Emergency Department Modeling and Staffing: Time-Varying Physician Productivity

Huiyin Ouyang
Faculty of Business and Economics, The University of Hong Kong, Pok Fu Lam Road, Hong Kong

Ran Liu
Department of Industrial Engineering and Management, Shanghai Jiao Tong University, China

Zhankun Sun
Department of Management Sciences, College of Business, City University of Hong Kong, Kowloon, Hong Kong
zhankun.sun@cityu.edu.hk

Motivated by an intriguing observation of a time-varying pattern in physician productivity (measured by the number of new patients seen per hour, or PPH), we study a continuous-time optimal control problem to understand the transient behavior of individual physicians within their shifts in emergency departments (EDs). By applying Pontryagin's maximum principle, we characterize the optimal policy and provide insights into physician capacity, productivity, and throughput. We conclude that individual physicians' transient behavior is intrinsic and mainly induced by shift-based scheduling. We leverage the insights from time-varying PPH to model a complex ED system as a time-varying multi-server queue with shift-hour-dependent service rates. Validated using data from two Canadian EDs, our simulation results show that our queueing model can accurately capture time-of-day-dependent patient waiting times with a simple parameter estimation procedure. In contrast, the simulated waiting times under constant service rates deviate significantly from the data. Hence, it is important to explicitly consider time-varying service rates to obtain accurate models of ED operations. The essence of our model is dimension reduction by state aggregation. As a result, the model allows for performance evaluation through the uniformization of a continuous-time Markov chain, which can be integrated with off-the-shelf algorithms for physician staffing. Our case study using data from a Canadian ED shows that the new shift schedules generated using our method can improve the current schedule in practice and result in substantial annual cost savings.

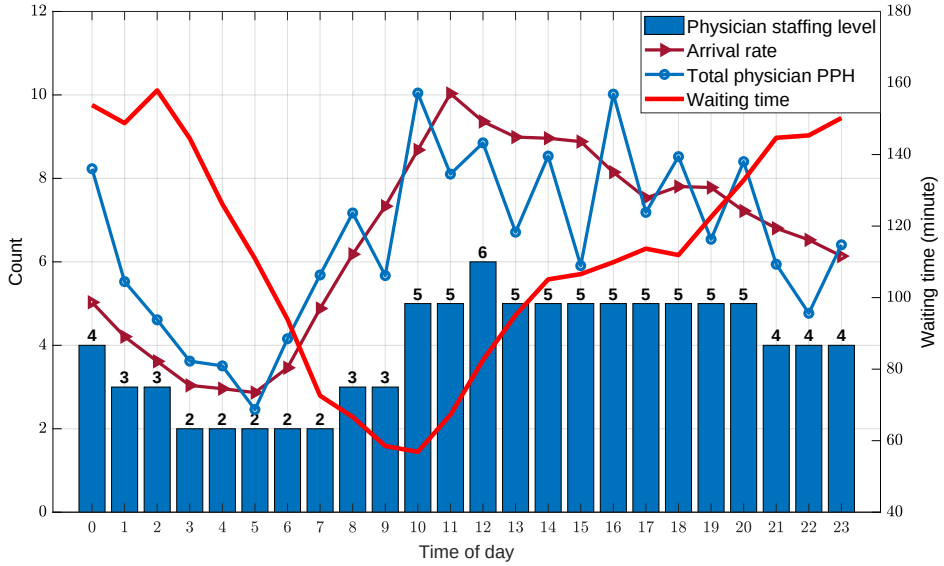
Key words: Emergency Department, Time-Varying Productivity, Simulation, Physician Staffing, Behavioral Queueing

1. Introduction

Emergency department (ED) overcrowding is a pressing issue for many countries around the world (Pines et al. 2011), impairing EDs' ability to provide timely care. Hence, the importance of modeling ED operations to reduce overcrowding cannot be overstated. ED is a complex network with time-varying demand and endogenous service rates, rendering it close to impossible to obtain any analytical results for system performance. Even numerical evaluation becomes a difficult task. For example, Campello et al. (2016) consider

EDs as *case-manager type systems*, where patients returning to physicians for reassessment is explicitly modeled. With proper Markovian assumptions, the system can be modeled by a continuous-time Markov chain. However, the system dimension grows exponentially and numerical evaluation of system performance becomes challenging, even for a small number of physicians under a stationary demand process.

Figure 1 Patient arrival rates, average waiting time, physician staffing level, and total new patients seen per hour (PPH) by all physicians on duty in the main treatment area (excluding fast-track area) of our study ED from January 3 to July 31, 2015. This period was chosen because physician staffing remained unchanged during this period.

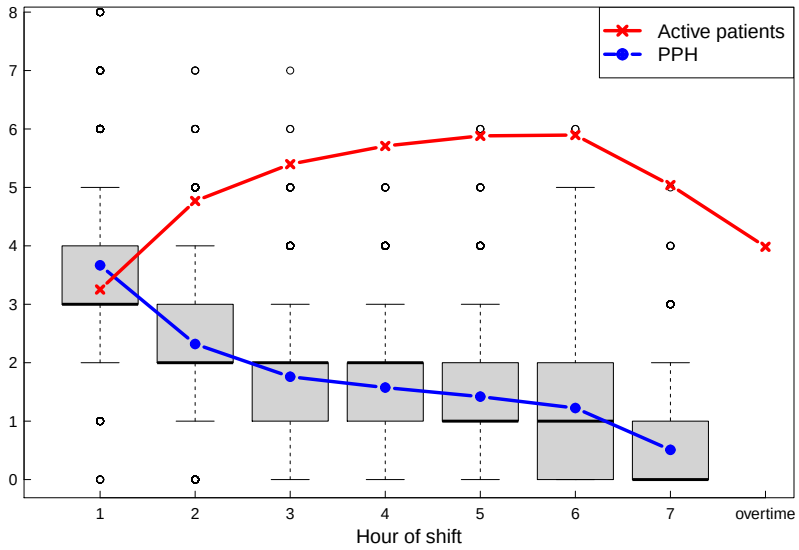


Despite the difficulty of evaluating time-dependent metrics (e.g., waiting times, throughput), they are crucial for system-level decision making, such as physician staffing. ED operations are naturally modeled as queueing systems, which requires a good understanding of the arrival and service processes. Using patient visit data from an urban tertiary hospital in Alberta, Canada, we calculate the average patient arrivals per hour (demand for emergency care), physician staffing levels (ED capacity), and average waiting times (from triage to first examination by a physician) by time of day, shown in Figure 1. An immediate observation is that the physician staffing level in our study ED is carefully designed so that the number of physicians on duty matches the time-varying demand. This is done by staggering shifts of different lengths; see a description of the shift patterns and lengths in Section 3.3. However, the outcome is less than satisfactory, as the average waiting time varies significantly over the course of the day and exceeds two hours at times.

A key determinant of the patient waiting time is physicians' speed in treating new patients, measured by PPH—the number of new patients seen by a physician per hour. PPH is often used as a measure of a physician's productivity, as in Joseph et al. (2018, 2020) and Zaerpour et al. (2021). We add the plot of the total physician PPH, i.e., the total number of new patients seen per hour by all physicians on duty, by time of

day to Figure 1. An intriguing observation is that the total physician PPH varies significantly, even when the staffing level remains constant. Take the 10:00 to 21:00 period as an example: there are 5 physicians on duty during this 11-hour period, except from 12:00 to 13:00. However, the total physician PPH varies from 5.9 to 10.0, a 69% difference. Interestingly, the highest total PPH level does not coincide with the peak staffing hour, which occurs between 12:00 to 13:00.[1] Motivated by this system-level behavioral anomaly and the wisdom from classical queueing theory that higher variations in service times lead to longer waiting times, we investigate this observation further by scrutinizing the PPH at the individual physician level.

Figure 2 The average number of new patients seen per hour (PPH) and the average number of active patients for all 7-hour shifts in the main ED area, estimated using data from January to July 2015. Active patients refer to all patients under a physician's care at any given time of the shift (see more details in Section 3.2). The extra data point for active patients outside the shift duration is due to physician work overtime for one hour at times.



Most shifts in our study ED have a length of 7 or 8 hours. Figure 2 shows the average PPH by shift hour of all 7-hour shifts in the non-fast-track area from our data (see the PPH plot for 8-hour shifts in Figure 9, Appendix A). Based on the time-varying structure of PPH observed from Figures 2 and 9, we divide a shift into three phases, within each phase PPH exhibits distinct patterns: the *start-of-shift* phase (the first two hours), the *end-of-shift* phase (the last hour), and the *middle-of-shift* phase (the remaining hours of the shift). We observe that PPH decreases exponentially during the start-of-shift phase—from 3.6 in the first hour to 1.94 in the third hour (a 46% drop) in Figure 2; then, it plateaus during the middle-of-shift phase; after

[1] One might conjecture that the variation in total physician PPH is a result of physician idling due to no patient waiting to be seen during certain time periods. However, our data show that there were always new patients waiting to be seen during the high-load period (10:00 to 21:00) in the study ED during our study period (January to July 2015).

which, it drops to near zero in the end-of-shift phase. The pattern becomes even more significant for 7- and 8-hour shifts using half an hour as the time resolution; see Figure 10 in Appendix A. When we further plot the PPH for each individual physician or for a specific type of shift, we observe a similar pattern. The PPH for fast-track shifts also shows a similar decreasing pattern, although the rates are higher and the magnitude of the decrease is lower than for non-fast-track shifts. Similar structures can be observed using data from U.S. hospitals (Joseph et al. 2018, 2020). Hence, we conclude that this time-varying pattern of physician PPH within a shift is highly robust.

The physician-level PPH determines the service speed of the ED, which has a crucial impact on system-level performance metrics, such as the average patient waiting time and queue length. To the best of our knowledge, this time-varying pattern of physician PPH has not been thoroughly investigated in the literature. Hence, we aim to shed light on its cause and impact. Specifically, we focus on three research questions: What is the mechanism behind time-varying physician PPH? How does time-varying PPH help us model complex ED operations? More importantly, how can we leverage our findings to support ED decision making such as physician staffing?

Our study makes the following contributions. First, we model a physician's decision problem within a finite shift by an optimal control framework. We obtain closed-form expressions for the time-varying PPH under the optimal policy, which explain the underlying drivers of the exponential decay during the start-of-shift phase and the sharp drop in the end-of-shift phase. We conclude that time-dependent physician behavior during a shift is a result of (i) the repetitive nature of emergency service, which leads to physician multitasking; and (ii) the discrete nature of shift-based scheduling, which induces *patient handoff*, i.e., transfer of patient care from one physician to another. We also provide an estimate of physician capacity and support the practice of setting a common cutoff for signing up new patients approaching the end of shift. Hence, this study advances our understanding of individual physicians' within-shift behavior and therefore contributes to the behavioral queueing literature.

Second, the insights into the time-varying pattern of PPH motivate our modeling of the complex ED system by an $M(t)/M^{\mathrm{PPH}}(t)/s(t)$ queue, i.e., a multiserver queue with nonstationary Poisson arrivals and exponential service times with time-varying rates (PPH). The model parameters can be estimated by simply calculating the average of the hourly arrivals and the PPH for each shift. Simulation results show that our model produces time-of-day-dependent performance metrics that closely match the data from two Canadian EDs. To the best of our knowledge, this is the first ED model that can accurately validate against real data using time-of-day-dependent metrics. Our results highlight the importance of explicitly considering time-varying physician service rates in the modeling of ED operations. Hence, this study contributes to the literature on ED modeling and simulation.

Third, EDs alternate between over-staffing and under-staffing during the course of a day, which renders most of the performance evaluation algorithms that depend on stationary approximations impracticable. The essence of our $M(t)/M^{\mathrm{PPH}}(t)/s(t)$ model is dimension reduction through state aggregation. As a result, it allows the transient analysis of system performance through uniformization of a continuous-time Markov chain (CTMC) with jumps at discrete time epochs, which can be integrated with off-the-shelf algorithms for physician staffing. Our case study using data from a Canadian ED shows that the new shift schedules generated using our method can improve the current schedule in the study ED and result in significant annual cost savings. Hence, this study contributes to ED practice and the physician staffing literature.

The rest of this paper is organized as follows. We discuss the relevant literature in Section 2 and introduce the study setting in Section 3. We study physicians' within-shift behavior in Section 4. Our results motivate a novel queueing model for ED operations in Section 5, and we show how it can support physician staffing decisions in Section 6. In Section 7, we conclude the paper, discuss the managerial insights, and point to future research directions. All proofs and additional results are given in the appendices.

2. Literature

Recent years have seen wide applications of operations research/management tools to improve healthcare access and reduce costs (see Saghafian et al. 2015 and Dai and Tayur 2020 for overviews). Our work aims to understand ED physicians' decision making underlying their time-varying productivity and thus is relevant to studies of healthcare workers' behavioral issues. Evidence shows that healthcare workers respond to system crowding and high workload by adjusting their behavior and capacity rationing decisions, such as service speedup (KC and Terwiesch 2009), patient undercoding (Powell et al. 2012), early patient discharge (Berry Jaeker and Tucker 2016), and early task initiation (Batt and Terwiesch 2016). Physicians may also adapt their patient prioritization behavior (Ding et al. 2019, Li et al. 2021) and admission decisions (Kim et al. 2015, 2020) to the level of system congestion. The aforementioned studies are mostly empirical in nature. In contrast, our study is based on an optimal control framework.

Our model explicitly captures the interactions between testing and reassessment during treatment of a patient by a physician and thus is relevant to studies that model the repetitive services provided to customers in service systems (see Table 1 in Ingolfsson et al. 2020 for a summary). Among which, our study is most relevant to works that use fluid models to study healthcare systems with patient returns. Yom-Tov and Mandelbaum (2014) propose an Erlang-R model to study the return-to-service phenomenon and find that Erlang-R is preferable to the classical Erlang-C model in a time-varying environment or over a finite horizon. They propose a square-root staffing policy based on the modified offered load. One important takeaway from Yom-Tov and Mandelbaum (2014) is that the repetitive nature of service should be considered to

achieve better staffing decisions, which aligns with our finding that it is crucial to account for time-varying productivity (partly driven by patient return) in ED modeling and physician staffing. Chan et al. (2014) use a fluid model to examine when to use speedup in service systems with state-dependent service times and return probabilities and identify scenarios where speedup is helpful or speedup should never be used. Ingolfsson et al. (2020) study a variant of the model in Chan et al. (2014) by assuming that the return to service occurs at the end of the delay rather than before the delay—which can be more realistic in certain settings—and find that their model and that in Chan et al. (2014) have identical equilibrium points but significant differences in transient behavior. We note that Duan et al. (2020) use a fluid model similar to ours to infer the initial assessment time. We use the optimal control framework to capture the trade-off between throughput and patient handoff to understand individual physicians' transient behavior during a shift. Hence, our work differs from the aforementioned studies in both the modeling framework and the study objectives.

The insights into physicians' time-varying PPH obtained from the optimal control framework motivate our novel $M(t)/M^{\mathrm{PPH}}(t)/s(t)$ model for ED operations. Hence, our work is also relevant to the literature on ED modeling and patient flow management; see, e.g., Dobson et al. (2013), Huang et al. (2015), Campello et al. (2016), Çağlayan et al. (2019). The dimension reduction of the complex ED network in our queueing model is via the aggregation of each physician node, which is similar to the $\mathcal{T}$ approximation in Campello et al. (2016). The differences between that study and ours are also significant. For example, the patient arrival process is modeled as a stationary Poisson process in Campello et al. (2016), whereas we consider a nonstationary Poisson process. We use the time-varying PPH as the service rate of each physician, whereas Campello et al. (2016) use the throughput rate of a single-server finite-source queue in steady state. Moreover, Campello et al. (2016) assume that customers are immediately "pushed" to a server upon arrival by a dispatcher unless the server has reached her maximum caseload. In contrast, physicians in our study ED "pull" a patient from the waiting room for treatment when they are available.

We apply the uniformization method to evaluate the performance of the $M(t)/M^{\mathrm{PPH}}(t)/s(t)$ model for physician staffing. Hence, our study is also relevant to the extensive literature on workforce management; see Green et al. (2007) for an overview. Hu et al. (2020) use an optimal control framework to study decisions on the allocation of resources for proactive care when considering patient condition deterioration. They obtain optimal scheduling policies when the system is (i) in a normal state of operation and (ii) under a random shock. Chan et al. (2021) study the dynamic assignment of nurses in EDs at the beginning of discrete shifts by a fluid control model. They obtain insights on the structure of "good" policies and use simulation to show that their heuristics on nurse reassignment can significantly reduce the system cost compared to without reassignment. Zaerpour et al. (2021) empirically examine the factors determining the time-varying physician productivity and then leverage this knowledge to assign physicians to predetermined shifts. Our work differs

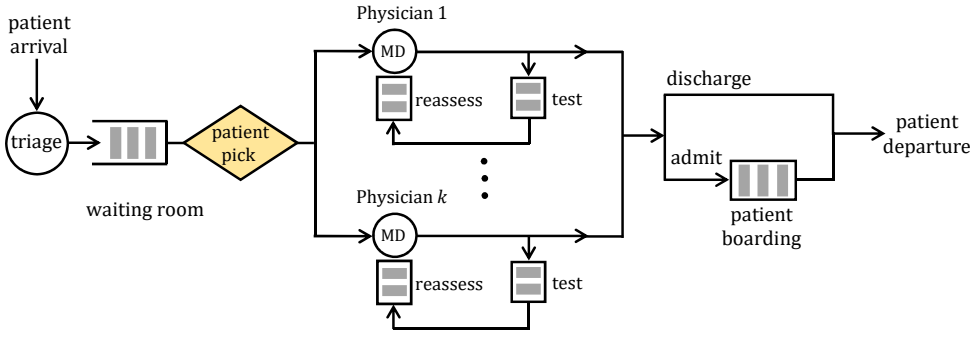
from these studies in that we focus on optimizing physician staffing levels by determining the shift start times to better match capacity with demand. Similar to our study, Liu and Xie (2021) use the uniformization method to optimize the shift start and end times in an ED setting. However, the time-varying service time is not considered in Liu and Xie (2021), in contrast to our work.

Finally, we note that both Chan (2018) and Batt et al. (2019) empirically investigate the impact of patient care handoff in EDs, which is another key driver of time-varying PPH, especially during the end-of-shift phase. Hence, their work is relevant to our study. The emergency medicine community has also observed time-varying pattern of physician productivity levels (Joseph et al. 2018, 2020). Joseph et al. (2018) find that estimating physician productivity as a simple average substantially misestimates physicians' capacity and suggest that the time-varying pattern should be factored into physician staffing. Joseph et al. (2020) find that a decrease in PPH does not reflects a decreasing workload. These studies differ from ours in both the study objectives and framework.

3. ED Operations and Patient Flow

In this section, we describe the patient flow process in the main area of our study ED. The fast-track area, a separate ED area with dedicated medical teams, is not the focus of this study. Note that our description is based on EDs in Alberta, Canada, and the operations in EDs of other regions may be different. However, we believe that the key features (such as patients return for service) are shared with most EDs. A depiction of the patient flow in the main ED area is provided in Figure 3.

Figure 3 A depiction of the patient flow process in the main area of an emergency department with k physicians.



3.1. Patient Flow

Upon arrival, patients are triaged into one of five levels with a lower level indicating higher urgency. After triage, patients wait in the waiting room. When a physician becomes available, she will choose a

patient for initial assessment[2] based on a given prioritization rule (Ding et al. 2019, Li et al. 2021). After initial assessment, some patients may leave the ED, while others may undergo diagnostic tests or medical procedures. (For simplicity, we hereafter use *tests* to represent all tasks performed by non-physician staff.) Those patients will join the queue for testing (see Figure 3) and return to the same physician for reassessment when the test results are ready. We refer to patients waiting to be seen in the waiting room as *new patients* and those waiting for reassessment as *return patients*. A patient may return to the same physician for service several times during his sojourn in the ED. A patient departs the ED if he is discharged; otherwise, the patient is admitted and becomes a boarding patient, waiting in an ED bed until being transferred into an inpatient bed.

3.2. Physician Multitasking and Capacity

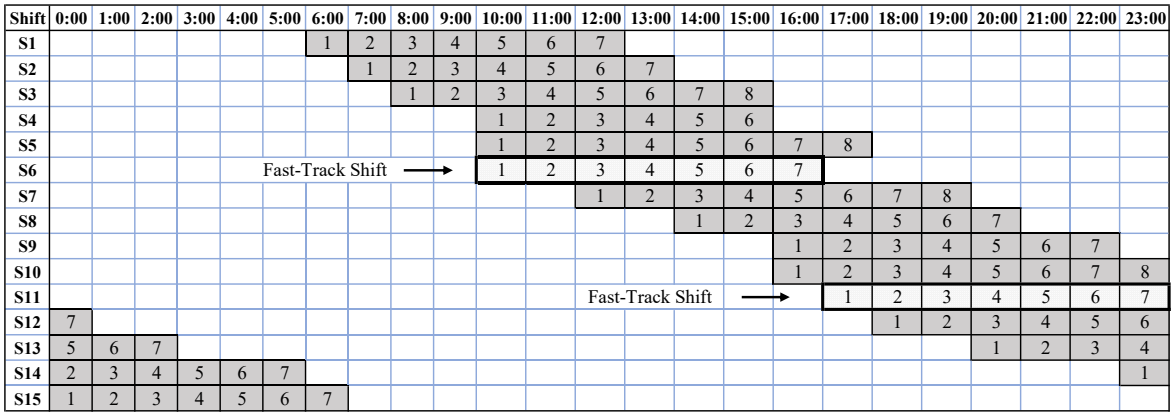
It is well known that ED physicians are multitasking (KC 2013, Song et al. 2018, Li et al. 2021); i.e., at any given time, a physician is responsible for the care of multiple patients, some of whom are undergoing testing in the test queue while others are waiting for reassessment (see Figure 3). The total patients under a physician's care at any given time are also referred to as the *active patients* of this physician (Joseph et al. 2020). See Figure 2 for an illustration of the average number of active patients by shift hour calculated using our data. Physicians do not need to discharge a patient before they start working on a new patient; however, a physician generally does not take on more patients than her *capacity*, i.e., the maximum number of patients that she can simultaneously care for (Campello et al. 2016). Although it seems unlikely that physicians keep a fixed number in mind as their capacity, Saghafian et al. (2012) observe from their study ED that an individual physician's capacity is generally no more than seven. KC (2013) find that the upper quartile of a physician's workload is five in their study hospital.

3.3. Shifts and Patient Care Handoff

EDs provide care 24 hours a day; however, no healthcare provider can work around the clock. As a result, shift-based scheduling is a necessity. Figure 4 shows the daily physician shifts from January to July 2015 in our study ED. During this period, there were 15 shifts (and hence 15 physicians) scheduled in the ED each day, two of which were fast-track shifts and the remainder were scheduled for the main area. The shift lengths in our study ED were 6, 7, or 8 hours. We observe that physicians started their shifts at staggered times during the day so as to better match physician capacity with time-varying patient demands. Moreover,

[2] Note that the mechanisms for routing patients to physicians could be different in other EDs. For example, Campello et al. (2016) describes an ED where a dispatcher assigns patients to physicians with available caseload after triage, whereas in Song et al. (2015) patients are routed to physicians by a round-robin policy, independent of physician speed or idle time.

Figure 4 The daily physician shifts in our study ED from January to July 2015.



Note. There are 15 shifts each day, with one 6-hour shift, 10 7-hour shifts (two out of the 10 are fast-track shifts), and four 8-hour shifts. The numbers in each row represent the shift hour of the corresponding shift.

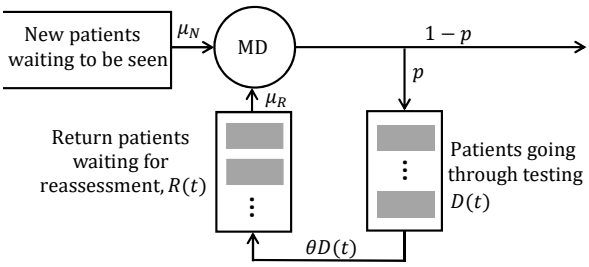
the staggered shifts avoided the undesirable situation of too many physicians leaving work at the same time and thus made the end-of-shift transition easier. We elaborate below.

When approaching the end of shift, a physician needs to transfer the care of unfinished patients to other physicians on duty. This practice is referred to as *patient handoff*, which is unsafe and undesirable because it causes discontinuity of care and creates opportunities for medical errors. Handoff has been linked to up to 24% of ED malpractice claims (Cheung et al. 2010), longer patient length of stay (Epstein et al. 2010), and higher 72-hour revisit rate (Batt et al. 2019). A recent study suggests that physicians should "slack off" approaching the end of their shift, i.e., stop signing up new patients, to avoid handoff and improve ED efficiency (Chan 2018). This aligns with the practice in the U.S. ED studied by Song et al. (2015), where new patients will not be assigned to physicians in the last two hours of their shifts. Physicians in another U.S. ED stated that "they are less likely to pick up new patients in the last hour or so of their shifts" (section 6.2.2 in Batt et al. 2019). Similarly, in our study ED in a Canadian hospital, physicians can choose *not* to see new patients in the last hour of their shifts, even they have to stay idle. It should not be interpreted literally when we say physicians slack off or stay idle. Physicians may perform non-clinical duties such as student mentoring, administration, among others.

4. Physician Within-Shift Behavior Behind Time-Varying Productivity

In this section, we model the patient treatment process of any individual physician using the optimal control framework. We obtain closed-form expressions for a physician's productivity, throughput, and capacity under the optimal policy, which help explain the time-dependent behavior of physicians. Understanding individual-level behavior provides fundamental insights into system-level performance, which motivates our modeling of the complex ED system in Section 5.

Figure 5 A reentrant queue to describe the patient treatment process by a single physician during a shift. MD = medical doctor.



4.1. Model Description

We consider a fluid model with returns to describe the patient treatment process by a single physician during her shift $[0,T]$, where $T>0$ denotes the shift length. A schematic depiction of the patient flow is shown in Figure 5. We assume that there are always new patients waiting to be seen in the waiting room. Our data analysis shows that this assumption holds for most of the time in our study period. The rate of serving new patients (i.e., initial assessment) is denoted by $\mu_N>0$. With probability p, a patient needs to undergo testing after assessment. Otherwise, the treatment is completed and the patient leaves the ED. We assume that the test queue has infinitely many servers and the mean testing time is $1/\theta>0$. This infinite-server assumption aligns with Yom-Tov and Mandelbaum (2014) and Campello et al. (2016). When the test results are ready, the patient returns to the same physician for reassessment unless they have been handed over to another physicians (see Section 3.3). Let μ_R denote the rate at which return patients are served. Reassessment generally takes less time than initial assessment, i.e., $\mu_R \geq \mu_N$. After reassessment, the patient may need another test with the same probability p, independent of the number of tests that have already been performed for this patient, which implies that the total number of tests that a patient undergoes upon leaving the ED follows a geometric distribution with success probability $1-p$. This assumption has been adopted in the literature; see, e.g., Yom-Tov and Mandelbaum (2014), Campello et al. (2016) and Li et al. (2021). The service and reassessment times, testing times, and return probability are assumed to be independent of each other and of the lengths of the test and reassessment queues.

Assume that a unit reward is earned when a patient's treatment at the ED is completed. At the end of shift, a physician's patients whose care is incomplete become handoff patients (see Section 3.3). Let $h(x)$ denote the cost function when there are x handoff patients, $x\geq 0$. The handoff cost represents (i) the health risk due to the information loss during handoff communication, and (ii) the time and effort spent on handoff communication to transfer essential information from one physician to another. Let $D(t)$ and $R(t)$ denote the number of patients in the test and reassess queues at time t, respectively. Then, the handoff cost at the end of a shift is $h(D(T)+R(T))$. Assume that return patients are prioritized over new patients.[3] We further assume

[3] In our study ED, when a physician finishes an ongoing task, she logs into the ED information system through a terminal. The

that physicians do not idle when there are patients waiting for reassessment, which aligns with the practice in our study ED. However, physicians can choose not to see new patients, even they have to stay idle. Let $\alpha_N(t)$ and $\alpha_R(t)$ denote the percentage of time that the physician spends on processing new and return patients at time t, respectively. The physician's objective is to maximize the total net reward by controlling $\alpha_N(t)$ and $\alpha_R(t)$, $t \in [0,T]$. This problem can be formulated within the optimal control framework as follows:

$$\begin{aligned} \max_{\alpha_N(t),\alpha_R(t)} & \left\{ \int_0^T (1-p)\left[\alpha_N(t)\mu_N + \alpha_R(t)\mu_R\right] \mathrm{d}t - h(D(T)+R(T)) \right\} \\ \text{s.t.} \quad & D'(t) = p\left[\alpha_N(t)\mu_N + \alpha_R(t)\mu_R\right] - \theta D(t),\ D(t) \geq 0,\ D(0) = 0, \\ & R'(t) = \theta D(t) - \alpha_R(t)\mu_R,\ R(t) \geq 0,\ R(0) = 0, \\ & 0 \leq \alpha_N(t) + \alpha_R(t) \leq 1,\ \alpha_R(t) = \min\left\{1, \theta D(t)/\mu_R\right\},\ \alpha_N(t) \geq 0. \end{aligned} \tag{1}$$

The constraints on $D'(t)$ and $R'(t)$ respectively describe the dynamics of the test and reassessment queues; $\alpha_N(t) + \alpha_R(t) \leq 1$ implies that the total percentage of time spent on initial assessment and reassessment should not exceed 100% at any time; $\alpha_R(t) = \min\{1, \theta D(t)/\mu_R\}$ captures that reassessment is prioritized over initial assessment and that physicians do not idle as long as there are patients waiting for reassessment. The initial condition $D(0) = R(0) = 0$ implies that a physician who just began her shift has no patient undergoing testing or waiting for reassessment. However, as it should be clear in the proof of the optimal policy (see Appendix B), the initial state conditions do not change the structure of the optimal policy.

4.2. The Optimal Policy

Next, we solve the optimal control problem in (1) by applying Pontryagin's maximum principle. Theorem 1 provides closed-form expressions for the optimal controls, denoted by $\alpha_N^*(t)$ and $\alpha_R^*(t)$, respectively.

THEOREM 1. *Assume that $h(\cdot)$ is an increasing differentiable function. Then, we have the following for the optimal control problem defined in (1):*

(i) $R(t) = 0$, $\theta D(t)/\mu_R \leq 1$, and $\alpha_R^*(t) = \theta D(t)/\mu_R$ for all $t \in [0,T]$.

(ii) The optimal control $\alpha_N^*(t)$ is of threshold type. More specifically, there exists an optimal switching time $t^* \in [0,T]$ such that $\alpha_N^*(t) = 1 - \theta D(t)/\mu_R$ if $t \in [0,t^*]$; $\alpha_N^*(t) = 0$ if $t \in (t^*,T]$, where

$$t^* = \min\left(T, \max\left(0, T - \frac{\ln[p(1+h'(D(T)))]}{(1-p)\theta}\right)\right). \tag{2}$$

upper half of the screen shows the reassessment requests from her active patients and the lower half shows the new patients waiting to be seen. The upper half is visible to this physician only, whereas the information on the lower half is available to all physicians. In general, a physician processes all of the reassessment requests before signing up a new patient so as to limit patients' length of stay. Physicians may also follow the shortest processing time rule because reassessment is generally faster than treating a new patient. We note that Huang et al. (2015) prove that it is optimal to prioritize return patients over new patients subject to adhering to their deadline constraints in ED settings under heavy traffic.

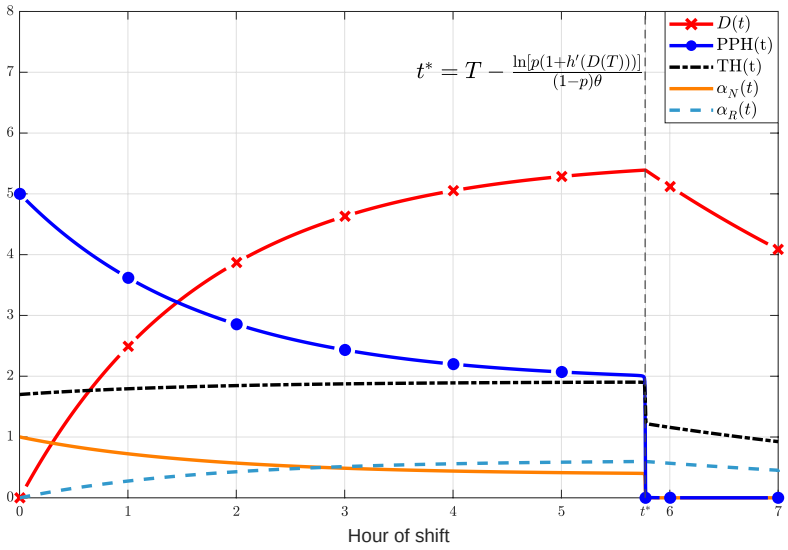
(iii) Furthermore, under the optimal policy, we have

$$D(t) = \frac{p\mu_N}{\theta\left(1-p+p\mu_N/\mu_R\right)}\left(1-e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t}\right),\ t\in[0,t^*],\ \textit{and} \tag{3}$$

$$D(t) = D(t^*)e^{-(1-p)\theta(t-t^*)},\ t\in(t^*,T]. \tag{4}$$

Theorem 1 completely characterizes the optimal policy for Problem (1). Under the optimal policy, there exists an optimal switching time t^* such that (i) when the shift hour is before t^*, the physician is always busy serving patients ($\alpha_R^*(t)+\alpha_N^*(t)=1$), and priority is given to return patients over new patients; (ii) when the shift hour exceeds t^*, it is optimal for the physician to stop signing up new patients and focus on serving return patients—even if the physician has to stay idle—so as to reduce the number of handoff patients. A numerical illustration of the optimal policy is shown in Figure 6. While the structure of the optimal policy is intuitive, the closed-form results provide rich insights into the behavior of individual physicians and the management of ED operations, as discussed in the rest of this section and the next section.

Figure 6 A numerical illustration of $\alpha_N^*(t)$, $\alpha_R^*(t)$, $D(t)$, TH(t), and PPH(t) under the optimal policy for Problem (1) when $\mu_N=5$, $\mu_R=6$, $\theta=0.6$, $p=0.66$, $h(x)=x$, and $T=7$ (i.e., 7-hour shifts). The optimal switching time $t^*=5.8$.



The optimal switching time t^* is a function of the test probability p, the mean testing time $1/\theta$, and the derivative of the handoff cost function $h'(\cdot)$. In fact, the proof of Theorem 1 does not require $h(\cdot)$ to be increasing. However, if $h'(\cdot)\le 0$, it is easy to see that $t^*=T$. In other words, it is optimal to serve new patients at any time in the shift if more handoffs lead to lower costs, which is trivial but unrealistic and less interesting. Note that when $h'(\cdot)$ is a constant, i.e., the handoff cost depends linearly on the number of handoff patients, t^* does not depend on μ_N and μ_R—measures of physicians' speed in treating patients.

This insight provides a justification for setting a *common switching time* for all physicians—despite being aware of the heterogeneity in physician speeds—when each handoff patient is perceived to contribute the same cost. The common switching time has been observed in practice; for example, physicians can choose *not* to see new patients in the last hour of their shifts in our study ED, and new patients will not be assigned to physicians in the last two hours of their shifts in the California ED studied by Song et al. (2015).

4.3. Time-Varying PPH, Physician Capacity, and Throughput

Let $\mathrm{PPH}(t)$ denote a physician's productivity rate, i.e., the rate of seeing new patients, and let $\mathrm{TH}(t)$ denote the throughput rate, i.e., the rate at which a patient's ED treatment is completed, at time t in a shift, where $t \in [0,T]$. Then, we have $\mathrm{PPH}(t) = \alpha_N(t)\mu_N$ and $\mathrm{TH}(t) = (1-p)[\alpha_N(t)\mu_N + \alpha_R(t)\mu_R]$, which gives the following result immediately.

Corollary 1. *Under the optimal policy for Problem (1), we have*

$$PPH(t) = \mu_N - \frac{p\mu_N^2}{p\mu_N + (1-p)\mu_R}\left(1 - e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t}\right),\ t \in [0,t^*], \tag{5}$$

$$PPH(t) = 0,\ t \in (t^*,T], \tag{6}$$

$$TH(t) = (1-p)\mu_N + \frac{p(1-p)\mu_N(\mu_R-\mu_N)}{p\mu_N + (1-p)\mu_R}\left(1 - e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t}\right),\ t \in [0,t^*],\ \textit{and} \tag{7}$$

$$TH(t) = (1-p)\theta D(t^*)e^{-(1-p)\theta(t-t^*)},\ t \in (t^*,T], \tag{8}$$

where t^* and $D(t^*)$ are given in (2) and (3), respectively.

Theorem 1 and Corollary 1 not only prove the structure of the optimal policy, but also provide closed-form expressions for $\alpha_N(t)$, $\alpha_R(t)$, $D(t)$, $\mathrm{PPH}(t)$, and $\mathrm{TH}(t)$ under the optimal policy. The analytical results allow us to obtain several interesting insights into a physician's within-shift behavior, as elaborated below.

4.3.1. Time-Varying Physician Productivity. The expressions of $\mathrm{PPH}(t)$ in (5) and (6) shows that a physician's productivity within a shift is time-varying. A numerical illustration of $\mathrm{PPH}(t)$ is shown in Figure 6. The fact that $\mathrm{PPH}(t)$ is an exponential function of the shift hour with a negative exponent explains the exponential decay of a physician's productivity during the start-of-shift phase observed from data (see Figure 2). Our model and results suggest that the dramatic reduction in physician productivity is mainly due to multitasking, i.e., physicians need to spend time processing the reassessment requests from returning patients and thus have less time to treat new patients. The exponential term in $\mathrm{PPH}(t)$ diminishes as t increases. Correspondingly, the productivity rate plateaus during the middle-of-shift phase; see Figures 2 and 6. During the end-of-shift phase, $\mathrm{PPH}(t)$ drops to zero as a result of the physician's decision not to sign up new patients so as to reduce patient handoff. This is also suggested by Chan (2018) and Batt et al. (2019)

and has been shown to be optimal under our model setting. Note that the PPH observed from data in the last shift hour is small but above 0 (see, e.g., Figure 2), because physicians occasionally sign up new patients in the last hour of their shifts. Physicians usually work overtime in these shifts.

4.3.2. Physician Capacity. By Theorem 1 (i), the length of the reassessment queue $R(t) = 0$, $\forall t \in [0, T]$. Hence, $D(t)$ becomes the total number of active patients of the physician, i.e., the physician's workload at t. This explains the similarity in shape between $D(t)$ in Figure 6 and the average number of active patients during a physician's shift observed from data in Figure 2. Note that $\theta D(t)/\mu_R \leq 1$ implies $D(t) \leq \mu_R/\theta$, suggesting that a physician's maximum workload, or her service capacity, is the maximum number of patients that she can take care of simultaneously so that the reassessment requests from her active patients do not take up all of her time. From our data, the mean testing time $(1/\theta)$ is 90.2 minutes and the mean reassessment time $(1/\mu_R)$ is 14.8 minutes,[4] which yields $\mu_R/\theta \approx 6.1$. Surprisingly, this estimated physician capacity matches closely with observations from the data; see Figures 2, 9, and 10. It also aligns with descriptions in the literature. For example, a physician's workload is generally no more than 7, as observed by Saghafian et al. (2012), while the upper quartile of a physician's workload is 5 in the study hospital of KC (2013). Both examples suggest that μ_R/θ is a fairly accurate estimate of a physician's capacity.

4.3.3. Physician Throughput. From the expressions of $\mathrm{PPH}(t)$ and $\mathrm{TH}(t)$ in Corollary 1 and the numerical example in Figure 6, we observe that $\mathrm{PPH}(t)$ differs significantly from $\mathrm{TH}(t)$, especially during the start-of-shift and end-of-shift phases. During the middle-of-shift phase, both $\mathrm{PPH}(t)$ and $\mathrm{TH}(t)$ plateau and converge to a constant. Let $\mathrm{PPH}_\infty \triangleq \lim_{t\to\infty} \mathrm{PPH}(t)$ and $\mathrm{TH}_\infty \triangleq \lim_{t\to\infty} \mathrm{TH}(t)$. Then, we have

$$\mathrm{PPH}_\infty = \frac{(1-p)\mu_N\mu_R}{p\mu_N + (1-p)\mu_R} = \frac{1}{\tau_N + n_R\tau_R} = \mathrm{TH}_\infty,$$

where $\tau_i \triangleq \mu_i^{-1}$, $i \in \{N, R\}$, and $n_R \triangleq p/(1-p)$ is the expected number of reassessments, i.e., the mean of a geometric distribution with success probability $1-p$. Hence, during the middle-of-shift phase, the productivity and throughput rates of a physician converge to the same limit. Interestingly, the limit is the reciprocal of the expected time that the physician spends on a patient and does not depend on the testing rate θ. This may be due to the infinite server assumption at the test queue. We further notice that when the service rates for new and returning patients are the same, i.e., $\mu_N = \mu_R$, then the throughput rate becomes a constant at any time t prior to the optimal switching time t^*, i.e., $\mathrm{TH}(t) = (1-p)\mu_N$, $t \in [0, t^*]$. However, $\mathrm{PPH}(t)$ remains time-dependent. Hence, the throughput rate may not serve as a substitute for the productivity rate, even though they are similar during the middle-of-shift phase.

[4] In our data, the start time of a reassessment is available but not its end time. We use the start time of the next activity of the same physician to approximate the end time of the reassessment, which may overestimate the reassessment time.

4.4. Connection to Load-Dependent Service Times

Our findings suggest that the service rates at which physicians treat new patients decrease with shift hours, which connects our study to the literature on service speedup or slowdown; see, e.g., KC and Terwiesch (2009), KC (2013), Batt and Terwiesch (2016), Berry Jaeker and Tucker (2016), and Deo and Jain (2019). Interested readers are referred to Delasay et al. (2019) for a review through unified frameworks. We identify the phenomenon of physician slowdown in treating new patients during a shift. Our results suggest that physician multitasking (i.e., treating both new and return patients) and "slacking off" are the main drivers of time-varying productivity. However, there may be other mechanisms at work. For example, fatigue may reduce a physician's speed, resulting in lower values of μ_N and μ_R; higher congestion level in the waiting room may increase a physician's service speed; the queue configuration and information disclosure at EDs may lead to physician speedup due to increased ownership (Song et al. 2015) or social pressure (Song et al. 2018). Hence, the patterns in Figures 2 and 9 may be an aggregation of several lower-level mechanisms. Such mechanisms deviate from the focus of this study. We note that Zaerpour et al. (2021) conclude that shift hours and physician heterogeneity are the most important driving factors when modeling the physician PPH based on regression results.

5. Model of ED Operations

In this section, we consider the problem of modeling ED operations. A distinguishing feature of emergency care is that a patient may return to the same physician multiple times for service during his sojourn in the ED (see Figure 3). With proper Markovian assumptions, the system dynamics can be represented by a Markov chain, where the system state is a vector that includes the number of patients waiting to be seen in the waiting room, and the number of patients going through tests and waiting for reassessment for each physician. Unfortunately, the state space grows exponentially with the number of physicians on duty. Even with five physicians (a common number in our study ED; see Figure 1), the dimension of the state space can easily exceed 20 million (see a similar discussion in Campello et al. 2016), which makes the model analysis both theoretically and computationally challenging. Hence, we seek dimension reduction techniques to simplify the problem. Our aim is to identify a model that can balance between details and tractability, model parameters are easy to estimate, and system performances are able to match with real data.

5.1. A Queueing Model with Time-Varying Service Rates

A main takeaway from Corollary 1 is that a physician's productivity rate is time-varying and decreases significantly over the course of a shift. Furthermore, a descriptive analysis based on our data shows that approximately 48% of new patients were seen during the start-of-shift phase (2 hours), 49% were seen during

the middle-of-shift phase (4 hours for 7-hour shifts and 5 hours for 8-hour shifts), and only 3% were seen during the end-of-shift phase (the last hour). Hence, it is important to account for time-varying physician productivity in the modeling of ED operations.

Motivated by the insights above, we model the ED operations as an $M(t)/M^{\text{PPH}}(t)/s(t)$ queue, i.e., a time-varying queueing system with heterogeneous servers and shift-hour-dependent service rates, where t is the time in hours. The first $M(t)$ represents a nonstationary Poisson arrival process with time-dependent rate $\{\lambda(t), t \geq 0\}$, which has been shown to be a reasonable assumption (Kim and Whitt 2014). The number of servers (physicians) is time-varying, denoted by $\{s(t), t \geq 0\}$, where $s(t)$ is a nonnegative integer. The $M^{\text{PPH}}(t)$ represents exponentially distributed service times with time-varying rates, which can be estimated by the PPH of each of the $s(t)$ physicians on duty at t. We assume that the arrival rate is periodic with a daily cycle, so is the physician scheduling. Hence, $\lambda(t) = \lambda(t+24), s(t) = s(t+24), \forall t \geq 0$. We chose a daily cycle for ease of presentation. Moreover, physician shift schedules often repeat on each day during a planning period in practice, which is the case in our study ED. However, our model can be extended in a straightforward manner to model schedules with different cyclic patterns, such as weekly cycles.

Let $\mathcal{S} = \{S_1, S_2, \cdots, S_k\}$ denote the physician shift schedule in an ED with k shifts scheduled to commence each day, where S_i represents the ith shift. Due to the shift-based scheduling, the number of physicians on duty is time-varying (see, e.g., Figure 1). We assume that an exhaustive discipline is applied whenever the number of physicians decreases, i.e., an outgoing physician will complete the service in progress before leaving (Ingolfsson et al. 2007). This is consistent with the practice in our study ED.

We assume that patients are served in a first-come-first-served (FCFS) manner, despite being aware that the patient prioritization process is highly complex and dependent on patients' triage levels, waiting times, and even ED resource availability (Ding et al. 2019, Li et al. 2021). However, we expect that the queueing discipline has a stronger impact on metrics beyond first-moment information, such as the waiting-time-based service levels (Green et al. 2007, Ingolfsson et al. 2007), but has little impact on the average patient waiting time or queue length, especially given that the composition of patients at each triage level does not vary significantly over the course of the day; see Figure 11 in Appendix A.

Finally, there may be more than one physician available to serve an arriving patient. Because physicians may be in different phases of their shifts and thus have heterogeneous service rates, we need to specify which physician to serve the patient. We choose to route the patient to the physician who most recently started her shift, which usually is the physician with the highest service rate at the moment. However, one would reasonably expect that this assumption does *not* make much difference compared to routing the patient to an available physician randomly, because EDs usually are critically loaded in a daily cycle, i.e., the daily arrivals—excluding patients who left without being seen (LWBS)—are approximately equal to the daily

total physician PPH. As a result, the chance that more than one physician is idling simultaneously is small. Our simulation results confirm this conjecture.

5.2. Model Accuracy: Validation Using Data from Two EDs

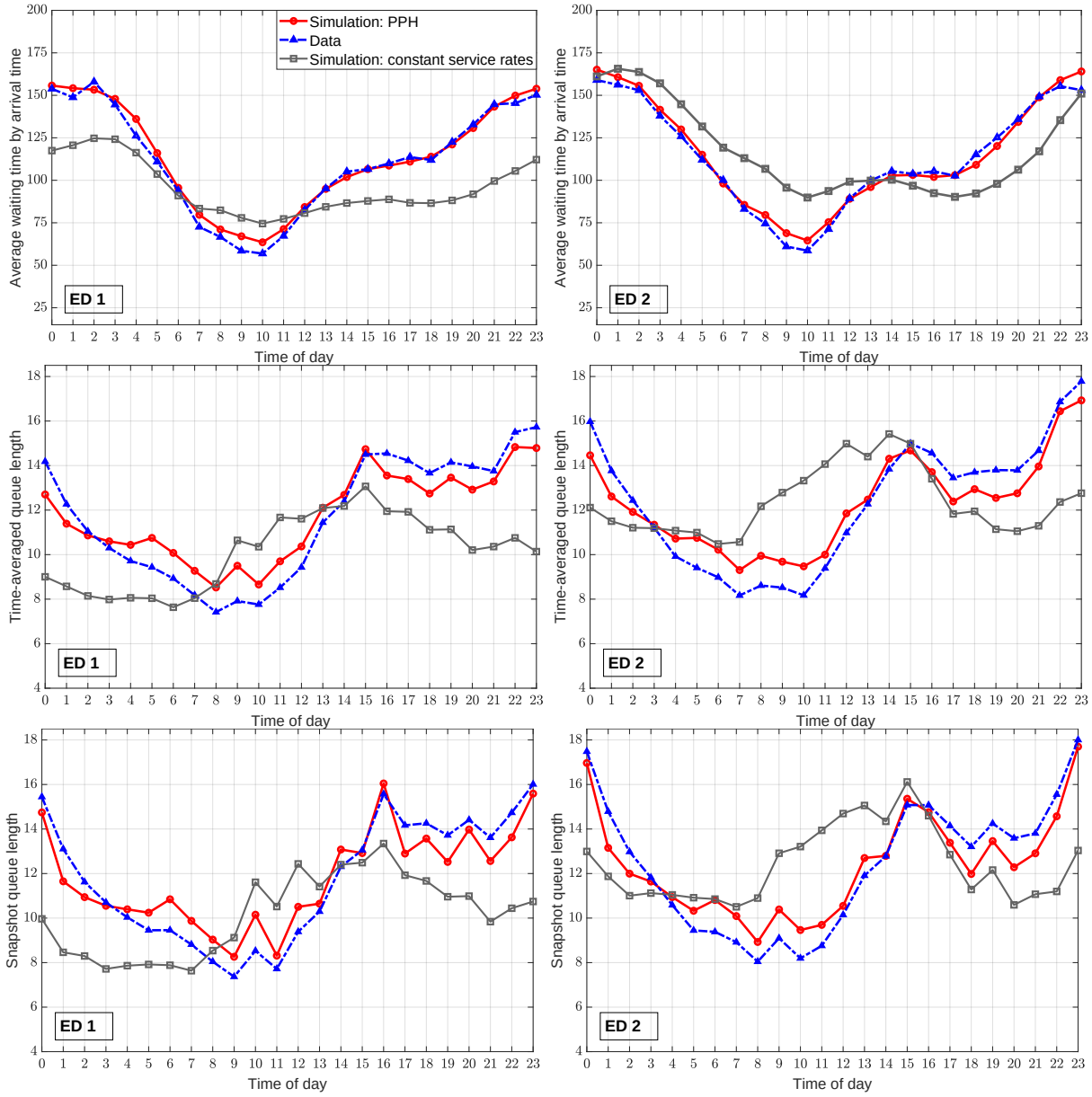
Next, we simulate the $M(t)/M^{\mathrm{PPH}}(t)/s(t)$ queue with parameters estimated using the data from our study ED (referred to as *ED 1*) between January and July 2015. We then compare the simulated time-of-day-dependent average waiting times and queue lengths with the data. To further demonstrate the generality and robustness of our results, we repeat the study using data from another ED (referred to as *ED 2*) in Alberta, Canada during a different study period. We provide the comparison results but not the details of the second dataset to avoid repetition.

The shift schedule at ED 1 from January to July 2015 is shown in Figure 4, including the start and end times (and thus the shift length) of each shift. We focus on the 13 shifts in the main ED area. The estimations of the arrival rates and the PPH for each shift are based on hourly resolution. The simulation model is built using SimPy 4.0.[5] The inter-arrival times of the nonstationary Poisson process are generated by the thinning algorithm. In the simulation, a physician immediately starts to serve patients at the shift start time. The service times are exponentially distributed with rates given by the PPH of the corresponding shift hour. An exhaustive discipline is applied at the shift end time. We focus on waiting room dynamics and choose three time-of-day-dependent performance metrics: (i) the average waiting time of patients who arrived in the same hour of day; (ii) the time-averaged number of patients in the waiting room (referred to as *time-averaged queue length*); and (iii) the average number of patients in the waiting room observed at the end of each hour (referred to as *snapshot queue length*).

We ran the simulation for 5 replications, each with 500 weeks, and we identify the first 200 weeks as the warm-up period. The results of comparing the simulated performance measures with the data are shown in Figure 7. We observe that the average waiting times from the simulation nicely match those from the data from both Canadian EDs—both in terms of the patterns and the magnitudes. An accurate evaluation of the average waiting time is essential for physician staffing planning, because it directly measures the timeliness of emergency care and is an important performance indicator in the Canadian Triage and Acuity Scale protocol—the triage algorithm in our study ED. The time-averaged and snapshot queue lengths also match the data reasonably well for both EDs; see the plots in the second and third rows of Figure 7. Furthermore, the aggregated average waiting time of all patients from ED 1 (ED 2) is 106.4 (109.5) minutes, whereas the simulated counterpart is 108.3 (110.9) minutes, which further shows the accuracy of our model.

[5] SimPy is a process-based discrete-event simulation framework based on Python. See `https://simpy.readthedocs.io/`.

Figure 7 The average waiting time, time-averaged queue length, and snapshot queue length from simulation with PPH (red line with circles), simulation with constant service rates (gray line with squares), and data (blue dashed line with triangles) by time of day. The plots on the left use data from our primary study hospital (ED 1), and those on the right use data from another Canadian hospital (ED 2).



We conducted an extensive search of the ED simulation literature (over 100 publications since the 1970s) and went over the references in a recent review paper on ED simulation (Vanbrabant et al. 2019), and find that our $M(t)/M^{\mathrm{PPH}}(t)/s(t)$ model is the first to produce time-of-day-dependent performance metrics that closely match real data, despite of the simple procedure to parameterize the model. The validation using data from a second ED demonstrates the feasibility of extending our model and results to other hospitals.

5.3. Constant Service Rates

Most previous works on ED modeling and physician staffing explicitly or inexplicitly assume a single-stage physician service with a constant service rate (see, e.g., Savage et al. 2015, Liu and Xie 2018). To examine whether this assumption is appropriate, we re-ran the simulation model with the same parameter setting, except that the physician service rate is a constant, calculated using the total number of new patients seen divided by the total shift hours. Hence, the service rate of all physicians on duty at any time is determined by the staffing level alone. The simulated average waiting times and queue lengths under constant service rates deviate significantly from the data; see Figure 7 (the gray line with squares). Interestingly, the variation in the simulated average waiting times between different hours of day is smaller than in the data. In other words, the simulated average waiting time curve under constant service rates is smoother. A plausible explanation is that the current physician shift schedules in both EDs were carefully designed to match the staffing level with patient demand under the assumption of constant service rates, so that the waiting times do not vary significantly over the course of the day. However, the outcome is less than satisfactory, potentially due to that the time-varying physician service rates were not considered by the scheduler.

To summarize, our results show that individual physicians' behavior is crucial to the modeling of system behavior. In particular, it is important to account for the shift-hour-dependent service rate (i.e., PPH) when modeling ED operations. Ignoring it is likely to fail to accurately capture the dynamics of patient flow. Our insight from Section 4.3.3 implies that the throughput rate from a physician node cannot serve as a substitute for PPH, which differentiates our model from the $\mathscr{T}$ approximation in Campello et al. (2016) that uses the throughput rate of a single-server finite-source queue in steady state as the service rate of a physician node. Furthermore, the $\mathscr{B}$ approximation in Campello et al. (2016) may be not applicable to our setting because the caseload of a physician who is just starting her shift can be quite different from that of a physician approaching the end of a shift, whereas the $\mathscr{B}$ approximation requires that the difference in caseload between physicians is at most 1. Both approximations work well in Campello et al. (2016), possibly because the case-manager setting is different from ours in terms of the arrival and patient-to-physician assignment processes; for more details, see Section 2.

Finally, we comment on the parameter estimation of the $M(t)/M^{\mathrm{PPH}}(t)/s(t)$ model. In principle, one simply needs to count the number of arrivals per hour (we use triage time as the arrival time) and the number of initial assessments done during each hour of a shift by the physician assigned to this shift, i.e., the PPH. However, one needs to be careful when dealing with real data. For example, our data cleaning identified issues including physician shift switching, system downtime due to maintenance, and physician no-shows, all of which create noise in the estimation. In addition, 1.65% of patients cannot be matched with a particular shift in the data from ED 2. As a result, the total daily PPH is, on average, slightly lower than the total daily arrivals. Hence, we proportionally adjust the arrival rates downward by multiplying by 98.35%.

6. Application to ED Physician Staffing

Physician staffing is a key decision affecting ED resource planning. In this section, we demonstrate that our model also enables numerical performance evaluation of the ED due to the reduced system dimensionality. The evaluation algorithm serves as a subroutine for optimizing physician staffing in our case study.

6.1. Improving Physician Staffing: A Case Study

In our study hospital, a scheduler first determines the start and end times of each shift every six months (more or less); then, physicians are allocated to each shift following required scheduling rules. Figure 4 shows the 15 physician shifts from January to July 2015 in our study ED. Among these, S6 and S11 are fast-track shifts, and all others are dedicated to serving patients in the main area. Next, we use the shift schedule in the main ED area (referred to as the *baseline schedule* hereafter) to demonstrate how our model can help improve physician staffing decisions. Specifically, we evaluate the effectiveness of optimizing the shift start times to reduce the average ED waiting time.

We adjust the start times of the 13 shifts to better match ED capacity with patient demand. The shift lengths remain the same as in Figure 4. The assignment of physicians to shifts is a second-stage problem, which is not the focus of this study. Hence, we assume that the assignment is the same as in the data. Interested readers are referred to Brunner and Edenharter (2011), Liu and Xie (2018), and Zaerpour et al. (2021) for the physician-to-shift assignment problem. In theory, the start time of each shift can be any time during the day. However, for practical relevance, we assume that physician shifts can only start at one of the 24 hours $\{0, 1, \cdots, 23\}$. Note that the adjustments in shift start times also affect the corresponding physicians' work schedule, which may violate certain scheduling rules and make the physician-to-shift assignment infeasible. Hence, we add constraints so that the baseline schedule will not be changed dramatically. In particular, we consider three scenarios and solve the corresponding staffing optimization problem under each scenario.

Scenario 1: The physician shifts must satisfy the following constraints: (i) the two night shifts, S14 and S15, remain unchanged because night shifts often complicate physician-shift assignment; (ii) the start times of the other 11 shifts can be adjusted to be earlier or later than the baseline schedule by at most two hours; (iii) the start times of the other 11 shifts cannot be later than 20:00 or earlier than 6:00; (iv) there must be at least two physicians on duty at any time of day.

Scenario 2: The same as in Scenario 1, except that constraint (iv) is relaxed; more specifically, we require the staffing level to be at least one physician on duty at any time of day.

Scenario 3: The same as in Scenario 2, except that we relax constraints (i) and (ii) so that all 13 shifts can be adjusted to be at most three hours earlier (or later) than the start times in the baseline schedule.

Our objective is to minimize the average ED waiting time under each scenario, because reducing waiting times achieves better health outcomes for patients (Guttmann et al. 2011) and cost reduction for hospitals (Woodworth and Holmes 2020). One may apply simulation optimization techniques to solve the staffing problem as there is no closed-form expression for the objective function. Indeed, we have shown that our novel simulation model can accurately capture ED waiting times. However, our attempts revealed that a commercial solver takes days to solve the optimization due to the large solution space. Hence, we propose a method that combines a local search algorithm (i.e., tabu search, see Liu and Xie 2021) with the uniformization method (discussed below) for the evaluation of each candidate schedule. Numerical experiments show that our method takes less than two hours for each scenario.

6.2. Performance Evaluation Through Uniformization

The operating regime of our study ED alternates between over-staffing and under-staffing over the course of a day under the current staffing plan, where *over-staffing* means that the hourly arrival rate exceeds the total physician PPH (e.g., 11:00–15:00 in Figure 1) and *under-staffing* refers to otherwise (e.g., midnight to 4:00 in Figure 1), which renders performance evaluation algorithms that depend on stationary approximations impracticable. In this section, we model the $M(t)/M^{\mathrm{PPH}}(t)/s(t)$ queue by a CTMC with state jumps at discrete time epochs and apply the uniformization method (which is also referred to as *randomization* in the literature) for the performance evaluation. The uniformization method achieves similar accuracy as solving the Kolmogorov forward equations while it only takes half the computational time, as demonstrated by Ingolfsson et al. (2007) in the setting of $M(t)/M/s(t)$ queues.

We consider a daily cycle and divide the 24 hours into periods of length l, where $((j-1)l, jl]$ represents the jth period, $j = 1, \cdots, 24/l$. For staffing purpose, l is often chosen to be one hour or half an hour. We assume that the staffing level changes only at the end of each period. Let $\mathcal{S}^j$ be the set of shifts that are ongoing during the entire period j, s_j be the cardinality of $\mathcal{S}^j$, and $\mu_j(u)$ be the service rate in period j of shift $u \in \mathcal{S}^j$. We estimate $\mu_j(u)$ by the PPH in the corresponding shift hour of shift u. We further consider piece-wise constant arrival rate and let λ_j denote the arrival rate of period j. The stochastic process in period j is the same as an $M/M/s_j$ queue with heterogeneous servers except that at the end of period j, ongoing shifts may end and new shifts may begin, causing instantaneous transitions of system states.

Next, we model the dynamics in the jth period by a time-homogeneous CTMC. Assume that the system has been running for a sufficiently long period of time such that the probability distribution of system states at any time of day is identical for every day. Let t be the time of day and $(x(t), \mathbf{y}(t))$ be the system state at t, where $x(t)$ is the number of patients waiting to be seen, and $\mathbf{y}(t)$ is a s_j-dimensional vector whose ith element $y_i(t)$ represents the status of the physician working on the ith shift in $\mathcal{S}^j$. Specifically,

$y_i(t)$ equals 0 if the physician is idling and 1 otherwise. Assume there is no unforced idling, then we have $x(t)(s_j - \sum_{i=1}^{s_j} y_i(t)) = 0$ for all t. Hence, the dimension of the state space is significantly reduced. Consider an ED with five physicians on duty and assume that the number of patients waiting to be seen is capped at 300. Then, the dimension of the state space is $301 + 2^5 = 333$, whereas that of the model that considers patient returns explicitly exceeds 20 million, as we discussed at the beginning of Section 5.

We apply the uniformization method to the $M/M/s_j$ queue with the uniformization constant $\Lambda_j \triangleq \lambda_j + \sum_{u \in \mathcal{S}^j} \mu_j(u)$. Let $\boldsymbol{\pi}(t)$ be the vector that represents the probability distribution of system states at t. Then, for any pair of t_1, t_2 such that $(j-1)l < t_1 < t_2 < jl$, we have

$$\boldsymbol{\pi}(t_2) = \sum_{n=0}^{\infty} p_j(n) \boldsymbol{\pi}(t_1) P_{1j}^n, \tag{9}$$

where $p_j(n)$ is the Poisson probability mass function with mean $(t_2 - t_1)\Lambda_j$ and P_{1j} is the transition probability matrix of the uniformized system. When $t = jl$, an instantaneous state jump will occur when there are shifts scheduled to begin or end at t. Assume that the instantaneous state transitions are governed by P_{2j}, then $\boldsymbol{\pi}(t) = \boldsymbol{\pi}(t^-)P_{2j}$, where t^- represents the time epoch just before t. Note that P_{2j} is an identity matrix if no shift begins or ends at t. We can calculate $\boldsymbol{\pi}(t)$ for any t with proper truncation of the state space and the sum of the infinite series in (9). With the availability of $\boldsymbol{\pi}(t)$, we can compute the long-run average ED waiting time. The waiting time calculation and the specifications of P_{1j} and P_{2j} are standard but tedious; thus, they are deferred to Appendix C.

6.3. Shift Extension

Our simulation model can also be used to perform what-if analysis. Consider a situation in which the hospital has extra budget and decides to extend the 6-hour shift in our data by one hour, i.e., convert shift S4 to a 7-hour shift by delaying the shift end time to 17:00. The ED manager might be interested in evaluating the impact of one additional physician hour on the average ED waiting time. This question can be addressed by our simulation model, but one challenge is how to decide the PPH rate for the added shift hour. Based on our partition of shifts into three phases, the added hour extends the middle-of-shift phase. Our insight from Section 4.3 states that the PPH rate converges to a constant during the middle-of-shift phase. Hence, we use the average of the fourth and fifth hour's PPH as an estimation of the PPH rate for the added hour.

6.4. Results and Discussion

After solving the optimization problems, we evaluate the average patient waiting time with and without shift extension under the baseline shift schedule and the optimized schedules for the three scenarios by simulation. Hence, a total of eight shift schedules are evaluated. We run the simulation for 500 replications.

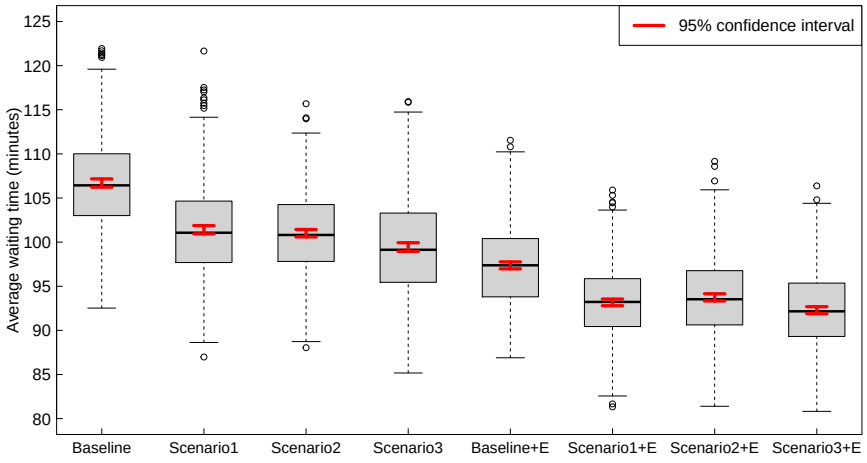
For each replication, we simulate the system for 500 weeks and identify the first 200 weeks as the warm-up period; thus they are removed from the output. We use the remaining 300 weeks to compute the average patient waiting time for each of the 500 replications.

Table 1 The start times of the optimized physician shift schedules under scenarios 1–3. The last two columns show the average patient waiting time (in minutes), the absolute reduction, and the percentage reduction of the optimized schedules over the baseline schedule with and without shift extension, respectively.

	Shift Start Times													Waiting Time (mins)	
	S1	S2	S3	S4	S5	S7	S8	S9	S10	S12	S13	S14	S15	Average	Reduction (%)
Without shift extension															
Baseline (from data)	6	7	8	10	10	12	14	16	16	18	20	23	0	106.7	N/A
Scenario 1	6	7	**10**	**9**	**11**	**14**	**12**	16	**15**	**20**	**18**	23	0	101.4	5.3 (5.0%)
Scenario 2	6	**9**	8	**11**	10	12	14	16	16	18	20	23	0	101.0	5.7 (5.3%)
Scenario 3	**8**	**10**	**6**	**12**	**9**	**11**	**15**	**17**	**14**	18	20	**1**	**22**	99.4	7.3 (6.8%)
With shift extension															
Baseline+E †	6	7	8	10	10	12	14	16	16	18	20	23	0	97.4	9.3 (8.7%)
Scenario 1+E †	6	7	**10**	**9**	**11**	**14**	**12**	16	**15**	**20**	**18**	23	0	93.2	13.5 (12.7%)
Scenario 2+E †	6	**9**	8	**11**	10	12	14	16	16	18	20	23	0	93.7	13.0 (12.1%)
Scenario 3+E †	**8**	**10**	**6**	**12**	**9**	**11**	**15**	**17**	**14**	18	20	**1**	**22**	92.3	14.4 (13.5%)

Note. The shift start times that are different from the baseline schedule are highlighted and marked in bold.
† "+E" indicates that the 6-hour shift is extended by one hour for the optimized shift in the corresponding scenario.

Figure 8 Box plots and 95% confidence intervals for the simulated average patient waiting time based on 500 replications.



The results are shown in Table 1 and Figure 8. We first observe that by adjusting the shift start times, we can achieve a better match between patient demand and ED capacity. As a result, the average patient waiting time can be reduced by 5.0% to 6.8% compared to the baseline schedule, which is equivalent to 13.8 to 19.0 hours of waiting for all patients in the ED per day. (The calculation is based on an average of 156.4

patients arriving daily to the main ED area.) Figure 8 shows that the reductions over the baseline schedule are statistically significant at the 5% level.

Among the three scenarios without shift extension, the schedule from Scenario 2 is particularly interesting, as it achieves a 5.3% reduction in the average waiting time by simply deferring the start times of shifts S2 and S4 by 2 hours and 1 hour, respectively, compared with the baseline schedule. A closer look at the total physician PPH by time of day before and after the changes finds that the adjustments cause the total physician PPH to lag behind the arrival rate function instead of matching it closely. This finding aligns with the observation in Yom-Tov and Mandelbaum (2014). We further observe that extending the 6-hour shift by one hour in the baseline schedule and the schedules from scenarios 1–3 can reduce the average waiting time by up 13.5%, equivalent to 37.5 hours for all patients in the ED per day.

Next, we discuss the practical value of our results. Woodworth and Holmes (2020) find that EDs could save the total healthcare cost approximately 2% to 4% from reducing each patient's waiting time by 10 minutes. Based on public data from a government website,[6] the average cost per ED visit in Alberta, Canada was CA$449.2 in 2015–2016. Table 1 shows that adding one hour to the 6-hour shift in the baseline schedule can reduce the average waiting time by 9.3 minutes (from 106.7 to 97.4 minutes). With 57,086 visits to the main ED area per year (156.4 visits/day multiplied by 365 days), this reduction in waiting time generates annual cost savings for our study hospital from CA$476,960 to CA$953,920. In contrast, the average hourly rate for emergency physicians in Canada is CA$126 in 2021,[7] which means that the annual cost for adding one shift hour per day is CA$45,990. Hence, shift extension is beneficial to the hospital. A similar calculation finds that using the adjusted schedule in scenario 2 over the baseline schedule saves CA$292,330 to CA$584,661 per year. This is achieved by simply deferring the start times of two shifts without adding additional resources. One can calculate the cost savings for other schedules in Table 1 in a similar fashion. Note that these are only rough estimates, as the study by Woodworth and Holmes (2020) is based on a U.S. hospital; moreover, the distribution of the waiting time reductions among different triage levels is unclear in our results, which may affect the calculation. However, we believe that these numbers can still provide insights into the benefits of shift extensions and shift adjustments.

7. Conclusions and Future Research

Motivated by an intriguing observation of a time-varying pattern in physician productivity, we use an optimal control framework to study the decision making of ED physicians within their shifts. We find that the shift-hour-dependent structure is due to the transient behavior of individual physicians during their *finite-length*

[6] Accessed via the Interactive Health Data Application at `www.ahw.gov.ab.ca/IHDA_Retrieval/` on November 3, 2021.

[7] Information obtained from `https://ca.talent.com/salary?job=emergency+physician`; accessed on November 3, 2021.

shifts in a *non-terminating* service system—a healthcare system that operates 24 hours per day and 7 days per week. The behavior at individual physician level has a significant impact on the operational metrics at the ED level. Using data from a Canadian hospital, we demonstrate how to leverage time-varying physician productivity for ED modeling and staffing.

7.1. Managerial Insights

Our study provides several useful insights. First, our findings suggest that time-varying is intrinsic to a physician's productivity, mainly driven by the physician's efforts to maximize throughput by multitasking and minimize patient handoff by strategic idling. Our data analysis finds that nearly half of all new patients were seen in the first two hours of a shift. Therefore, it is crucial to consider shift-hour-dependent service rates in ED modeling and staffing. Using a constant rate—a common practice in the healthcare operations literature—does not capture ED dynamics accurately and can lead to a discrepancy between the expected and actual performances of any staffing plan. Second, our results suggest that the optimal switching time for a physician to stop signing up new patients in a shift does not depend on physician characteristics, which supports the practice of setting a common switching time for all physicians. Third, this study advances our understanding of physician service capacity, which has drawn increasing attention from the operations management community (Saghafian et al. 2012, KC 2013, Campello et al. 2016, Li et al. 2021). We provide a fairly accurate estimate of a physician's capacity. Lastly, our proposed staffing algorithm, which accounts for the time-varying demand of physician productivity, is practically useful for ED managers who aim to improve physician staffing planning.

7.2. Relevance to Non-Healthcare Settings

Shift changes and task handoffs are unavoidable in any non-terminating system where human workers are involved, such as nuclear reprocessing and oil-refining plants (Lardner 1996), aviation maintenance facilities (Parke and Kanki 2008), and auditing firms (The Audit Commission 2012). Worker multitasking is also common in today's workplace, as multiple tasks compete for the same worker's attention. Compared with idly waiting on a pending task, it may be better to switch to process another task, which may increase worker utilization and productivity (Ophir et al. 2009, KC 2013). Hence, the trade-off between productivity and handoff may also exist in other settings, which relates our findings to the management of non-healthcare systems. Our results suggest that understanding the behavior of individual workers during a shift is crucial for predicting system-level performance, which is key to workforce management. We also mark that the dimension reduction technique in our study can be applied to the modeling of case-manager type systems, such as online chat systems in contact centers (Tezcan and Zhang 2014) and case management in social work (Campello et al. 2016).

7.3. Future Research

There are a number of opportunities for future research. First, it would be of interest to extend our proposed queueing model to account for patient LWBS behavior. LWBS is an important aspect of ED operations and depends on the patient's waiting time and the order in which patients are seen (Batt and Terwiesch 2015). Therefore, to model the LWBS behavior, one needs to understand the patient prioritization mechanism used in EDs, which has been found to depend on both clinical and operational factors (Li et al. 2021). Second, our queueing model and the corresponding simulation model focus on the waiting room dynamics and did not consider the subsequent treatment and disposition process in detail such as patient boarding. It would be interesting to integrate our model with the simulation model in Shi et al. (2015), which focuses on the interface between ED and inpatient units, to capture the complete patient journey through the healthcare system. Third, our model is validated using data from two large urban hospitals in Canada. It is important to evaluate the accuracy of our model using data from hospitals of different sizes and in other regions. Finally, Green et al. (2007) point out that the true nature of ED service times remains unclear because physician multitasking, i.e., serving multiple patients at a time, causes disruptions to the service provided to a given patient. Moreover, the estimation of service times from observational data is challenging because the end times for physician activities are usually not recorded (Duan et al. 2020). Our findings suggest that the aggregated service rates (i.e., PPH)—which usually can be easily counted from data—are perhaps adequate for ED modeling and staffing. It would be valuable to investigate when using the aggregated service rates is sufficient and when it is not.

Acknowledgments

The authors would like to thank Professor Nan Liu for early discussions and helpful comments.

References

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

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

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

J. A. Berry Jaeker and A. L. Tucker. Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Science*, 63(4):1042–1062, 2016.

J. O. Brunner and G. M. Edenharter. Long term staff scheduling of physicians with different experience levels in hospitals using column generation. *Health Care Management Science*, 14(2):189–202, 2011.

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

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

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

C. W. Chan, M. Huang, and V. Sarhangian. Dynamic server assignment in multiclass queues with shifts, with applications to nurse staffing in emergency departments. *Operations Research*, 2021.

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

D. S. Cheung, J. J. Kelly, C. Beach, R. P. Berkeley, R. A. Bitterman, R. I. Broida, W. C. Dalsey, H. L. Farley, D. C. Fuller, D. J. Garvey, et al. Improving handoffs in the emergency department. *Annals of Emergency Medicine*, 55 (2):171–180, 2010.

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

M. Delasay, A. Ingolfsson, B. Kolfal, and K. Schultz. Load effect on service times. *European Journal of Operational Research*, 279(3):673–686, 2019.

S. Deo and A. Jain. Slow first, fast later: Temporal speed-up in service episodes of finite duration. *Production and Operations Management*, 28(5):1061–1081, 2019.

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

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

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

K. Epstein, E. Juarez, A. Epstein, K. Loya, and A. Singer. The impact of fragmentation of hospitalist care on length of stay. *Journal of Hospital Medicine*, 5(6):335–338, 2010.

L. V. Green, P. J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39, 2007.

A. Guttmann, M. J. Schull, M. J. Vermeulen, and T. A. Stukel. Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from ontario, canada. *BMJ*, 342, 2011.

Y. Hu, C. W. Chan, and J. Dong. Optimal scheduling of proactive service with customer deterioration and improvement. *Management Science*, 2020.

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

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

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

J. W. Joseph, S. Davis, E. H. Wilker, M. L. Wong, O. Litvak, S. J. Traub, L. A. Nathanson, and L. D. Sanchez. Modelling attending physician productivity in the emergency department: a multicentre study. *Emergency Medicine Journal*, 35(5):317–322, 2018.

J. W. Joseph, S. R. Davis, E. H. Wilker, B. A. White, O. Litvak, L. A. Nathanson, and L. D. Sanchez. Emergency physicians' active patient queues over the course of a shift. *The American Journal of Emergency Medicine*, 2020.

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

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

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

S.-H. Kim, C. W. Chan, M. Olivares, and G. Escobar. ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science*, 61(1):19–38, 2015.

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

R. Lardner. Effective shift handover - A literature review. Technical report, Health & Safety Executive, 1996.

W. Li, Z. Sun, and L. J. Hong. Who is next: Patient prioritization under emergency department blocking. *Operations Research, forthcoming*, 2021.

R. Liu and X. Xie. Physician staffing for emergency departments with time-varying demand. *INFORMS Journal on Computing*, 30(3):588–607, 2018.

R. Liu and X. Xie. Weekly scheduling of emergency department physicians to cope with time-varying demand. *IISE Transactions*, pages 1–30, 2021.

E. Ophir, C. Nass, and A. D. Wagner. Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences*, 106(37):15583–15587, 2009.

B. Parke and B. G. Kanki. Best practices in shift turnovers: Implications for reducing aviation maintenance turnover errors as revealed in asrs reports. *The International Journal of Aviation Psychology*, 18(1):72–85, 2008.

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

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

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

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

D. W. Savage, D. G. Woolford, B. Weaver, and D. Wood. Developing emergency department physician shift schedules optimized to meet patient demand. *Canadian Journal of Emergency Medicine*, 17(1):3–12, 2015.

S. Sethi. *Optimal Control Theory—Applications to Management Science and Economics*. Springer, third edition, 2019.

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

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

H. Song, A. L. Tucker, K. L. Murrell, and D. R. Vinson. Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices. *Management Science*, 64(6): 2628–2649, 2018.

T. Tezcan and J. Zhang. Routing and staffing in customer service chat systems with impatient customers. *Operations Research*, 62(4):943–956, 2014.

The Audit Commission. Guidance on Handover of Audits, 2012. URL `https://assets.publishing.service.gov.uk/media/5329db7040f0b60a73000065/audit_commission_extract_from_standing_guidance_on_handover_of_audits.pdf`. Accessed on October 27, 2021.

L. Vanbrabant, K. Braekers, K. Ramaekers, and I. Van Nieuwenhuyse. Simulation of emergency department operations: A comprehensive review of KPIs and operational improvements. *Computers & Industrial Engineering*, 131: 356–381, 2019.

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

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

F. Zaerpour, M. Bijvank, H. Ouyang, and Z. Sun. Scheduling of physicians with time-varying productivity levels in emergency departments. *Production and Operations Management*, 2021.

Appendices.

Appendix A. Further results from descriptive data analysis

Figure 9 The average new patients seen per hour (PPH) for 8-hour shifts in the main ED area with time resolution of 1 hour. The extra point on the curve for active patients outside the shift duration is due to physician overtime for one hour.

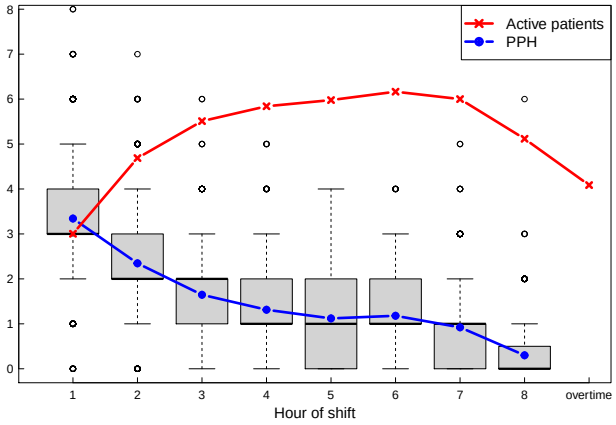
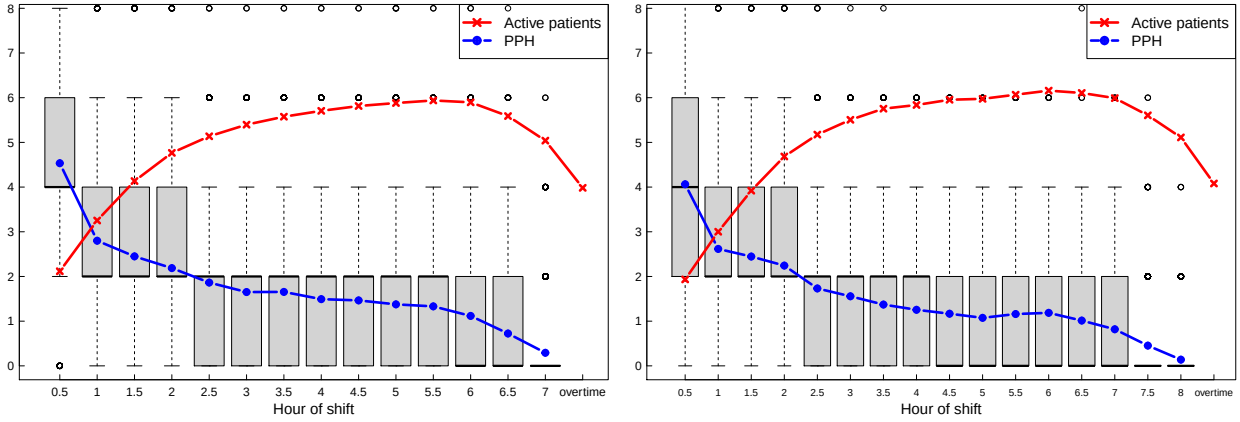


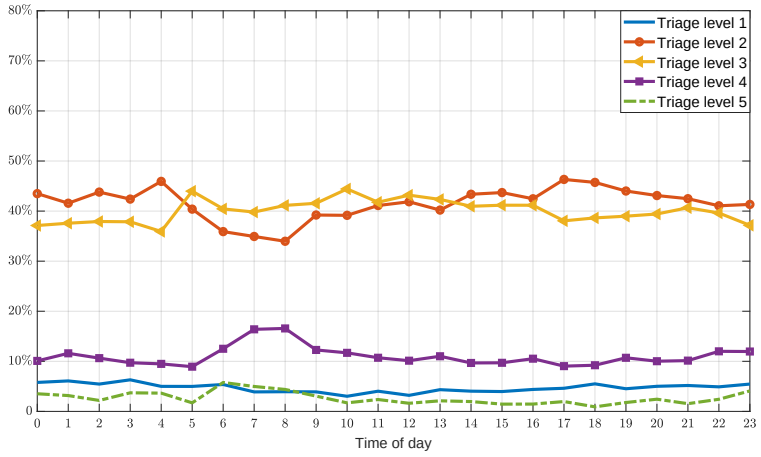
Figure 10 The average new patients seen per hour (PPH) for 7- and 8-hour shifts with time resolution of 30 minutes. The extra point for active patients outside the shift duration is due to physician overtime for one hour.



Appendix B. Proofs for Theorem 1 and Corollary 1

Proof of Theorem 1. We first show that $R(t) = 0,\ \forall t \in [0,T]$. For any t such that $\theta D(t)/\mu_R < 1$, we have $\alpha_R(t) = \theta D(t)/\mu_R$ and hence $R'(t) = 0$. When $\theta D(t)/\mu_R = 1$, we have $\alpha_R(t) = 1$ and $\alpha_N(t) = 0$, which yields $D'(t) = p\mu_R - \theta D(t) = -(1-p)\mu_R < 0$. Because $D(0) = 0$, we conclude that $\theta D(t)/\mu_R \le 1,\ \alpha_R(t) =$

Figure 11 The percentage of patients from each triage level in the main ED area by the time of day.



$\theta D(t)/\mu_R$, and $R'(t) = 0,\ \forall t \le T$. Combining with $R(0) = 0$, we conclude that $R(t) = 0,\ \forall t \le T$. This completes the proof for Theorem 1 (i).

Furthermore, whenever $D(t) = 0$, we have $D'(t) = p\alpha_N(t)\mu_N \ge 0$. Combining with $D(0) = 0$, we conclude that the pure-state constraint $D(t) \ge 0,\ \forall t \in [0,T]$ holds naturally, which simplifies Problem (1) into the following:

$$\begin{aligned} &\max_{\alpha_N(t)} \left\{ \int_0^T (1-p)[\alpha_N(t)\mu_N + \theta D(t)]\,\mathrm{d}t - h(D(T)) \right\} \\ &\text{s.t.} \quad D'(t) = p\alpha_N(t)\mu_N - (1-p)\theta D(t),\ D(0) = 0,\ 0 \le \alpha_N(t) \le 1 - \theta D(t)/\mu_R. \end{aligned} \tag{10}$$

Next, we apply Pontryagin's maximum principle to Problem (10). Denote the co-state variable of $D(t)$ by λ_D. The Hamiltonian is

$$\begin{aligned} H(D,\alpha_N,\lambda_D,t) &= (1-p)[\alpha_N(t)\mu_N + \alpha_R(t)\mu_R] + \lambda_D\left[p\mu_N\alpha_N(t) - (1-p)\theta D(t)\right] \\ &= (1-p)\theta(1-\lambda_D)D(t) + \mu_N\left(p\lambda_D + 1 - p\right)\alpha_N(t). \end{aligned} \tag{11}$$

Note that the Hamiltonian (11) is linear in $\alpha_N(t)$. The Pontryagin's maximum principle requires the Hamiltonian be maximized for all $t \in [0,T]$. Hence, the optimal policy to Problem (10) is bang-bang, i.e., $\alpha_N^*(t)$ is equal to either 0 or $1 - \theta D(t)/\mu_R$. Due to the existence of the mixed inequality constraint, we need to define a Lagrangian by appending the Hamiltonian with the mixed constrains (see Chapter 3 in Sethi 2019). Let μ_L and μ_U be the Lagrange multipliers for the lower and upper constraints on the control $\alpha_N(t)$, respectively. The Lagrangian is

$$\begin{aligned} L(D,\alpha_N,\lambda_D,\mu_L,\mu_U,t) &= H(D,\alpha_N,\lambda_D,t) + \mu_L\alpha_N(t) + \mu_U\left[1 - \theta D(t)/\mu_R - \alpha_N(t)\right] \\ &= \left[(1-p)\theta(1-\lambda_D) - \theta\mu_U/\mu_R\right]D(t) + \left[\mu_N\left(p\lambda_D + 1 - p\right) + \mu_L - \mu_U\right]\alpha_N(t) + \mu_U. \end{aligned} \tag{12}$$

The optimal policy to Problem (10) needs to satisfy the conditions below by Pontryagin's maximum principle:

(i) Maximum Conditions:

$$\alpha_N(t) = 0 \Leftrightarrow p\lambda_D + 1 - p < 0,\ \alpha_N(t) = 1 - \theta D(t)/\mu_R \Leftrightarrow p\lambda_D + 1 - p > 0.$$

(ii) First-Order Conditions: $\mu_N\left(p\lambda_D + 1 - p\right) + \mu_L - \mu_U = 0.$

(iii) Complementary Slackness: $\mu_L \alpha_N(t) = \mu_U\left[1 - \theta D(t)/\mu_R - \alpha_N(t)\right] = 0,\ \mu_L \geq 0,\ \mu_U \geq 0.$

(iv) Adjoint Conditions: $\lambda_D'(t) = \theta\mu_U/\mu_R - (1-p)\theta(1-\lambda_D(t)),\ \lambda_D(T) = -h'(D(T)).$

Because Problem (10) does not contain pure-state constraints, the co-state variable λ_D is continuous in t under optimality. Consider any time t, where $1 - \theta D(t)/\mu_R > 0$ and $\alpha_N(t) = 0$, such that $\alpha_N(t) < 1 - \theta D(t)/\mu_R$. Because of the complementary slackness, $\mu_U = 0$. The adjoint equation for $\lambda_D(t)$ becomes

$$\lambda_D'(t) = (1-p)\theta\lambda_D(t) - (1-p)\theta \Rightarrow \lambda_D(t) = Ce^{(1-p)\theta t} + 1, \tag{13}$$

where C is a constant. We argue that in a nontrivial setting, $C < 0$; otherwise, $\lambda_D(t) > 0$, and hence $\alpha_N(t) = 1 - \theta D(t)/\mu_R$ as a result of $p\lambda_D + (1-p)r > 0$, which contradicts with $\alpha_N(t) = 0$. Hence, $\lambda_D'(t) = (1-p)\theta Ce^{(1-p)\theta t} < 0$ whenever $\alpha_N(t) = 0$. This implies that once $\alpha_N(t) = 0$, then $p\lambda_D(\hat{t}) + 1 - p < 0$ and thus $\alpha_N(\hat{t}) = 0,\ \forall \hat{t} \in [t, T]$. In other words, once it is optimal for the physician to choose idling at t, i.e., $\alpha_N(t) = 0$, then it is optimal to stay idle during the remaining time of her shift.

Solving (13) together with the boundary condition in the adjoint conditions yields

$$\lambda_D(t) = 1 - \left[1 + h'(D(T))\right] e^{(1-p)\theta(t-T)},\ t \in [0, T]. \tag{14}$$

Let t^* denote the optimal threshold under the optimal control policy. The maximum conditions imply that $p\lambda_D(t^*) + 1 - p = 0$. Solving this equation yields

$$t^* = T - \frac{\ln[p(1 + h'(D(T)))]}{(1-p)\theta}. \tag{15}$$

Since the right-hand side of (15) is not necessarily between 0 and T, we let $t^* = 0$ if the right-hand side of (15) is less than 0, and let $t^* = T$ if the right-hand side of (15) is greater than T. This completes the proof for Theorem 1 (ii).

Assume that the physician starts idling at $t^* \in [0, T]$. Then, $\alpha_N(t) = 1 - \theta D(t)/\mu_R$ when $t \in [0, t^*]$ and $\alpha_N(t) = 0$ when $t \in (t^*, T]$. The system dynamics can be described as follows:

$$\frac{\mathrm{d}D(t)}{\mathrm{d}t} = p\left(\mu_N - \frac{\mu_N}{\mu_R}\theta D(t)\right) - (1-p)\theta D(t),\ D(0) = 0,\ t \in [0, t^*],\ \textit{and} \tag{16}$$

$$\frac{\mathrm{d}D(t)}{\mathrm{d}t} = -(1-p)\theta D(t),\ t \in (t^*, T]. \tag{17}$$

Solving the ordinary differential equations in (16) and (17) yields:

$$D(t) = \frac{p\mu_N}{\theta\left(1-p+p\mu_N/\mu_R\right)}\left(1-e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t}\right),\ t\in[0,t^*],\ and \tag{18}$$

$$D(t) = D(t^*)e^{-(1-p)\,\theta(t-t^*)},\ t\in(t^*,T], \tag{19}$$

which completes the proof for Theorem 1 (iii). □

Proof of Corollary 1. Because $\mathrm{PPH}(t) = \alpha_N(t)\mu_N$, we get

$$\mathrm{PPH}(t) = \mu_N - \frac{p\mu_N^2}{p\mu_N+(1-p)\mu_R}\left(1-e^{-\theta\left(1-p+p\mu_N/\mu_R\right)t}\right),\ \forall t\in[0,t^*].$$

It is obvious that $\mathrm{PPH}(t)=0,\ \forall t\in(t^*,T]$. Similarly, since $\mathrm{TH}(t) = (1-p)[\alpha_N(t)\mu_N+\theta D(t)]$, we get

$$\mathrm{TH}(t) = (1-p)\mu_N + \frac{p(1-p)\mu_N(\mu_R-\mu_N)}{p\mu_N+(1-p)\mu_R}\left(1-e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t}\right),\ t\in[0,t^*],\ and$$

$$\mathrm{TH}(t) = (1-p)\theta D(t^*)e^{-(1-p)\,\theta(t-t^*)},\ t\in(t^*,T],$$

where t^* and $D(t)$ are given in (15) and (18), respectively, which completes the proof. □

Appendix C. Specifications of P_{1j}, P_{2j}, and the average ED waiting time

We first specify the transition probability matrix P_{1j}. The transition probability of P_{1j} from $(x_1,\mathbf{y}_1)$ to $(x_2,\mathbf{y}_2)$, denoted by $p_{(x_1,\mathbf{y}_1)\to(x_2,\mathbf{y}_2)}$, is defined as follows:

$$p_{(x_1,\mathbf{y}_1)\to(x_2,\mathbf{y}_2)} = \begin{cases} \frac{\lambda_j}{\Lambda_j} & \text{if } x_1\ge 0, x_2 = x_1+1, \mathbf{y}_1=\mathbf{y}_2=\mathbf{1}_{s_j}; \text{ or } x_1=x_2=0, \mathbf{y}_1\neq\mathbf{1}_{s_j},\\ & \quad \mathbf{y}_2=\mathbf{y}_1+\mathbf{e}_i, \text{where } i=\arg\max_{1\le m\le s_j}\{\mu_j(S_m)(1-\mathbf{y}_1^{(m)})\},\\ \frac{\sum_{m=1}^{s_j}\mu_j(S_m)}{\Lambda_j} & \text{if } x_1\ge 1, x_2=x_1-1, \mathbf{y}_1=\mathbf{y}_2=\mathbf{1}_{s_j},\\ \frac{\mathbf{y}_1^{(m)}\mu_j(S_m)}{\Lambda_j} & \text{if } x_1=x_2=0, \mathbf{y}_2=\mathbf{y}_1-\mathbf{e}_m\mathbf{y}_1^{(m)}, \mathbf{y}_1=\mathbf{y}_2\neq\mathbf{1}_{s_j}, 1\le m\le s_j,\\ 1-\frac{\lambda_j+\sum_{m=1}^{s_j}\mathbf{y}_1^{(m)}\mu_j(S_m)}{\Lambda_j} & \text{if } x_1=x_2=0, \mathbf{y}_1=\mathbf{y}_2\neq\mathbf{1}_{s_j},\\ 0 & \text{otherwise,}\end{cases}$$

where $\mathbf{e}_i$ is the ith row of s_j-dimensional identity matrix, $\mathbf{1}_{s_j}$ is a s_j-dimensional vector with all elements equal to 1, $\mathbf{y}_1^{(m)}$ is the mth element of $\mathbf{y}_1$, and S_m is the mth shift in $\mathcal{S}^j$.

Next, we specify P_{2j} to describe the instantaneous transition of system states at the end of the jth period due to that physicians may go off-duty or begin new shifts. Let ξ_j and η_j be the numbers of physicians who go off-duty and begin new shifts at the end of period j, respectively. Note that ξ_j and η_j are known for any given schedule. If $\xi_j=\eta_j=0$, then P_{2j} is an (s_j+1)-dimensional identity matrix; otherwise, let $p_{(x_1,\mathbf{y}_1)\to(x_2,\mathbf{y}_2)}$ denote the transition probability from $(x_1,\mathbf{y}_1)$ to $(x_2,\mathbf{y}_2)$ of P_{2j}, and let Q be an $s_j\times(s_j-\xi_j)$ matrix whose column corresponds to one of the $(s_j-\xi_j)$ physicians who continue working in period $j+1$.

Each column of Q is a s_j-dimensional vector whose elements are all 0 except that the gth element is 1 where g is the index of the corresponding physician in $\mathbf{y}_1$. Then, when $\eta_j \geq 1$,

$$p_{(x_1,\mathbf{y}_1)\to(x_2,\mathbf{y}_2)} = \begin{cases} 1, & \text{if } x_1 \geq \eta_j, x_2 = x_1 - \eta_j, \mathbf{y}_1 = \mathbf{1}_{s_j}, \mathbf{y}_2 = (\mathbf{y}_1 Q, \mathbf{1}_{\eta_j}); \text{ or } x_1 \leq \eta_j - 1, x_2 = 0, \\ & \quad \mathbf{y}_2 = (\mathbf{y}_1 Q, \mathbf{1}_{x_1}, \mathbf{0}_{\eta_j - x_1}); \text{ or } x_1 = x_2 = 0, \mathbf{y}_2 = (\mathbf{y}_1 Q, \mathbf{0}_{\eta_j}), \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbf{1}_g$ and $\mathbf{0}_g$ are g-dimensional vectors with all elements equal to 1 and 0, respectively. Note that when $\eta_j > x_1$, i.e., the number of waiting patients is less than the number of physicians who begin their shifts, we need to specify the rule to assign patients to physicians (see Section 5.1). When $\eta_j = 0$, then

$$p_{(x_1,\mathbf{y}_1)\to(x_2,\mathbf{y}_2)} = \begin{cases} 1, & \text{if } x_1 = x_2, \mathbf{y}_2 = \mathbf{y}_1 Q, \\ 0, & \text{otherwise,} \end{cases}$$

Finally, we explain the calculation of the average ED waiting time. The total expected patient waiting time in the jth period $((j-1)l, jl]$, denoted by W_j, can be computed as

$$W_j = \sum_{n=0}^{\infty} \left(\frac{(\Lambda_j l)^n}{n!} e^{-\Lambda_j l} \sum_{m=0}^{n} \sum_{i=1}^{\infty} \boldsymbol{\pi}_i((j-1)l + ml/(n+1)) \frac{il}{n+1} \right),$$

where $\boldsymbol{\pi}_i((j-1)l + ml/(n+1))$ is the probability at state $(i, \mathbf{1}_{s_j})$ at $t = (j-1)l + ml/(n+1)$. Hence, the average ED waiting time is $\sum_{j=1}^{24/l} W_j / \sum_{j=1}^{24/l} \lambda_j$.