initial assessment, a holding cost h'_j, $j \in \{3,4\}$, for each unit of time that an available IP patient of class $i = j-2$ waits for a reassessment, and a holding cost h'_j, $j \in \{5,6\}$, for each unit of time that an IP patient of class $i = j-4$ remains in

a test facility. Moreover, there is a penalty cost c_i associated with each patient of class i that is handed off to another physician at the end of the shift ($i = 1, 2$). The objective is to find a patient selection policy Π^* that minimizes the expected total cost over the course of a physician's shift, namely,

$$\min_{\Pi \in \Omega} \quad \mathbb{E}\left[\int_0^T \sum_{j=1}^{6} h_j' Q_j^{\Pi}(t)dt \; + \; \sum_{i=1}^{2} c_i \left(Q_{i+2}^{\Pi}(T) + Q_{i+4}^{\Pi}(T)\right)\right] . \tag{1}$$

Finding the optimal patient selection policy can be formulated as an MDP. MDPs are widely used to improve various aspects of ED patient flow (see, e.g., Section 3.4 of Saghafian et al. (2015) for a review). However, in our setting, an exact analysis to characterize the structure of the optimal policy is difficult due to the curse of dimensionality and the fact that the policy needs to be derived with respect to the transient behavior of the system. Therefore, we use fluid approximations to analyze the associated stochastic process and find the optimal policy on the fluid limits of the system. Several papers in the literature use a similar approach (see, e.g., Zychlinski (2023) for a recent review). This reformulation is presented in the next section.

4. Fluid Model Approximation and Optimal Control

Since a solution to the problem formulation in Eq. (1) cannot be found, we introduce a fluid model to approximate it in Section 4.1. To find an optimal patient selection policy based on the fluid limits obtained in Section 4.1, we reformulate the problem as an optimal control problem in Section 4.2.

4.1. Fluid Model

To formulate a fluid model that serves as an approximation for our original problem, we replace the discrete stochastic process in Eq. (1) with continuous deterministic average rates. Under the Functional Strong Law of Large Numbers (FSLLN), we derive the fluid model by letting the arrival rates and service rates grow proportionally to infinity. To achieve this, we introduce a family of scaled-up stochastic processes with parameter η denoted by $Q^{\eta}(t) = (Q_j^{\eta}(t), j \in \{1, 2, 3, 4, 5, 6\})$, where $\lambda_i^{\eta}(\cdot) = \eta\lambda_i(\cdot)$, $\mu_i^{\eta}(\cdot) = \eta\mu_i(\cdot)$, and $\mu_i'^{\eta}(\cdot) = \eta\mu_i'(\cdot)$ for $i = 1, 2$. Furthermore, we let $q(t) = (q_j(t), j \in \{1, 2, 3, 4, 5, 6\})$ represent a vector of fluid limits, where $q_j(t)$ corresponds to the fluid limit of processes $Q_j(t)$ for $j = 1, 2, 3, 4, 5, 6$. Based on the FSLLN, $q(t) = \lim_{\eta \to \infty} \frac{Q^{\eta}(t)}{\eta}$. Accordingly, these fluid limits can be obtained by the following set of ordinary differential equations (ODEs):

$$\dot{q}_1(t) = \lambda_1 - \mu_1 u_1(t) \tag{2}$$

$$\dot{q}_2(t) = \lambda_2 - \mu_2 u_2(t) \tag{3}$$

$$\dot{q}_3(t) = \delta q_5(t) - \mu_1' u_3(t) \tag{4}$$

$$\dot{q}_4(t) = \delta q_6(t) - \mu_2' u_4(t) \tag{5}$$

$$\dot{q}_5(t) = \theta_1 \mu_1 u_1(t) + \theta_1' \mu_1' u_3(t) - \delta q_5(t) \tag{6}$$

$$\dot{q}_6(t) = \theta_2 \mu_2 u_2(t) + \theta_2' \mu_2' u_4(t) - \delta q_6(t) \; , \tag{7}$$

where $u_j(t)$ is the counterpart of control variable $U_j(t)$ in the fluid model. Note that in fluid models, the capacity of the server can be divided among different patients (Maglaras 2006). Therefore, $u_j(t)$ can take any real number between 0 and 1, with the sum of all $u_j(t)$'s being less than or equal to 1. Consequently, the cost minimization problem in Eq. (1) can be reformulated as

$$\min_{u} \quad \left[\int_0^T \sum_{j=1}^{6} h'_j q_j(t) dt + \sum_{j=1}^{2} c_j (q_{j+2}(T) + q_{j+4}(T)) \right] , \tag{8}$$

subject to Eqs. (2) to (7), $q_j(t) \geq 0$, $u_j(t) \geq 0$, and $\sum_{j=1}^{4} u_j(t) \leq 1$.

There are three sets of constraints in this minimization problem: (*i*) *dynamic constraints* ensure that the trajectory of optimal states follows the ODEs stated in Eqs. (2) to (7); (*ii*) *control-dependent constraints* state that each control variable $u_j(t)$ must be non-negative, while the sum of all control variables must be less than or equal to one; and (*iii*) *state-dependent constraints* require that each state variable $q_j(t)$ must be non-negative. To solve the corresponding transient optimal control problem, we use Pontryagin's Maximum Principles, which provide necessary optimality conditions (Sethi 2019).

4.2. Optimal Control Problem

We transform the minimization problem of Eq. (8) into a maximization problem, where the objective function is the negative cost. The transient optimal control problem is formulated as follows:

$$\max_{u} \quad \left[\int_0^T -F(q(t)) dt - \Psi(q(T)) \right] , \tag{9}$$

subject to

$$\dot{q}(t) = f(q(t), u(t)) \tag{10}$$

$$g(u(t)) \geq 0 \tag{11}$$

$$h(q(t)) \geq 0 , \tag{12}$$

where $F(q(t)) = \sum_{j=1}^{6} h'_j q_j(t)$ represents the immediate cost function, and $\Psi(q(T)) = \sum_{j=1}^{2} c_j (q_{j+2}(T) + q_{j+4}(T))$ represents the terminal cost function. The notations for the three types of constraints in Eqs. (10) to (12) are discussed in the remainder of this section.

The dynamic function in Eq. (10) is denoted by $f(q(t), u(t)) = (f_j(q(t), u(t)), j \in \{1, 2, 3, 4, 5, 6\})$, where

$$f_j(q(t), u(t)) = \lambda_j - \mu_j u_j(t) \qquad j = 1, 2 \tag{13}$$

$$f_j(q(t), u(t)) = \delta q_{j+2}(t) - \mu'_{j-2} u_j(t) \qquad j = 3, 4 \tag{14}$$

$$f_j(q(t), u(t)) = \theta_{j-4} \mu_{j-4} u_{j-4}(t) + \theta'_{j-4} \mu'_{j-4} u_{j-2}(t) - \delta q_j(t) \qquad j = 5, 6 . \tag{15}$$

The control-dependent function in Eq. (11) is denoted by $g(u(t)) = (g_j(u(t)), j \in \{1,2,3,4,5\})$, where

$$g_j(u(t)) = u_j(t) \qquad j = 1,2,3,4 \tag{16}$$

$$g_5(u(t)) = 1 - \sum_{j=1}^{4} u_j(t) \ . \tag{17}$$

The state-dependent function in Eq. (12) is denoted by $h(q(t)) = (h_j(q(t)), j \in \{1,2,3,4,5,6\})$, where $h_j(q(t)) = q_j(t)$ for $j = 1,2,3,4,5,6$.

To analyze the optimal policy in the fluid approximation under heavy traffic conditions, we introduce the following assumption to the problem. Such an assumption is common in a fluid analysis of a queueing system (see, e.g., Zychlinski (2023)).

Assumption 1. *There are always new patients to be seen in the waiting area (i.e., $q_1(t) > 0$ and $q_2(t) > 0$ for $t \in [0,T]$).*

Based on Assumption 1, the state-dependent constraints related to $q_1(t)$ and $q_2(t)$ are always satisfied. As a result, we can omit them from the optimal control problem. Furthermore, we present the following lemma which states the trajectory of $q_5(t)$ and $q_6(t)$ over the entire planning horizon. All proofs are provided in Appendix EC.2.

Lemma 1. *The quantities $q_5(t)$ and $q_6(t)$ are non-negative (i.e., $q_j(t) \geq 0, j = 5,6$) for all t.*

Based on Lemma 1, the state-dependent constraints for $q_5(t)$ and $q_6(t)$ are satisfied and can be removed from the optimal control formulation. As a result, the only state-dependent constraints that should be explicitly considered in the optimal control formulation are the ones related to $q_3(t)$ and $q_4(t)$ (i.e., $q_j(t) \geq 0, j = 3,4$). These state-dependent constraints add significant complexity to the problem, and the difficulty arises from the fact that these constraints lack any explicit control variable (Hu et al. 2022). To overcome this challenge, we take an indirect approach (Sethi 2019, Chapter 4), according to which, we select control variables such that $\dot{q}_j(t) \geq 0$ when $q_j(t) = 0$ for $j = 3,4$. The optimality conditions to solve our optimal control problem with this indirect approach are presented in Appendix EC.1.

5. Optimal Policy

In this section, we characterize the optimal policy for the optimal control problem as formulated in Section 4.2. We first present a proposition that specifies that the optimal policy is a "bang-bang" policy. Next, we specify that the structure of the optimal policy follows a time-threshold policy.

Proposition 1. *The optimal control policy to the problem in Eqs. (9) to (12) is a bang-bang policy (i.e., $u_j(t)$ is either 0 or 1 for all j) in an interior interval (i.e., in an interval (τ_1, τ_2) where $q_j(t) > 0$ ($j = 3,4$) for all $t \in (\tau_1, \tau_2)$).*

Based on Proposition 1, it is optimal to allocate all the capacity of a physician to only one patient type, if all patient types are present. However, which patient type to use for this allocation can vary over time and is not fixed during the shift of a physician. To gain insights into which patient type should be selected by a physician, we introduce a family of patient selection policies, termed *time-threshold-type policies*. Such policies specify the patient type that should be selected next at any point during a physician's shift.

Definition 1 (Time-Threshold-Type Policy)**.** *A time-threshold-type policy, denoted as* $\pi^t(\bar{t}_1, \bar{t}_2, \tilde{t}, \hat{t})$*, operates as follows:*

- *For* $t \leq \bar{t}_1$*, new patients get priority over IP patients. For* $t > \bar{t}_1$*, the priority switches to IP patients over new patients.*
- *For* $t \leq \tilde{t}$*, new patients of a class* ν_1 *get priority over new patients of another class* $\overline{\nu}_1 = 3 - \nu_1$*. Conversely, for* $t > \tilde{t}$*, the priority shifts to new patients of class* ν_2 *over new patients of class* $\overline{\nu}_2 = 3 - \nu_2$*.*
- *For* $t \leq \hat{t}$*, IP patients of a class* κ_1 *get priority over IP patients of another class* $\overline{\kappa}_1 = 3 - \kappa_1$*. For* $t > \hat{t}$*, the priority shifts to IP patients of class* κ_2 *over IP patients of class* $\overline{\kappa}_2 = 3 - \kappa_2$*.*
- *No new patients are selected for their initial treatment if* $t > \bar{t}_2$*.*

Note that new and IP patients within their patient class are selected based on a FCFS regime. Figure 2 presents an illustration of a time-threshold-type policy $\pi^t(\bar{t}_1, \bar{t}_2, \tilde{t}, \hat{t})$ where $\nu_1 = \kappa_1 = 1$ and $\nu_2 = \kappa_2 = 2$. According to this policy, the physician prioritizes new patients over IP patients until time $\bar{t}_1$, after which IP patients are prioritized. Among new patients, class 1 is prioritized over class 2 until time $\tilde{t}$, and class 2 is prioritized over class 1 thereafter. After time $\bar{t}_2$, no new patients are selected. Similarly, among IP patients, class 1 is prioritized over class 2 until time $\hat{t}$, and class 2 is prioritized over class 1 after that. Note that the relative order of the time thresholds $\bar{t}_1$, $\bar{t}_2$, $\tilde{t}$, and $\hat{t}$ may differ from what is illustrated in Figure 2 depending on the system parameters.

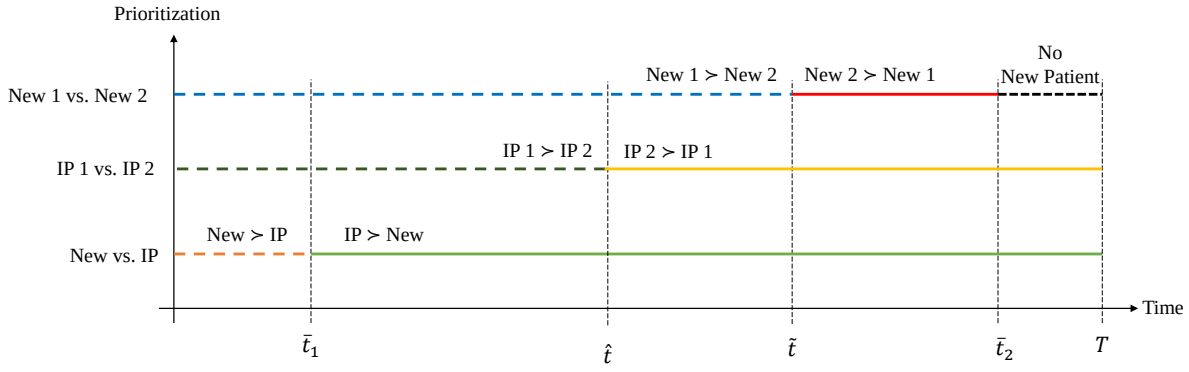


Figure 2 **Visualization of a time-threshold-type policy** $\pi^t(\bar{t}_1, \bar{t}_2, \hat{t}, \tilde{t})$ **where** $\nu_1 = \kappa_1 = 1$ **and** $\nu_2 = \kappa_2 = 2$

Our first theorem shows that time-threshold-type policies are optimal based on the formulation in Section 4.2. As presented in Appendix EC.1, we introduce dual variables for each constraint in the optimal control. The dual variables corresponding to the differential equations in Eqs. (2) to (7) are called *adjoint vectors*, denoted as $p(t) = (p_j(t), j \in \{1,2,3,4,5,6\})$. These functions play a crucial role in shaping the behavior of the optimal policy, with further details provided in Appendix EC.1.

Theorem 1. *There exist optimal time thresholds $\bar{t}_1^*$, $\bar{t}_2^*$, $\hat{t}^*$, and $\tilde{t}^*$ such that a time-threshold-type policy is optimal.*

The next proposition presents conditions under which these thresholds are within the planning horizon. Otherwise, one or more of these thresholds does not play a role in the optimal policy.

Proposition 2. *The following conditions result in the proposed thresholds to exist within the planning horizon $[0,T]$:*

(i) $\bar{t}_1^* \in [0,T]$, if and only if $(\theta_i\mu_i - \theta_i'\mu_i')p_{i+4}(0) + h_i'\mu_i T + p_{i+2}(0)\mu_i' \geq 0$ for $i = 1,2$.

(ii) $\bar{t}_2^* \in [0,T]$, if and only if $c_i > 0$ and $h_i' \geq -\frac{\theta_i p_{i+4}(0)}{T}$ for $i = 1,2$.

(iii) $\hat{t}^* \in [0,T]$, if and only if $(1-\theta_{\kappa_1}')\mu_{\kappa_1}' c_{\kappa_1} \leq (1-\theta_{\kappa_2}')\mu_{\kappa_2}' c_{\kappa_2}$ and $p_{\kappa_1+2}(0)\mu_{\kappa_1}' - p_{\kappa_2+2}(0)\mu_{\kappa_2}' \leq \theta_{\kappa_1}'\mu_{\kappa_1}' p_{\kappa_1+4}(0) - \theta_{\kappa_2}'\mu_{\kappa_2}' p_{\kappa_2+4}(0)$ for $\kappa_1 \neq \kappa_2$.

(iv) $\tilde{t}^* \in [0,T]$, if and only if $\theta_{\nu_1}\mu_{\nu_1}c_{\nu_1} \geq \theta_{\nu_2}\mu_{\nu_2}c_{\nu_2}$ and $h_{\nu_1}'\mu_{\nu_1} - h_{\nu_2}'\mu_{\nu_2} \geq -\frac{\theta_{\nu_1}\mu_{\nu_1}p_{\nu_1+4}(0) - \theta_{\nu_2}\mu_{\nu_2}p_{\nu_2+4}(0)}{T}$ for $\nu_1 \neq \nu_2$.

In the following sections, we explore more properties of the optimal patient selection policy. In Sections 5.1 and 5.2, we investigate two simpler models that provide insightful perspectives on the optimal policy. Then we return to the original model formulation to generalize these insights in Section 5.3. Two possible extensions beyond the original model formulation are introduced in Section 5.4.

In the remainder of this section, we employ notations and terminologies that are presented in more details in the appendix. Specifically, the switching curve functions $\psi_j(t)$ for $j = 1,2,3,4$ serve as utility functions for the optimal selection policy. This implies that at time t, priority is given to index $k = \text{argmax}_j(\psi_j((t))$, where indices 1 and 2 correspond to new patients of class 1 and 2, and indices 3 and 4 correspond to IP patients of class 1 and 2, respectively. Adjoint vectors $p_j(t)$ serve as dual variables associated with dynamic constraints $\dot{q}_j(t)$ for $j = 1,2,3,4,5,6$. The notation t_e^l is used to specify the time in a physician's shift when a state-dependent constraint becomes active for the l-th time since the start of the shift. In the context of the simpler models explored in Sections 5.1 and 5.2, it is demonstrated that a state-dependent constraint becomes active only once over the planning horizon $[0,T]$, where we use t_e to denote this time during the physician's shift.

5.1. Homogeneous Patients

In this section, we study the special case where there is no distinction between patients of class 1 and class 2, i.e., $\mu_1 = \mu_2$, $\mu_1' = \mu_2'$, $\theta_1 = \theta_2$, $\theta_1' = \theta_2'$, $h_1' = h_2'$, $h_3' = h_4'$, $h_5' = h_6'$, and $c_1 = c_2$. Without

loss of generality, we consider all patients to be of priority class 1. As a result, the patient selection policy is reduced to determining the prioritization between IP and new patients (i.e., the time-threshold-type policy $\pi^t(\bar{t}_1, \bar{t}_2, \hat{t}, \tilde{t})$ reduces to $\pi^t(\bar{t}_1, \bar{t}_2)$). In the following theorem, we show that $\bar{t}_1^* = 0$, meaning that it is optimal to give strict (non-preemptive) priority to IP patients over new patients, unless the holding cost of the new patients is sufficiently large. In that case, $\bar{t}_1^*$ must be sufficiently large such that $t_e > T$. In the proof of this theorem in Appendix EC.2.5, we show that in all cases where $t_e \leq T$, having a strictly positive value for $\bar{t}_1$ only increases the average number of IP patients in the shift while having no effect on the average number of new patients. This will only decrease the value of the objective function as expressed in Eq. (9).

Theorem 2. *When priority classes are homogeneous, $\bar{t}_1^* = 0$ unless all of the following conditions are satisfied:*

(i) $h_1'\mu_1 > h_3'(\theta_1\mu_1 + (1-\theta_1')\mu_1')$,

(ii) $c < \left(h_1'\mu_1 - h_3'(\theta_1\mu_1 + (1-\theta_1')\mu_1')\right) \cdot \frac{T}{\theta_1\mu_1 + (1-\theta_1')\mu_1'}$, *and*

(iii) $\bar{t}_1^*$ *is sufficiently large such that* $t_e > T$.

5.2. Homogeneous IP Patients and Heterogeneous New Patients

In this section, we study the special case where IP patients are homogeneous. This means that no distinction is made between patients of class 1 or class 2 after their initial assessment, i.e., $\mu_1' = \mu_2'$, $\theta_1' = \theta_2'$, $h_3' = h_4'$, $h_5' = h_6'$, and $c_1 = c_2$. As a result, instead of utilizing both $q_3(t)$ and $q_4(t)$, we exclusively employ $q_3(t)$ to represent the number of (all) IP patients awaiting reassessment. Similarly, rather than using $q_5(t)$ and $q_6(t)$ separately, we only use $q_5(t)$ to describe the number of (all) IP patients under delay (before reassessment). In this case, the patient selection policy must consider the prioritization between new patients of class 1 and 2 alongside the prioritization between new and IP patients. Therefore, the time-threshold-type policy $\pi^t(\bar{t}_1, \bar{t}_2, \hat{t}, \tilde{t})$ reduces to $\pi^t(\bar{t}_1, \bar{t}_2, \tilde{t})$.

In the next two theorems, we determine the prioritization of patient classes between new patients.

Theorem 3. *If patient class i has the largest value of $h_i'\mu_i$ and the lowest value of $\theta_i\mu_i$ compared to the other patient classes (i.e., if $h_i'\mu_i > h_j'\mu_j$ and $\theta_i\mu_i < \theta_j\mu_j$ for all $j \neq i$), then patients of class i are always prioritized during the entire shift of a physician when a new patient has to be selected.*

This theorem only specifies which patient class to prioritize if the prioritization of patient classes between new patients does not change (i.e., if $\tilde{t}^* \notin [0, T]$). The following theorem specifies which patient class to prioritize otherwise.

Theorem 4. *Under an optimal time-threshold-type policy $\pi^t(\bar{t}_1^*, \bar{t}_2^*, \tilde{t}^*)$ with $\tilde{t}^* \in [0, T]$:*

(a) When $t \leq \tilde{t}^*$, new patients of class $\nu_1 = \text{argmax}_i\{h_i'\mu_i\}$ are prioritized.

(b) When $t > \tilde{t}^*$, new patients of class $\nu_2 = \text{argmin}_i\{\theta_i\mu_i\}$ are prioritized.

One can easily extend Theorem 4 to cases with more than two patient classes, where new patients are prioritized in a decreasing order of $h_i'\mu_i$ when $t \le \tilde{t}^*$, and new patients are prioritized in an increasing order of $\theta_i\mu_i$ when $t > \tilde{t}^*$. We refer to this class of patient selection rules as the $h'\mu$-rule and the $-\theta\mu$-rule.

We next extend Theorem 2 to Theorem 5 when heterogeneous new patients are allowed. Finding explicit conditions on the value of h_i' is more complicated in these cases compared to the previous section, since such conditions depend on the prioritization of the classes of new patients.

Theorem 5. *When IP patients are homogeneous, $\bar{t}_1^* = 0$ unless the holding costs of new patients are large enough and $\bar{t}_1^*$ is sufficiently large such that $t_e > T$.*

5.3. Heterogeneous IP and New Patients

In this section, we study the optimal patient selection policy for the most general problem formulation as presented in Section 3, where both new and IP patients are both heterogeneous. We examine the prioritization rule between two classes of patients (for both new and IP patients) in the following theorem. It is not possible to derive a simple index-based rule similar to the case of homogeneous IP patients (Section 5.2) due to the intractability that is inherent to the complexities of the model formulation. In particular, the prioritization depends on the adjoint vectors $p_j(t)$, $j = 1, 2, \ldots, 6$, as introduced prior to Theorem 1 and in Appendix EC.1.

Theorem 6. *Prioritization between patients of class 1 and class 2 follows the following time-dependent prioritization rules:*

(a) New Patients: The prioritization between classes of new patients at time $t \in [0, T)$ follows a decreasing order of $\theta_i\mu_i p_{i+4}(t) + h_i'\mu_i(T-t)$, $i = 1, 2$. At the end of the shift (i.e., when $t = T$), prioritization is given to the new patient class with the lowest value of $\theta_i\mu_i c_i$.

(b) IP Patients: The prioritization between classes of IP patients at time $t \in [0, T)$ follows a decreasing order of $\theta_i'\mu_i' p_{i+4}(t) + p_{i+2}(t)$, $i = 1, 2$. At the end of the shift (i.e., when $t = T$), prioritization is given to the IP patient class with the highest value of $(1-\theta_i')\mu_i' c_i$.

Regarding the prioritization between new and IP patients in terms of the existence of a time threshold $\bar{t}_1^*$ within the shift, finding simple inequalities similar to Theorems 2 and 5 is not tractable for this extended model since the conditions depend on the order of the time thresholds (especially the order of $\bar{t}_1^*$, $\hat{t}^*$, $\tilde{t}^*$, t_e^1, and t_e^2). While it is possible to derive explicit inequalities for each given trajectory of the optimal policy, considering the order of time thresholds and the values of t_e^l, there are more than 150 feasible possibilities for the trajectory of the optimal policy, such an approach would be intricate and of limited value.

Instead of a cumbersome proof, we claim that similar results as Theorem 2 and 5 still hold for the general problem formulation, which is that $\bar{t}_1^*$ should always be either zero or large enough such

that $t_e^2 > T$. To make this claim, we first observe that the integral expressions $\int_0^T (q_3(t) + q_4(t))\,dt$ and $\int_0^T (q_5(t) + q_6(t))\,dt$ and the number of patient hand-offs are always increasing in $\bar{t}_1$. When $t_e < T$, then $\int_0^T (q_1(t) + q_2(t))\,dt$ is independent of $\bar{t}_1$. However, when $t_e > T$, then $\int_0^T (q_1(t) + q_2(t))\,dt$ is decreasing. This can be shown similar to the proof of Theorem 2. Consequently, in the case where $t_e < T$, we have $\bar{t}_1^* = 0$. When $t_e > T$, if the holding cost of new patients is significantly higher compared to the holding cost of IP patients and the hand-off penalty cost, it may be optimal to select a strictly positive value for $\bar{t}_1^*$. This value should be sufficiently large such that $t_e > T$, which results in reduced wait times but a substantial increase in patient hand-offs.

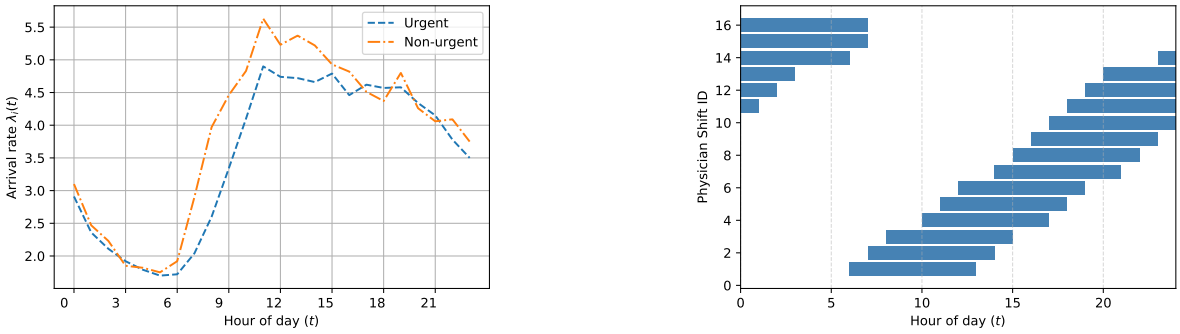
5.4. Model Extensions

Two possible extensions to the model formulation of Section 3 are studied in Appendix EC.3. First, some new ED patients might abandon the system while waiting for their initial assessment and are often labeled as left without being seen (LWBS). For the special case with heterogeneous new and homogeneous IP patients (similar to Section 5.2), we (a) prove that time-threshold-type policies are still optimal in the setting where new patients can abandon the ED (similar to Theorem 1), and (b) extend the prioritization rules for new patients (similar to Theorems 3 and 4). Second, when the holding cost function for new patients is nonlinear, we show that the optimal policy is not a bang-bang policy anymore, and therefore, time-threshold-type policies may not be optimal.

6. Numerical Study

We next perform a comprehensive numerical study to evaluate the performance of our proposed time-threshold-type policy. Since any ED setting will have multiple physicians working at the same time and the patient arrival process will vary throughout the day, we use simulation analyses calibrated with hospital data to create more realistic settings.[4] In particular, we are inspired by a case study from an ED in Canada. In this ED, patients with CTAS levels 1 and 2 are classified as urgent (or class 1), and patients with CTAS levels 3 to 5 as non-urgent (or class 2). The patient arrivals for both classes are assumed to follow a non-homogeneous Poisson process, where the time-dependent arrival rate per hour matches that of actual arrivals illustrated in Figure 3a. In our analyses, we only include patients assigned to the main ED and excluded those who were diverted to the fast track area, since each area in the ED has its own dedicated pool of physicians. The average total number of patients arriving to the main ED is almost 60,000 per year. To provide emergency care to these patients, a total of 16 physicians are scheduled to work during a single day. Figure 3b depicts the physician shift schedule that we adopted in our numerical experiments. Of note, each shift has a duration of seven hours ($T = 7$).

The average duration for initial assessments and reassessments for both patient classes as well as the average delay due to further examinations before reassessments are based on actual data



(a) Time-varying patient arrival pattern

(b) Physician shift schedule with $T = 7$

Figure 3 Patient arrival and physician shift schedule at our partner hospital

as presented in Table 1. The table also presents the estimated probability of a reassessment for both patient classes. Another characteristic commonly observed in the ED is LWBS as discussed in Section 5.4. In particular, some patients leave if their waiting time for an initial assessment by a physician exceeds a certain threshold, a value known as the patience time. In our simulation model, we assume that the patience time for patients in class 1 and class 2 follows an exponential distribution similar to Kamali et al. (2019) and Zayas-Caban et al. (2019). Since the data that is collected only reveals the patience time for LWBS patients (i.e., censored data), we employ the Kaplan-Meier estimator to estimate the average patience times (Kaplan and Meier 1958). This turns out to be $5,119.49$ and $1,836.70$ minutes for patients in class 1 and class 2, respectively. Even though this may seem high, usually less than 5% of the patients become LWBS, and the associated 5-th percentile of the estimated distributions is 4.38 and 1.57 hours for patients in class 1 and class 2, respectively. Additionally, similar parameter values for mean patience times under comparable waiting times are found in the literature (Zayas-Caban et al. 2019), and our simulation produces LWBS percentages comparable to those observed at our partner hospital (Table 2).

We evaluate the performance of any patient selection policy in terms of four main and common ED performance measures: average wait time for initial assessment (denoted by W); average length of care defined as the time from the initial assessment until a disposition decision is made (denoted

Parameter	Notation	Case Study		Baseline
		$i = 1$	$i = 2$	Scenario
Avg. Initial Assessment Time of Class i (minutes)	$1/\mu_i$	20.97	20.98	20.97
Avg. Reassessment Time of Class i (minutes)	$1/\mu_i'$	12.35	11.95	12.13
Avg. Delay Before Reassessment (minutes)	$1/\delta$	13.31	13.31	13.31
Avg. Reassessment Probability for New Patients of Class i	θ_i	62.3%	54.9%	60.0%
Avg. Reassessment Probability for IP Patients of Class i	θ_i'	42.9%	37.1%	40.0%

Table 1 Summary of the parameter values used for the numerical experiments

by LOC)[5]; average percentage of patients who leave without being seen (denoted by $LWBS$); and average number of patient hand-offs per day (denoted by HO). All four measures are studied both at and across patient class levels, where a subscript is used to denote the performance for a particular patient class. More details about our simulation study are presented in Appendix EC.4.

Table 2 summarizes the performance measures observed in our dataset alongside those obtained from simulations of optimal and alternative policies. To better understand the current patient selection policy in practice, we replicated the performance measures from our dataset (row 1) using a simulation model that is specifically designed to mimic the performance of current practices (row 2). Our observations indicate that the behavior of physicians in selecting their next patients can be approximated using a randomization approach between different patient classes and between IP and new patients. Prioritization between IP and new patients in this approach occurs with a probability of 0.5. The probabilities of prioritizing patients of class 1 over class 2 are 0.56 and 0.6 for new and IP patients, respectively. Furthermore, we find that physicians implement a cutoff time of around 90 minutes before the end of the shift after which no new patients are accepted. With these, our simulation results (row 2 in Table 2) are fairly similar to the values observed in the data (row 1).

Before analyzing the performance of our proposed time-threshold-type policies, we also consider the following simple alternative policies. In Policy 1, patients of class 1 are given strict (non-preemptive) priority over patients of class 2 (while the prioritization between new and IP patients remains randomized). In Policy 2, IP patients are given strict (non-preemptive) priority over new patients (while the prioritization between patients in class 1 and class 2 remains randomized). In Policy 3, patients of class 1 are given strict (non-preemptive) priority over patients of class 2 and IP patients are given strict (non-preemptive) priority over new patients. For Policy 1, the cutoff time is set to 90 minutes, consistent with current practice. For Policy 2 and Policy 3, we consider two variations with different cutoff times: 90 minutes for Policy 2A and 3A, and 45 minutes for Policy 2B and 3B. Finally, we explore eight settings with different cost parameters to determine optimal time-threshold-type policies, as presented in Table 3. Of note, the last columns in this table represent the threshold values of the corresponding optimal policies.

In all these settings $\bar{t}_1^* = 0$, which means that IP patients are prioritized over new patients for the entire duration of each shift. As discussed in Section 4 (see Theorems 2 and 5), a non-zero value for $\bar{t}_1^*$ results in scenarios where the waiting time for initial assessments of patients is low, but this comes at the expense of a long length of stay and many patient hand-offs. Therefore, we refrain from considering such scenarios in our numerical experiments. In settings where the holding cost for new patients of class 1 (h_1') is higher than that of class 2 (h_2') (Settings 1 to 5 and 8), we observe a large value of $\tilde{t}^*$. This indicates that class 1 patients are prioritized for most of the shift

Policy		W_1 (hrs)	W_2 (hrs)	W (hrs)	LOC_1 (hrs)	LOC_2 (hrs)	LOC (hrs)	HO_1	HO_2	HO	$LWBS_1$ (%)	$LWBS_2$ (%)	LWBS (%)
Current Practice	Data	1.40	2.03	1.72	2.97	1.94	2.43	10.72	7.04	17.76	2.65	6.76	4.85
	Simulation	1.41	2.05	1.72	2.10	1.97	2.04	12.45	11.38	23.83	1.79	7.01	4.38
Simple Policies	1	0.27	2.80	1.46	1.63	2.51	2.05	5.92	18.35	24.28	0.34	9.40	4.84
	2A	3.58	4.70	4.10	1.16	1.01	1.09	2.38	1.20	3.58	4.23	15.13	9.64
	2B	1.33	1.95	1.63	1.09	1.11	1.10	5.25	4.88	10.13	1.62	6.57	4.08
	3A	0.36	6.37	3.01	1.07	1.13	1.10	1.80	1.67	3.47	0.43	20.11	10.20
	3B	0.25	2.57	1.35	1.07	1.13	1.10	5.04	4.98	10.02	0.32	8.58	4.42
Optimal Policies	Setting 1	0.32	2.26	1.24	1.07	1.13	1.10	3.53	7.88	11.41	0.38	7.59	3.96
	Setting 2	0.35	2.26	1.26	1.07	1.13	1.10	3.22	8.36	11.58	0.44	7.60	4.00
	Setting 3	0.38	2.22	1.26	1.08	1.13	1.10	2.75	8.58	11.33	0.50	7.47	3.96
	Setting 4	0.29	2.43	1.31	1.07	1.13	1.10	4.18	6.48	10.66	0.38	8.11	4.22
	Setting 5	0.29	2.51	1.34	1.14	1.06	1.10	4.61	5.67	10.28	0.37	8.43	4.38
	Setting 6	3.64	0.24	1.92	1.06	1.15	1.11	5.81	6.61	12.42	4.49	0.86	2.69
	Setting 7	4.06	0.25	2.13	1.14	1.08	1.11	6.00	5.46	11.45	4.97	0.82	2.91
	Setting 8	0.29	2.59	1.38	1.07	1.13	1.10	3.89	5.82	9.71	0.34	8.73	4.51

Table 2 Comparison of various policies

	Cost Parameters								Optimal Thresholds			
	c_1	c_2	h_1'	h_2'	h_3'	h_4'	h_5'	h_6'	$\bar{t}_1^*$	$\tilde{t}^*$	$\hat{t}^*$	$\bar{t}_2^*$
Setting 1	20.00	15.00	2.50	2.00	1.00	0.90	0.10	0.10	0.00	5.70	7.00	6.35
Setting 2	30.00	15.00	2.50	2.00	1.00	0.90	0.10	0.10	0.00	5.45	7.00	6.35
Setting 3	45.00	15.00	2.50	2.00	1.00	0.90	0.10	0.10	0.00	5.26	7.00	6.35
Setting 4	20.00	19.00	2.50	2.00	1.00	0.90	0.10	0.10	0.00	5.86	6.74	6.29
Setting 5	20.00	20.00	2.50	2.00	1.00	0.90	0.10	0.10	0.00	5.92	0.00	6.28
Setting 6	20.00	15.00	2.50	2.50	1.00	0.90	0.10	0.10	0.00	0.00	7.00	6.40
Setting 7	20.00	20.00	2.50	2.50	1.00	0.90	0.10	0.10	0.00	0.00	0.00	6.33
Setting 8	35.00	25.00	2.90	2.00	1.00	0.90	0.10	0.10	0.00	5.92	7.00	6.23

Table 3 Cost parameters and optimal thresholds for different settings

when selecting new patients. This significantly reduces the average wait time for class 1 (urgent) patients while it increases the average wait time for class 2 (non-urgent) patients compared to the current practice. However, the overall average wait times are substantially shorter under our proposed policies. In addition, as the relative hand-off penalty cost for patients of class 1 and class 2 increases, the value of $\tilde{t}^*$ decreases. On the other hand, as the results for Settings 6 and 7 show, when the holding cost for new patients of the two classes is similar ($h_1' = h_2'$), the optimal policy strictly prioritizes new patients of class 2 over new patients of class 1 (i.e., $\tilde{t}^* = 0$). This occurs due to the lower hand-off penalty cost of class 2 patients, as well as a lower reassessment probability and average reassessment time for these patients. As a result, the average wait time for patients of class 1 increases significantly, while the average wait time for patients of class 2 decreases. Since the average patience time for patients of class 1 is significantly longer than for patients of class 2, these policies result in a lower LWBS, which increases the overall wait times due to an increase in the effective arrival rate.

In terms of prioritizing IP patients, we observe the following: In Settings 1, 2, 3, 6, and 8, where c_1 is much greater than c_2, patients of class 1 are strictly prioritized over patients of class 2

when selecting IP patients (i.e., $\hat{t}^* = 7$). In Setting 4, where c_1 and c_2 are close but not equal, the prioritization switches to patients of class 2 later in the shift (i.e., $0 < \hat{t}^* < 7$). In Settings 5 and 7, where $c_1 = c_2$, IP patients of class 2 are strictly prioritized over IP patients of class 1 (i.e., $\hat{t}^* = 0$). Changes in $\hat{t}^*$ have hardly any impact on the overall average LOC, while they only slightly affect class-level LOCs. Finally, we observe that the optimal cutoff time to stop accepting new patients is between 36 and 46 minutes before the end of the shift (since $6.23 < \bar{t}_2^* < 6.40$), which is significantly later than what occurs in practice. Later cutoff times are incorporated in Policies 2B and 3B.

Overall, when comparing our policies to current practices, we observe that our policies consistently perform better across all four performance measures when new patients of class 1 are prioritized over new patients of class 2 for most of the shift duration (Settings 1 to 5 and 8). Among the simple policies, only Policies 2B and 3B consistently outperform current practices, highlighting the potential benefit of strictly prioritizing IP patients over new patients while implementing a later cutoff time. Between these two simple policies, the latter outperforms the former, which indicates that strictly prioritizing patients of class 1 over patients of class 2 is more advantageous than a randomized prioritization between the patient classes. While Policy 3B achieves comparable results to our proposed optimal policies, we achieve fewer patient hand-offs while maintaining similar average wait times and LOC in Setting 8. Furthermore, our proposed policy is more dynamic and capable of balancing class-level performance measures.

6.1. Sensitivity Analysis and Policy Insights

As Table 1 shows, there is not much difference between the parameter values for patients in class 1 and class 2 for the ED in our case study. As a result, we observe minimal differences between the performance of different time-threshold-type policies in Table 2. To examine the impact of an imbalance between patient classes, we first specify a baseline scenario where all parameter values for both patient classes are identical (similar to the case of homogeneous patients from Section 5.1, with the exception that cost parameters have different values). This baseline scenario is inspired by our case study and is presented in the last column of Table 1. Next, we investigate the sensitivity of the patient selection policy with respect to the parameter values of the baseline scenario.

When there is an imbalance between the two patient classes, we are especially interested in scenarios where the following holds for patients in class 2 in comparison to class 1 : the arrival rates are greater, initial assessment and reassessment times are shorter, and reassessment probabilities are lower. To quantify the difference between patient classes, we consider four parameters: Δ_λ, $\Delta_{1/\mu}$, $\Delta_{1/\mu'}$ and Δ_θ, which denote the percentage deviation from the baseline scenario in arrival rates, initial assessment times, reassessment times, and reassessment probabilities, respectively. We let these percentage deviations vary between 0 to 50%. For example, when Δ_λ is 10%, $\lambda_1(t)$

is decreased by 10% and $\lambda_2(t)$ is increased by 10% compared to the baseline scenario. Since all other parameter values are balanced in the baseline scenario (i.e., they are the same across the two patient classes), the total workload of the queueing system in Figure 1 remains the same compared to the baseline scenario. Furthermore, we set $h_1' = 2.5$, $h_2' = 1.7$, $h_3' = 1.0$, $h_4' = 0.9$, $h_5' = h_6' = 0.1$, $c_1 = 60.0$, and $c_2 = 45.0$ in our numerical experiments. Consequently, $\bar{t}_1^* = 0$ for all instances.

6.1.1. Insights into the Patient Selection Behavior of the Proposed Policy For the baseline scenario, since all parameter values are the same for patients of class 1 and 2, but the cost parameters are greater for patients in class 1, the optimal policy will prioritize class 1 patients. For IP patients, we know from Theorem 6 that the prioritization at the end of the shift will be given to the patient class with the highest value of $(1-\theta_i')\mu_i' c_i$. Since $c_1 > c_2$ and the other two parameters are equal, IP patients in class 1 will still have priority at the end of the physician's shift. Consequently, $\hat{t}^* > 7$, which means that IP patients in class 1 get strict prioritization over IP patients in class 2 for the entire duration of the physician's shift in the baseline scenario. Regarding the prioritization between new patients, Theorem 6 states that the prioritization at the end of the shift will be given to the patient class with the lowest value of $\theta_i \mu_i c_i$. Since $c_1 > c_2$, new patients in class 2 will eventually get prioritization during the physician's shift (i.e., $\tilde{t}^* < 7$). However, despite the significant difference between c_1 and c_2 as well as the difference between h_1' and h_2', the value of $\tilde{t}^*$ is large enough such that $\tilde{t}^* > \bar{t}_2^*$ in the baseline scenario. This means that new patients in class 1 will always have strict priority over new patients in class 2 (similar to IP patients).

The optimal policy is independent of the arrival rates (due to Assumption 1), which means that changing Δ_λ does not impact the optimal policy. Table 4 presents the optimal policies associated with the baseline scenario as well as for different levels of $\Delta_{1/\mu}$, $\Delta_{1/\mu'}$ and Δ_θ. Note that in this table, the values highlighted in gray indicate a prioritization rule that is in the opposite direction of a regular (or intuitive) prioritization rule. This means that new patients in class 2 are prioritized at the start of the shift and the prioritization switches to new patients in class 1 after time threshold $\tilde{t}^*$. This only happens when changing initial assessment times (i.e., for $\Delta_{1/\mu}$). In contrast, when Δ_θ increases, the prioritization among new patients switches from class 1 to class 2 after $\tilde{t}^*$.

Time	Baseline	$\Delta_{1/\mu}$					$\Delta_{1/\mu'}$					Δ_θ				
Threshold	Scenario	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
$\bar{t}_2^*$	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.06	6.12	6.19	6.25	6.32
$\tilde{t}^*$	$> \bar{t}_2^*$	$> \bar{t}_2^*$	$> \bar{t}_2^*$	2.13	5.91	5.95	$> \bar{t}_2^*$	$> \bar{t}_2^*$	$> \bar{t}_2^*$	$> \bar{t}_2^*$	$> \bar{t}_2^*$	5.67	5.28	4.40	0.00	0.00
$\hat{t}^*$	$> T$	$> T$	$> T$	$> T$	$> T$	$> T$	$> T$	$> T$	$> T$	0.00	0.00	$> T$	$> T$	6.86	6.72	6.59

Table 4 Time thresholds for the optimal patient selection policies

Note: Gray cells indicate instances where new patients in class 2 are prioritized before time threshold $\tilde{t}^*$, and new patients in class 1 are prioritized thereafter.

In the remainder of this section, we highlight some insights into how the optimal patient selection policy changes when there is an imbalance between patient classes (supported by Table 4).

- The timing when to stop accepting new patients does not depend on the initial assessment or reassessment times (i.e., $\bar{t}_2^*$ remains constant in $\Delta_{1/\mu}$ and $\Delta_{1/\mu'}$). According to the definition of switching curves in Eq. (EC.7) in Appendix EC.1, we have $\psi_1(t) = (\theta_1 p_5(t) + p_1(t))\mu_1$. Recall from the proof of Theorem 1 that $\psi_1(t) = 0$ at $t = \bar{t}_2^*$. Consequently, $\theta_1 p_5(t) + p_1(t) = 0$ at $t = \bar{t}_2^*$, which is independent of μ_1 or μ_1', but it depends on other parameters such as the reassessment probabilities (see scenarios Δ_θ) and cost parameters.
- When the initial assessment times for patients in class 1 increase, the physician should prioritize patients in class 2 among the new patients at the start of the shift (according to part (a) of Theorem 6 and consistent with the $c\mu$-rule). However, due to the penalty cost for hand-offs at the end of the shift, new patients in class 1 will be given priority during the second half of the shift. The latter is also formalized by Theorem 6 (since $\theta_1\mu_1 c_1 < \theta_2\mu_2 c_2$). This can be seen as counter-intuitive. However, when the physician starts to prioritize patients with longer initial assessment times closer to the end of the shift, the pick-up rate of new patients will be lower, and therefore, the number of patient hand-offs decreases. In other words, one can see $\theta_i\mu_i$ as a proxy for the rate at which new patients of class i become IP patients close to the end of the shift.
- When the reassessment times of patients in class 1 increase, the physician needs to prioritize IP patients in class 2 at the end of the shift since it is likely that these patients can be dispositioned before the end of the shift, whereas it is very probable that the IP patients in class 1 end up as hand-offs regardless. At the same time, there is no impact on the prioritization of new patients (or $\tilde{t}^*$). As stated in the proof of Theorem 6 (Appendix EC.2.7), the patient class with the highest value of switching curve $\psi_i(t)$ gets priority among the new patients. Increasing reassessment times for a patient class i, will decrease $\psi_i(t)$ for that class. However, the time t where $\psi_1(t)$ equals $\psi_2(t)$ remains exactly the same. In contrast, this intersection time is more sensitive to initial assessment times, reassessment probabilities, and cost parameters (as seen in the other scenarios).
- When the likelihood of reassessments for patients in class 1 increases, the physician will start to prioritize patients from class 2 among the new patients earlier in the shift (and at some point even from the start of the shift). At the same time, the physician will continue to accept new patients for a longer period in the shift (since the physician prioritizes patients in class 2 at that time and these patients have a smaller probability of reassessment). Furthermore, there is less of an impact on how IP patients are prioritized (i.e., $\hat{t}^*$ is relatively stable).

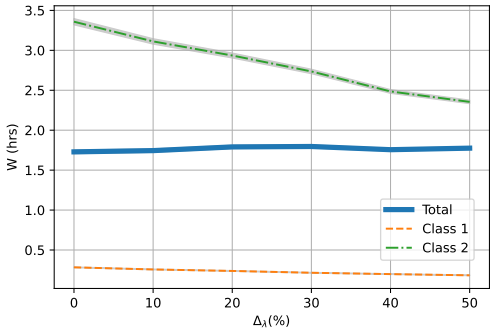
6.1.2. Performance Evaluation We next examine the performance impact of the optimal policy when the aforementioned imbalance is introduced. Figure 4 illustrates the performance

as we adjust the arrival rates according to Δ_λ, where the solid lines represent the performance measure as an average over all patients, and the dashed lines specify the performance per patient class. Additionally, the gray areas around these lines represent 95% confidence intervals since our numerical results are based on simulation. Since fewer patients of class 1 arrive to the system when Δ_λ increases, there are less class 1 patients who are handed off at the end of the shift (and the reverse holds for class 2 patients). However, the total number of patient hand-offs remains nearly constant across all scenarios (Figure 4d). In terms of average waiting time, it is obvious that $W_1 < W_2$ regardless of Δ_λ, since new and IP patients of class 1 always have priority over patients of class 2 during the entire shift. When more patients of class 2 arrive to the system (who have a longer wait time), one can expect the overall wait time (i.e., W) to increase as well. However, there are also fewer patients of class 1 who get served before class 2 patients. Consequently, more patients are seen based on a FCFS regime, which lowers W_2. This effect dampens the increase of W. It is interesting to observe that even though there are more patients with a shorter average patience time (since $\lambda_2(t)$ increases), the number of patients who leave the system without being seen (i.e., $LWBS$) remains constant. Finally, Figure 4b shows that the length of care (LOC_i) also decreases for both patient classes, with almost no effect on LOC, following a similar reason as the one described above for W.

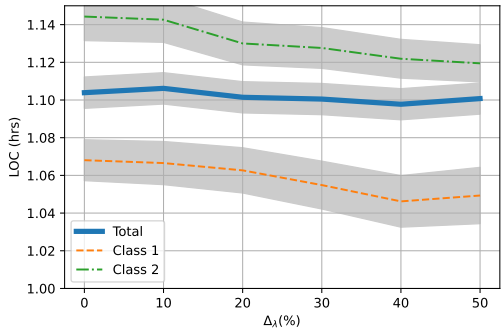
Figure 5 presents the changes in performance when the initial assessment times are varied. In the scenarios where $\Delta_{1/\mu} < 30\%$, patients of class 1 gain the priority among new patients such that $W_1 < W_2$ for these scenarios (Figure 5a). In the scenario where $\Delta_{1/\mu} = 30\%$, the optimal policy takes a balanced approach in prioritizing both patient classes among new patients over the duration of the shift. Consequently, W_1 and W_2 are similar. However, in scenarios where $\Delta_{1/\mu} > 30\%$, more priority is given to new patients of class 2 (see also Table 4) such that $W_2 < W_1$. When new patients of class 2 are prioritized even longer (i.e., when $\Delta_{1/\mu} > 30\%$), there will be more patients of class 2 who are handed off at the end of the shift again, and there is less time to accept new patients of class 1 at the end of the shift, which lowers HO_1 again (but it increases W_1 drastically). Changing initial assessment times has a direct impact on the length of care as a greater initial assessment time $1/\mu_1$ increases LOC_1 (and the reverse holds for patients in class 2; see Figure 5b). Finally, the results of varying the average reassessment times and reassessment probabilities follow a similar logic, and hence, are presented in Appendix EC.5.1.

6.2. Comparison to Alternative Policies

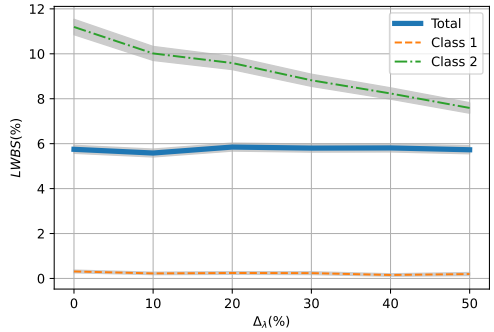
We next compare the performance of our proposed patient selection policy against three alternative policies from the literature. First, in the *static policy*, IP patients are strictly prioritized over new patients, and class 1 patients are strictly prioritized over class 2 patients. This corresponds to the



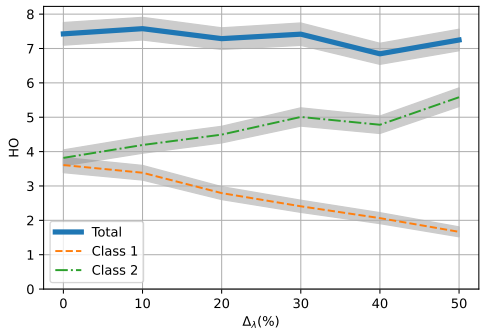
(a) Average wait time for initial assessment



(b) Average length of care



(c) Average percentage of patients left unseen

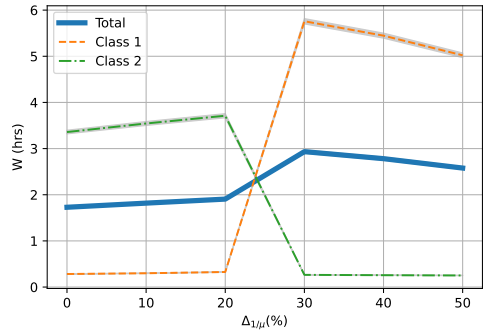


(d) Average number of patient hand-offs (per day)

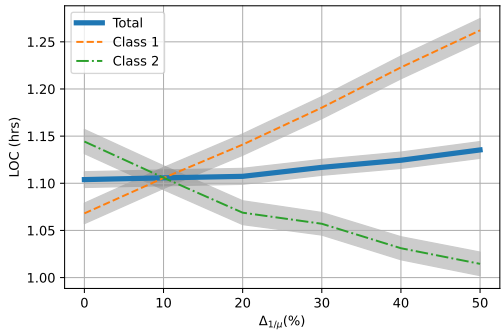
Figure 4 The impact of varying arrival rates: a decrease for class 1 and an increase for class 2 by Δ_λ

Note: Gray bands are not visible in some subplots because confidence intervals are too tight.

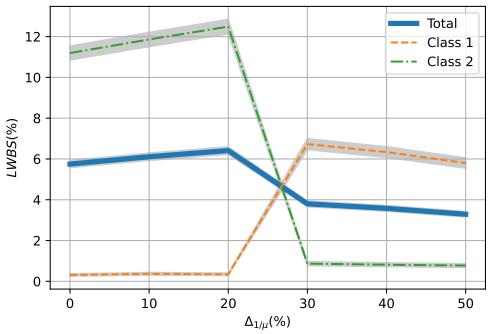
$c\mu$-rule (Cox and Smith 1961). The other two policies are based on Ouyang et al. (2021), which is the only patient selection policy in the literature that considers a time threshold such that no new patients are selected by the physician beyond it (similar to $\bar{t}_2$). Furthermore, IP patients are given strict priority over new patients. Since Ouyang et al. (2021) do not consider different patient classes, we study two settings: one where prioritization follows a first-come-first-serve (FCFS) discipline (regardless of patient classes) and another where patients in class 1 are given fixed priority. For our proposed policy, we use the same parameter values as the baseline scenario in Section 6.1. To allow for a fair comparison between the performance of our proposed policy and that of Ouyang et al. (2021), we set the penalty cost for hand-offs in the latter policy such that the threshold to stop accepting new patients results in the exact same number of patient hand-offs as our proposed policy. Consequently, our proposed policy and that of Ouyang et al. (2021) result in the exact same selection rules in the base scenario (since $\bar{t}_1 = 0$, $\hat{t}^* > T$ and $\tilde{t}^* > \bar{t}_2$). We perform the same sensitivity analysis as discussed in Section 6.1. The detailed results are presented in Appendix EC.5.2 and the insights gained are summarized in the remainder of this section.



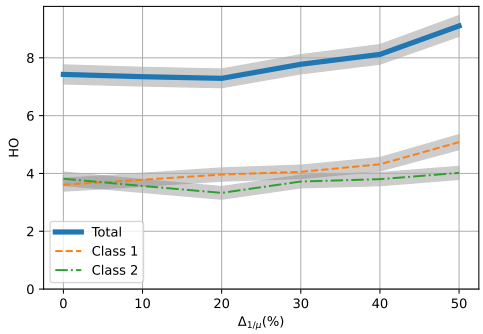
(a) Average wait time for initial assessment



(b) Average length of care



(c) Average percentage of patients left unseen



(d) Average number of patient hand-offs (per day)

Figure 5 The impact of varying initial assessment times: an increase for class 1 and a decrease for class 2 by $\Delta 1/\mu$

Note: Gray bands are not visible in some subplots because confidence intervals are too tight.

First, adopting a cutoff time $\bar{t}_2$ reduces the number of patient hand-offs significantly (compared to the static policy), but it increases the wait times and number of patients who leave without being seen ($LWBS$). Second, a FCFS regime in combination with the policy in Ouyang et al. (2021) results in noticeably longer wait times but less LWBS patients compared to Ouyang et al. (2021) with a fixed prioritization (i.e., always prioritizing class 1). Third, when comparing our proposed policy with Ouyang et al. (2021) with a fixed prioritization between patient classes, we observe that our proposed policy can reduce LWBS while maintaining similar levels of wait times and patient hand-offs in most scenarios. This is because our proposed policy switches to prioritizing patients in class 2 over class 1 for some period of time during the physician's shift (or even for the entire duration of the shift in some scenarios). Since patients in class 1 have more patience, there are fewer LWBS patients (i.e., the throughput of the ED increases), which prevents the average wait times to decrease. Lastly, we note that there is no strict prioritization of one patient class for the entire duration of the physician's shift under our proposed policy. Instead, the prioritization between patient classes switches during the shift. This lowers the 90-th percentile

for the performance measures (i.e., extreme values become less frequent under our proposed time-threshold-type policies), and the difference in average performance measures between the patient classes is smaller. This means that a mixture of prioritizing patients of the classes 1 and 2 provides a better balance than strictly prioritizing patients of class 1, but only as long as patients in class 1 are given priority for most of the time during the physician's shift.

7. Conclusion

Patient selection in EDs is a complicated task that can have a significant impact on operational performance and clinical outcomes. Factors such as patient heterogeneity, the need for reassessments, and patient hand-offs at the end of a physician's shift contribute to the complexity of this task. By adopting optimal control theory, our study shows that a time-threshold-type policy is optimal in deciding who should be seen next by a physician. This optimal policy employs two mechanisms to manage patient hand-offs: (1) a *cutoff mechanism*, which establishes a time threshold ($\vec{t}_2^*$) near the end of the shift to prevent from accepting new patients, and (2) a *switching mechanism*, which involves two time thresholds: one for new patients ($\tilde{t}^*$) and one for IP patients ($\hat{t}^*$), after which the priority switches in favor of the patient class that leads to a reduction in the number of patient hand-offs. This nuanced mixture of different patient classes being prioritized at different times during a physician's shift helps to improve the operational performance measures.

In addition, our proposed policy provides multiple insights into ways of supporting physician decisions in terms of patient selection. First, physicians should always prioritize IP patients over new patients when reducing patient hand-offs. Only when the wait time of new patients is most relevant, but not the length of care for IP patients or the number of patient hand-offs, new patients should be prioritized. Second, by studying a special case of our original problem formulation, we show that it is optimal to adopt an index-based policy rule to prioritize new patients with different triage levels (urgency levels). Specifically, at the start of the shift, the patient class with a higher value of $h_i'\mu_i$ should be prioritized, while the patient class with a lower value of $\theta_i\mu_i$ should be prioritized at the end of the shift. From our numerical experiments, we conclude that initial assessment times and reassessment probabilities are the most important factors that affect the optimal patient selection policy, thereby having a significant impact on the operational performance measures.

Sensitivity analysis and robustness checks show that our proposed patient selection policies perform well under more general settings than what is explicitly considered in our analysis. By varying the arrival rates of patient classes, we studied the performance under different system workloads (Section 6.1). Even tough our proposed policy does not depend on the state in terms of system or physician workload, our results show that the policy still performs well under these

different workload conditions. Additionally, the overall performance is also not significantly affected when patients are misclassified into an incorrect patient class during triage (Appendix EC.5.3).

In closing, we note that our study has several limitations. First, we only consider two patient classes, whereas patients are more diverse in reality. However, we believe that our prioritization rules and insights into patient selections extend to settings with multiple patient classes. Second, literature suggests that physicians tend to prioritize discharge patients over admit patients among the same triage level when there is a shortage in beds at inpatient wards (Li et al. 2023), and physicians tend to see more patients during a shift when there is greater familiarity among ED physicians who are scheduled to work at the same time (Niewoehner III et al. 2023). Consequently, disposition predictions and physician familiarity can also be considered in patient selection policies to shed more light on actionable insights that can further improve operational metrics in EDs.

Endnotes

1. In case of zero delay, we refer interested readers to Zhan and Ward (2014) and references therein.

2. While our focus is on EDs, we note that, the same formulation can be used for other environments that are highly congested and prioritization decisions between new and in-progress customers needs to be made.

3. We extend our analysis by considering the fact that the ED arrival rate is time-dependent in Section 6.

4. Of note, including multiple physicians does not impact our proposed policy.

5. This measure does not include any ED boarding time for patients admitted to the hospital post ED visit.

References

Abir M, Goldstick JE, Malsberger R, Williams A, Bauhoff S, Parekh VI, Kronick S, Desmond JS (2019) Evaluating the impact of emergency department crowding on disposition patterns and outcomes of discharged patients. *International Journal of Emergency Medicine* 12(4).

Atar R, Giat C, Shimkin N (2010) The cμ/θ rule for many-server queues with abandonment. *Operations Research* 58(5):1427–1439.

Atkinson MK, Saghafian S (2023) Who should see the patient? On deviations from preferred patient-provider assignments in hospitals. *Health Care Management Science* 26(2):165–199.

Barjesteh N, Abouee-Mehrizi H (2021) Multiclass state-dependent service systems with returns. *Naval Research Logistics (NRL)* 68(5):631–662.

Batt RJ, KC DS, Staats BR, Patterson BW (2019) The effects of discrete work shifts on a nonterminating service system. *Production and Operations Management* 28(6):1528–1544.

Bullard M, Musgrave E, Warren D, Unger B, Skeldon T, Grierson R, van der Linde E, Swain J (2017) Revisions to the Canadian emergency department triage and acuity scale (CTAS) guidelines 2016. *Canadian Journal of Emergency Medicine* 19(S2):S18–27.

Campello F, Ingolfsson A, Shumsky RA (2017) Queueing models of case managers. *Management Science* 63(3):882–900.

Carter EJ, Pouch SM, Larson EL (2014) The relationship between emergency department crowding and patient outcomes: A systematic review. *Journal of Nursing Scholarship* 46(2):106–115.

Chan CW, Yom-Tov G, Escobar G (2014) When to use speedup: An examination of service systems with returns. *Operations Research* 62(2):462–482.

Chan TC, Huang SY, Sarhangian V (2024) Dynamic control of service systems with returns: Application to design of postdischarge hospital readmission prevention programs. *Operations Research* .

Cox D, Smith W (1961) *Queues* (Methuen & Co. Ltd., London).

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

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

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

Ferrand YB, Magazine MJ, Rao US, Glass TF (2018) Managing responsiveness in the emergency department: Comparing dynamic priority queue with fast track. *Journal of Operations Management* 58:15–26.

Furman E, Diamant A, Kristal M (2021) Customer acquisition and retention: A fluid approach for staffing. *Production and Operations Management* 30(11):4236–4257.

He S, Sim M, Zhang M (2019) Data-driven patient scheduling in emergency departments: A hybrid robust-stochastic approach. *Management Science* 65(9):4123–4140.

Hu Y, Chan CW, Dong J (2022) Optimal scheduling of proactive service with customer deterioration and improvement. *Management Science* 68(4):2533–2578.

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

Ingolfsson A, Almehdawe E, Pedram A, Tran M (2020) Comparison of fluid approximations for service systems with state-dependent service rates and return probabilities. *European Journal of Operational Research* 283(2):562–575.

Janke AT, Melnick ER, Venkatesh AK (2022) Rates of patients who left before accessing care in US emergency departments, 2017-2021. *JAMA Netw Open* 5(9), URL `http://dx.doi.org/10.1001/jamanetworkopen.2022.33708`.

Jones PG, Mountain D, Forero R (2021) Review article: Emergency department crowding measures associations with quality of care: A systematic review. *Emergency Medicine Australasia* 33(4):592–600.

Kamalahmadi M, Bretthauer KM, Helm JE, Mills AF, Coe EC, Judy-Malcolm A, Kara A, Pan J (2023) Mixing it up: Operational impact of hospitalist caseload and case-mix. *Management Science* 69(1):283–307.

Kamali MF, Tezcan T, Yildiz O (2019) When to use provider triage in emergency departments. *Management Science* 65(3):1003–1019.

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282):457–481.

Li W, Sun Z, Hong LJ (2023) Who is next: Patient prioritization under emergency department blocking. *Operations Research* 71(3):821–842.

Liu Y, Whitt W (2017) Stabilizing performance in a service system with time-varying arrivals and customer feedback. *European Journal of Operational Research* 256(2):473–486.

Maglaras C (2006) Revenue management for a multiclass single-server queue via a fluid model analysis. *Operations Research* 54(5):914–932.

Mandelbaum A, Stolyar A (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research* 52(2):836–855.

Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L (2018) Emergency department crowding: A systematic review of causes, consequences and solutions. *PLoS One* 13(8), URL `http://dx.doi.org/10.1371/journal.pone.0203316`.

Niewoehner III RJ, Diwas K, Staats B (2023) Physician discretion and patient pick-up: How familiarity encourages multitasking in the emergency department. *Operations Research* 71(3):958–978.

Ouyang H, Liu R, Sun Z (2021) Emergency department modeling and staffing: Time-varying physician productivity. *Available at SSRN 3963226* .

Pearce S, Marchand T, Shannon T, Ganshorn H, Lang E (2023) Emergency department crowding: an overview of reviews describing measures causes, and harms. *Internal and Emergency Medicine* 1137–1158.

Rasouli HR, Esfahani AA, Nobakht M, Eskandari M, Mahmoodi S, Goodarzi H, Farajzadeh MA (2019) Outcomes of crowding in emergency departments; a systematic review. *Archives of Academic Emergency Medicine* 7(1).

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

Saghafian S, Hopp WJ, Iravani SM, Cheng Y, Diermeier D (2018) Workload management in telemedical physician triage and other knowledge-based service systems. *Management Science* 64(11):5180–5197.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* 60(5):1080–1097.

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

Saghafian S, Kilinic D, Traub S (2024) Dynamic assignment of patients to primary and secondary inpatient units: Is patience a virtue? *The Cambridge Handbook of Healthcare: Productivity, Efficiency, Effectiveness*, volume 2024, 612–656 (Cambridge University Press).

Sethi SP (2019) *Optimal Control Theory Applications to Management Science and Economics* (Springer Nature Switzerland AG), 3rd edition, ISBN 978-3-319-98236-6.

Shakeri M, Haji B, Farrokhvar L (2023) A partially flexible routing strategy for assigning emergency department patients to inpatient wards. *Computers & Industrial Engineering* 176:108810.

Sun BC, Hsia RY, Weiss RE, Zingmond D, Liang LJ, Han W, McCreath H, Asch SM (2013) Effect of emergency department crowding on outcomes of admitted patients. *Annals of Emergency Medicine* 61(6):605–611.

Traub SJ, Bartley AC, Smith VD, Didehban R, Lipinski CA, Saghafian S (2016a) Physician in triage versus rotational patient assignment. *The Journal of Emergency Medicine* 50(5):784–790.

Traub SJ, Stewart CF, Didehban R, Bartley AC, Saghafian S, Smith VD, Silvers SM, LeCheminant R, Lipinski CA (2016b) Emergency department rotational patient assignment. *Annals of Emergency Medicine* 67(2):206–215.

Van Mieghem JA (1995) Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *The Annals of Applied Probability* 5(3):809–833.

Wolf L, Ceci K, McCallum D, Brecher D (2023) Emergency severity index handbook, 5th edition .

Yankovic N, Green LV (2011) Identifying good nursing levels: A queuing approach. *Operations Research* 59(4):942–955.

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

Zaerpour F, Bijvank M, Ouyang H, Sun Z (2022) Scheduling of physicians with time-varying productivity levels in emergency departments. *Production and Operations Management* 31(2):645–667.

Zayas-Caban G, Xie J, Green LV, Lewis ME (2019) Policies for physician allocation to triage and treatment in emergency departments. *IISE Transactions on Healthcare Systems Engineering* 9(4):342–356.

Zhan D, Ward AR (2014) Threshold routing to trade off waiting and call resolution in call centers. *Manufacturing & Service Operations Management* 16(2):220–237.

Zychlinski N (2023) Applications of fluid models in service operations management. *Queueing Systems* 103(1-2):161–185.