Submitted to *Manufacturing & Service Operations Management*
manuscript (Please, provide the manuscript number!)

Time-Varying Physician Productivity and Implications for Emergency Department Modeling and Staffing

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

Problem definition: Motivated by an intriguing observation of a time-varying pattern in physician productivity in emergency departments (EDs), we examine the contributing factors to this time-varying pattern analytically and empirically. We then investigate the impact of incorporating time-varying service rates in ED modeling and physician staffing. ***Methodology/results:*** We model the behavior of individual physicians within their shifts using a continuous-time optimal control framework and characterize the structure of the optimal policy. We find that physician multitasking and handoff (or overtime) avoidance may drive individual physicians' transient behavior and contribute to the time-varying pattern in physician productivity. We also provide empirical evidence that shift hour is the most important factor in explaining the variations in physician productivity and predicting physician productivity. We then investigate the impact of incorporating the time-varying physician productivity in ED modeling and staffing. Validated using data from two Canadian EDs, our simulation results demonstrate that the multi-server queuing model with shift-hour-dependent service rates can accurately capture time-of-day-dependent patient waiting times. In contrast, the simulated waiting times under the assumption of constant service rates deviate significantly from the data. Furthermore, a case study using data from a Canadian ED shows that the optimized staffing plan considering the time-varying service rates can improve upon the current physician staffing in practice. In contrast, when the time-varying service rates are ignored and a constant service rate is assumed, the staffing plan generated using the same algorithm performs even worse than the existing one. These findings suggest that accounting for the time-varying physician productivity in ED staffing decisions can lead to substantial cost savings. ***Managerial implications:*** Our findings emphasize the importance and necessity of considering the time-varying nature of physician productivity in the planning and allocation of physician resources to improve ED operations.

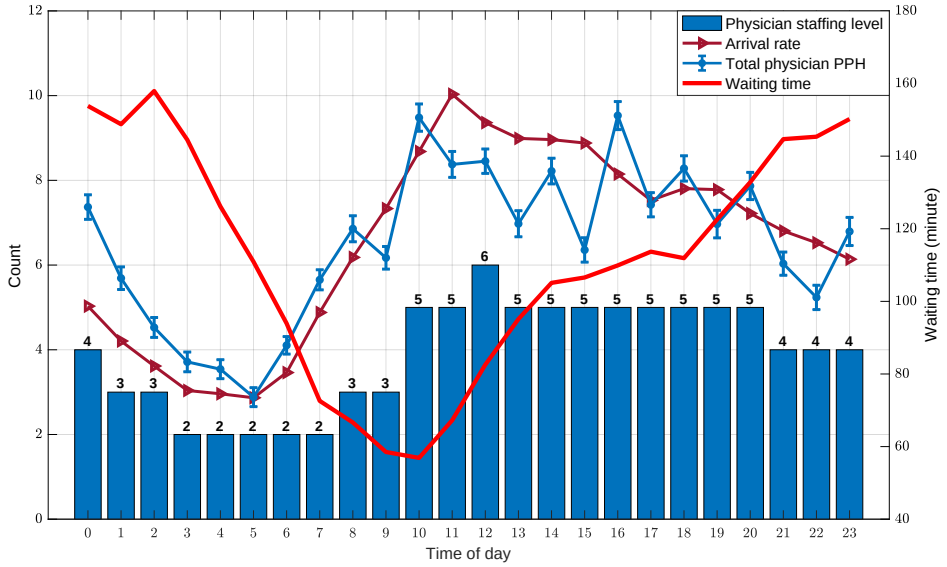
Key words: Healthcare, Emergency Department, Time-Varying Service Rate, Physician Staffing

1. Introduction

Emergency department (ED) overcrowding is a pressing issue facing many countries globally, significantly impacting EDs' ability to provide timely care (Pines et al. 2011), affecting EDs' ability to provide timely care. As a result, extended patient waiting times have become extremely common in many healthcare

systems around the world. Given the critical nature of this challenge, the importance of understanding the driving factors behind ED crowding and long patient waiting times cannot be overstated. The ED patient flow process is inherently complex, characterized by time-varying demand, staggered shift pattern, and a network structure, rendering it nearly impossible to obtain any analytical performance measures. Despite the difficulty in evaluating time-dependent metrics (e.g., waiting times, throughput), these metrics are crucial for system-level decision-making, such as physician staffing optimization.

Figure 1 **Patient arrival rates, average waiting times, physician staffing levels, and total new patients seen per hour (PPH) by all physicians in the main ED area (excluding fast-track area) of our study ED from January to July 2015. The error bars represent 95% confidence intervals.**



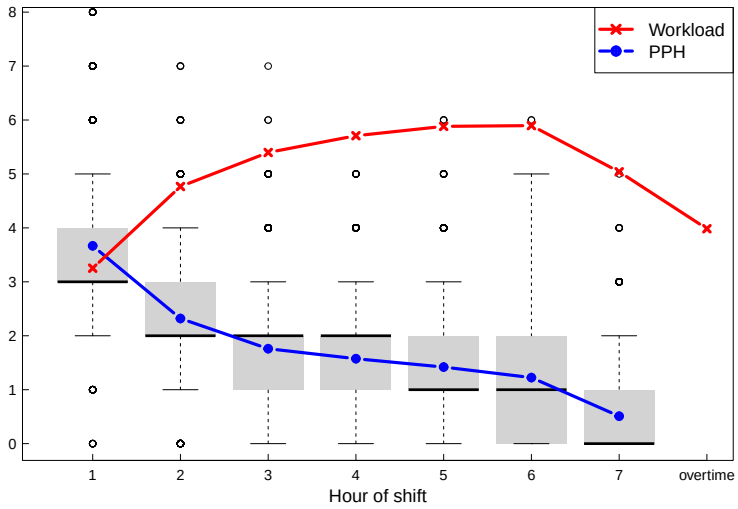
ED operations are naturally modeled as queueing systems, which requires a good understanding of the arrival and service processes. Using patient visit data from an urban tertiary hospital in Alberta, Canada, we plot the average patient arrivals per hour (demand for emergency care), physician staffing levels (ED capacity), and average waiting times (from triage to first assessment by a physician) by the time of day in Figure 1. We observe that the physician staffing level in our study ED is carefully designed to match the time-varying demand. This is done by staggering shifts of different lengths; see a detailed description of the shift structures in Section 3.2. However, the result 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 picking up new patients, measured by PPH—the number of new patients picked up by a physician per hour. Note that PPH is *not* the rate at which patients complete treatment in the ED, which is usually referred to as *throughput*; rather, it is the rate at which physicians pick up new patients and effectively end their waiting in the waiting room. PPH has been used as a measure of a physician's productivity; see, e.g., Joseph et al. (2018, 2021), Zaerpour et al.

2022, and Niewoehner et al. (2023). Hereafter, we use physician productivity and physician service rate interchangeably.

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 the 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 five 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 level of PPH does not coincide with the peak staffing hour, which occurs between noon to 13:00.[1] In addition, the total physician PPH displays a relatively small degree of variability at all times of day, as evidenced by the narrow 95% confidence intervals. This finding underscores the robustness of the PPH pattern. Moreover, we analyzed the total physician PPH by categorizing it into weekdays and weekends. Nevertheless, our statistical analysis shows that the differences between these two categories are statistically insignificant (at the 5% level) for most hours of the day. Motivated by this system-level behavioral anomaly and the wisdom from the 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 physician workload for all 7-hour shifts in the main ED area, using data from January to July 2015. Physician workload refers to all patients under a physician's care at any given time of the shift, which is also referred to as the level of multitasking (KC 2013).**



Most shifts in our study ED are 7 or 8 hours long. 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 10,

[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 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).

Appendix A). Based on the time-varying structure of PPH observed from Figures 2 and 10, we partition 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 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 11 in Appendix A. We observe a similar pattern when we further plot the PPH for each individual physician or a specific type of shift. Similar structures were observed by the emergency medicine community using data from U.S. hospitals (Joseph et al. 2018, 2021). Hence, we conclude that this time-varying pattern of physician PPH 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 patient waiting times and throughput. Hence, it is critical to understand the factors driving this time-varying pattern of physician PPH. To the best of our knowledge, this time-varying pattern of physician PPH has not been thoroughly investigated in the existing literature. Most prior studies have focused on analyzing ED performance under the assumption of constant service rates, overlooking the dynamic nature of physician productivity over the course of a shift; see, e.g., Ingolfsson et al. 2002, Savage et al. 2015, Wang et al. 2022. Therefore, we aim to identify factors that influence physician productivity and assess their respective impacts. Specifically, we focus on two questions: (i) What operational factors contribute to the time-varying physician PPH? (ii) What is the potential impact of incorporating time-varying service rates in ED modeling and physician staffing?

Our study makes the following contributions. First, to understand the contributing factors to the time-varying physician PPH, we use an optimal control framework to model a physician's decision problem within a finite-length shift, balancing the trade-off between throughput and patient handoff (i.e., transfer of patient care from one physician to another). We obtain closed-form expressions for the time-varying PPH under the optimal policy, and our results reveal a structural similarity between the PPH derived from our solution and the empirical PPH from data. This similarity motivates a plausible mechanistic explanation for the time-varying pattern of PPH: (i) physician multitasking behavior contributes to the exponential decay during the start-of-shift phase, and (ii) physician handoff and overtime avoidance behavior leads to the sharp drop in the end-of-shift phase. Another insight from our analytical results is that allocating additional resources to test centers not only enhances test turnaround times but also encourages physicians to attend to more new patients due to reduced concerns about overtime and patient handoffs. Our results also support setting a common cutoff time for signing up new patients approaching the end of their shift.

Second, we conduct an empirical study to investigate behavioral factors contributing to the time-varying physician productivity, such as physician workload and system congestion, which complements mechanistic factors identified by the optimal control framework. We find that shift hour is the most important feature in

explaining the variation in PPH and predicting PPH. In conclusion, our results suggest that time-varying is the nature of physician productivity and shift-hour-dependent service rates should be considered in ED modeling and staffing. Hence, this study advances our understanding of the time-varying physician productivity.

Third, we investigate the impact of considering the time-varying physician productivity in ED modeling and physician staffing. Building on the time-varying nature of PPH, we model the complex ED system by a multiserver queue with nonstationary Poisson arrivals and exponential service times with time-varying rates. The simulation results show that our model produces time-of-day-dependent performance metrics that closely match the data from two Canadian EDs. In contrast, the outputs of the same simulation model that ignores the time-varying physician service rates (i.e., use a constant rate) deviate significantly from the data. Furthermore, through a case study using data from a Canadian ED, we find that the optimized staffing plan that considers the time-varying service rates outperforms the current physician staffing. Conversely, when the time-varying service rate is ignored and a constant service rate is assumed (which is the prevalent practice), the staffing plan generated using the same algorithm performs even worse than the current staffing plan in practice. Hence, our results highlight the importance of considering time-varying physician service rates in ED modeling and staffing decisions.

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 investigate the contributing factors to physicians' time-varying productivity through an optimal control model in Section 4 and empirically in Section 5. We study the impact of considering time-varying service rates in ED modeling in Section 6 and physician staffing in Section 7. Section 8 concludes the paper and points 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 an overview). Our work aims to better understand the decision-making of ED physicians underlying their time-varying productivity and thus is relevant to studies of healthcare workers' behavioral issues. Evidence has shown that healthcare workers adjust their service rates when faced with a heavy workload (KC and Terwiesch 2009, Powell et al. 2012, Ding et al. 2024), high level of multitasking (KC 2013), and overcrowded systems (KC and Terwiesch 2012, Armony et al. 2015, Berry Jaeker and Tucker 2016, Batt and Terwiesch 2016). Physicians may also adapt their patient prioritization behavior (Ding et al. 2019, Li et al. 2023), admission decisions (Kim et al. 2015, 2020, Freeman et al. 2016), and routing decisions (Freeman et al. 2021, Lu and Lu 2018) to the level of system congestion. Many studies have investigated other behavioral factors and mechanisms in healthcare settings. Interested readers are referred to KC et al. (2020) and Cho et al. (2019) for overviews on this topic.

Among them, studies that explore physicians' behavior related to shifts are particularly relevant to our study. Using a parametric hazard model, Batt et al. (2019) study the rate at which physicians complete

patient treatments in EDs and find that the rate is lowest early in the shift and highest toward the end of shift. Moreover, handed-off patients experience a slightly higher treatment rate and 72-hour revisit rate than non-handed-off patients. Using a simulation study, Batt et al. (2019) examine how to reduce handoffs by adjusting shift length and new patient cutoff rules. Similarly, Chan (2018) find that ED physicians are less likely to accept new patients and tend to speed up the treatment of existing patients near the end of shift. Deo and Jain (2019) examine the change in system speed using data from an outpatient department, where patient treatments must be completed before the end of the service episode (unlike EDs). They find that the service speed of a patient is slower at the start and progressively increases toward the end of the service episode. The differences between the studies above and our work are twofold: First, we focus on the rate of ED physicians picking up new patients within their shift (i.e., PPH). We find that PPH is the highest at the start of a shift, plateaus in the middle, and drops to its lowest approaching the end of shift. Our results suggest that both physician multitasking due to the repetitive nature of emergency care and physicians' efforts to avoid patient handoffs may be the driving force behind the changes in PPH over a shift. Second, PPH is the rate that effectively ends a patient's waiting in the waiting room. Through a data-calibrated simulation model, we demonstrate that considering the time-varying PPH helps build accurate models for ED patient flow, and models ignoring it generate outputs that deviate significantly from data, which further differentiates our study from the literature. It is worth noting that a recent study Niewoehner et al. (2023) find that working with familiar peers can increase ED physicians' PPH in a shift which sheds light on the variation in PPH from an interesting organizational perspective, and Zaerpour et al. (2022) empirically identify factors correlated with PPH (without explaining the mechanism), and then leverage this knowledge to assign physicians to predetermined shifts. In contrast, our study investigates the factors that contribute to the time-varying nature of PPH using an optimal control framework and empirical data. We then demonstrate through a data-calibrated simulation model that it is critical to incorporate the time-varying nature of PPH in modeling ED patient flow and physician staffing. Thus, both studies are relevant to ours but different in their research methodologies and 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 patient flow. 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). Whitt and Zhang (2017) propose an infinite-server queueing model of the ED with a time-varying arrival process, where the length of stay is used as the patient service time. Simulation results show the importance of considering the time-dependent nature of the service time, which aligns with the insight from our study. In contrast to the infinite-server model in Whitt and Zhang (2017), our model explicitly accounts for the time-dependent physician staffing level in the model, which can support physician scheduling (as shown by the case study in Section 7).

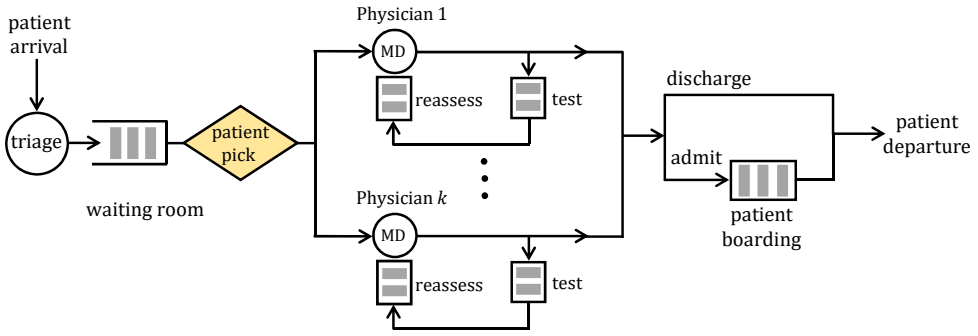
Methodology-wise, we use the optimal control framework to capture the trade-off between throughput and patient handoff to understand individual physicians' transient behavior during a shift. Hence, works that use fluid control models to support decision-making in healthcare systems (Hu et al. 2022, Chan et al. 2021) are most relevant to our study. Hu et al. (2022) use an optimal control framework to study decisions on allocating 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.

Finally, we note the emergency medicine community has also observed a time-varying pattern of physician productivity levels. Joseph et al. (2018) find that estimating physician productivity as a simple average substantially misestimates physicians' capacity and suggests that the time-varying pattern should be factored into physician staffing. Joseph et al. (2021) find that a decrease in PPH does not reflect 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. Nevertheless, we believe that the key features (such as patients' return for service) are shared among 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. In our study ED, the chief nurse decides which patient to move to the treatment room when an ED bed becomes available. When a physician becomes available, she will choose a patient from the roomed patients for initial assessment[2] based on a given prioritization rule (Ding et al. 2019, Li et al. 2023). Physicians occasionally select patients from the waiting room directly. After the 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.

In our study ED, when a physician finishes an ongoing task, she logs into the ED information system through a terminal. The upper half of the screen shows the reassessment requests from her existing patients, and the lower half shows all the new patients waiting to be seen. The information on 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 to limit patients' length of stay. Physicians may also follow the shortest processing time rule because reassessment generally takes less time than treating a new patient. This has important implications for our model in Section 4 as we assume physicians prioritize reassessment tasks over signing up a new patient. At last, a patient departs the ED if discharged; otherwise, the patient is admitted and becomes a boarding patient, waiting in an ED bed until being transferred to an inpatient bed.

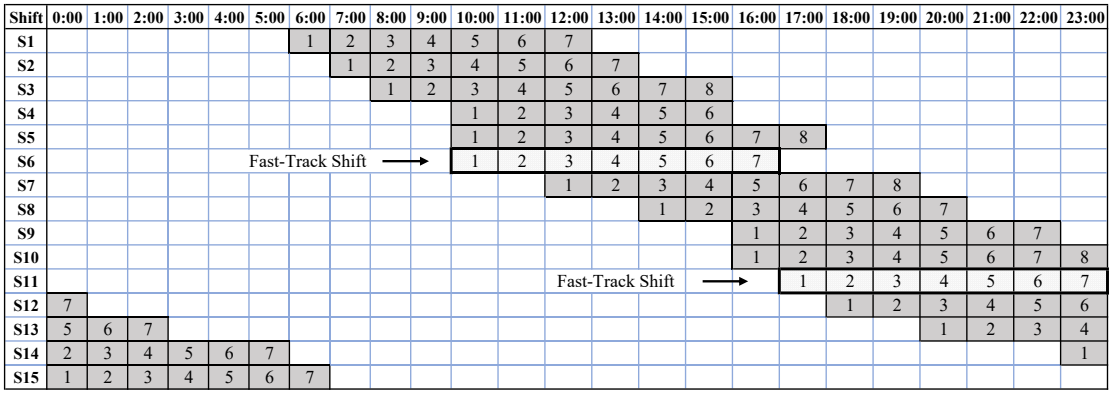
It is well known that ED physicians are multitasking (KC 2013, Song et al. 2018, Niewoehner et al. 2023); i.e., at any given time, a physician is responsible for the care of multiple patients simultaneously. Some of these patients are undergoing testing in the test queue while others are waiting for reassessment (see Figure 3). The number of patients under a physician's care at any given time is referred to as the physician's workload or this physician's level of multitasking (KC 2013). See Figure 2 for an illustration of the physician workload by shift hour calculated using our data.

3.2. 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 our study ED's daily physician shifts from January to July 2015. During this period, 15 shifts (and hence 15 physicians) were scheduled in the ED each day, two

[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 3 to July 31, 2015.



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

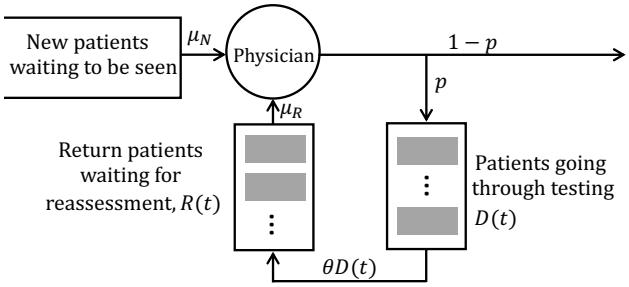
of which were fast-track shifts, and the remainder were scheduled in the main area. The shift lengths in our study ED were 6, 7, or 8 hours. We observe that physicians started their shifts at staggered times during the day to better match physician capacity with time-varying patient demands. Moreover, the staggered shifts avoid the undesirable situation of too many physicians leaving work at the same time and thus make the end-of-shift transition easier. We elaborate below.

When approaching the end of the shift, a physician must 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, physicians in our study ED in a Canadian hospital can choose *not* to see new patients in the last hour of their shifts, even if 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 or administration.

4. Physician Behavior Behind Time-Varying Productivity

In this section, we model the treatment process of any individual physician using an optimal control framework. We obtain closed-form expressions for a physician's productivity under the optimal policy. Understanding individual-physician within-shift behavior helps explain the time-varying productivity of physicians.

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



4.1. Model Description

We consider a fluid model with returns to describe the patient treatment process of 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. Let $D(t)$ and $R(t)$ denote the number of patients in the test and reassess queues at time t, respectively. Let μ_R denote the rate at which return patients are served. 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. (2023). The service and reassessment times, testing times, and return probability are assumed to be independent 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 the shift, if a physician still has patients with incomplete care, the physician either goes overtime to finish the treatment or hands off these patients to other physicians (see Section 3.2), or both could happen. Note that handoffs could happen before the shift is over in practice (Batt et al. 2019). We assume that handoffs do not happen before T to simplify our model and analysis. Let $h(x)$ denote the cost when there are x patients with incomplete care at the end of the shift, $x\geq 0$. The cost may represent the inconvenience caused by physician overtime, the time and effort required for handoff communication to transfer essential information from one physician to another, and/or the compromised quality of care due to handoffs (Cheung et al. 2010, Batt et al. 2019). It is expected that $h(\cdot)$ is a non-decreasing function. In this model, we do not account for the potential

impact of physician fatigue on the quality of care delivered over the course of a shift. We list it as one of the future research directions in Section 8.

We further assume that return patients are prioritized over new patients, which generally aligns with the practices in our study hospitals (see detailed descriptions in Section 3.1). We believe that the primary goal of a physician is to treat as many patients as possible within her shift without exceeding overtime limits or resorting to excessive patient handoffs. Hence, we further assume that physicians do not idle when there are patients waiting for reassessment. However, physicians can choose not to see new patients to avoid overtime and/or handoffs, even if 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 using the optimal control framework as follows:

$$\begin{aligned}
\max_{\alpha_N(t),\alpha_R(t)} \ & \left\{ \int_0^T (1-p)\left[\alpha_N(t)\mu_N + \alpha_R(t)\mu_R\right] \mathrm{d}t - h(D(T)+R(T)) \right\} \qquad (1)\\
\text{s.t.} \quad & D'(t) = p\left[\alpha_N(t)\mu_N + \alpha_R(t)\mu_R\right] - \theta D(t),\ D(t) \geq 0,\ D(0) = D_0 \geq 0,\\
& R'(t) = \theta D(t) - \alpha_R(t)\mu_R,\ R(t) \geq 0,\ R(0) = R_0 \geq 0,\\
& 0 \leq \alpha_N(t) + \alpha_R(t) \leq 1,\ \alpha_R(t) = \min\left\{1, \theta D(t)/\mu_R + \mathbf{1}_{\{R(t)>0\}}\right\},\ \alpha_N(t) \geq 0.
\end{aligned}$$

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; $\alpha_R(t) = \min\{1, \theta D(t)/\mu_R + \mathbf{1}_{\{R(t)>0\}}\}$ captures that physicians do not idle when there are return patients waiting. Specifically, $\mathbf{1}_{\{R(t)>0\}}$ is an indicator function which equals to 1 if $R(t) > 0$ and 0 otherwise. This constraint further implies that when $R(t) > 0$, i.e., there are return patients waiting for reassessment, then $\alpha_R(t) = 1$, i.e., the physician should only focus on reassessing return patients. In other words, return patients are prioritized over new patients. This assumption is not too restrictive as it aligns with the physician workflow (See Section 3.1 for more details). Furthermore, 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. The initial conditions $D(0) = D_0$ and $R(0) = R_0$ imply that a physician who just began her shift has D_0 patients in the test queue and R_0 patients waiting for reassessment, both of which are handoff patients from other physicians.

4.2. The Optimal Policy

We solve the optimal control problem (1) by applying the Pontryagin's maximum principle. We consider three cases: (i) $R_0 = 0, D_0 \leq \mu_R/\theta$; (ii) $R_0 = 0, D_0 > \mu_R/\theta$; (iii) $R_0 > 0$. In the main body of the paper, we present the results for Case (i). We focus on the first case because, in general, return patients who are waiting for reassessment will not be handed off to other physicians. The focal physician usually goes overtime to

finish the reassessment of return patients in our study hospital. In fact, we do not observe any patients waiting for reassessment being handed off in our data (i.e., $R_0 = 0$). Physicians tend to avoid excessive patient handoffs. We observe from our data that the number of handoffs taken by any physician at the beginning of their shifts is less than or equal to 6 for 97.7% of the shifts ($\mu_R/\theta \approx 6.1$ in our data).[3] Hence, Case (i) is the most relevant case. Nevertheless, we study the optimal policies for Cases (ii)&(iii) for mathematical completeness. The optimal controls for Cases (ii)&(iii) have the same threshold structure as that of Case (i) despite the solutions being more complicated. The results for Cases (ii)&(iii) and proofs for all three cases are deferred to Appendix B.

THEOREM 1. *Assume that $D_0 \le \mu_R/\theta$, $R_0 = 0$, and $h(\cdot)$ is an increasing differentiable function. Then, the optimal control $\alpha_N^*(t)$ for the optimal control problem defined in* (1) *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 $t \in (t^*,T]$. Furthermore, under the optimal policy, $\alpha_R^*(t) = \theta D(t)/\mu_R$ for all $t \in [0,T]$ where*

$$D(t) = \left(D_0 - \frac{p\mu_N}{\theta\left(1-p+p\mu_N/\mu_R\right)}\right) e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t} + \frac{p\mu_N}{\theta\left(1-p+p\mu_N/\mu_R\right)},\ t \in [0,t^*],\ \textit{and} \tag{2}$$

$$D(t) = D(t^*)e^{-(1-p)\theta(t-t^*)},\ t \in (t^*,T]. \tag{3}$$

Theorem 1 completely characterizes the optimal policy for Problem (1) under Case (i). The optimal policy implies that 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 chance of overtime and patient handoffs. In fact, the proof of Theorem 1 does not require $h(\cdot)$ to be increasing. However, if $h(\cdot)$ is decreasing, it is easy to see from the expression of t^* in (4) 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. A numerical illustration of the optimal controls is shown in Figure 6.

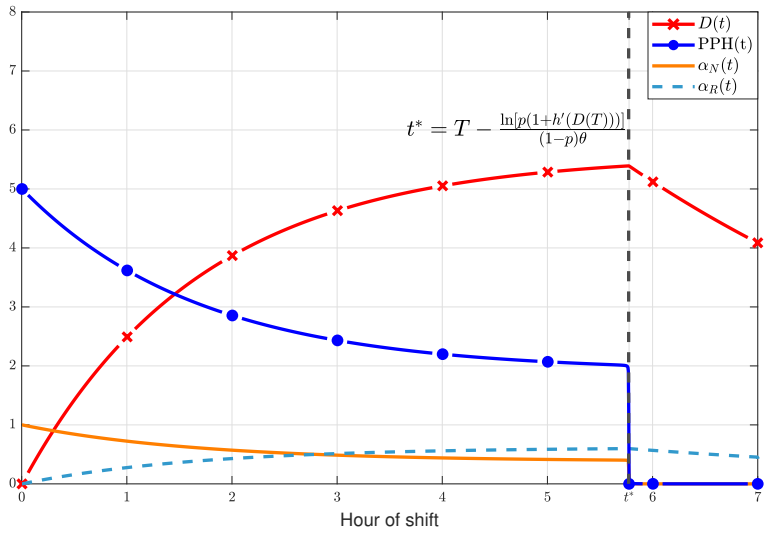
PROPOSITION 1. *Assume that $D_0 \le \mu_R/\theta$ and $R_0 = 0$. The optimal switching time t^* has a closed-form expression as follows:*

$$t^* = \min\left(T, \max\left(0, T - \frac{\ln[p(1+h'(D(T)))]}{(1-p)\theta}\right)\right). \tag{4}$$

Moreover, t^* is non-decreasing in θ and non-increasing in p when $h(\cdot)$ is a linear function.

[3] In our data, the start time of a reassessment is available but not its end time. We use the start time of the following activity of the same physician to approximate the end time of the reassessment, which may overestimate the reassessment time, or in other words, underestimate μ_R.

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 $D_0 = 0, R_0 = 0, \mu_N = 5,\ \mu_R = 6,\ \theta = 0.6,\ p = 0.66$, $h(x) = x$, and $T = 7$ (i.e., 7-hour shifts). The optimal switching time $t^* = 5.8$.**



When the cost function is linear, i.e., $h'(\cdot)$ is a constant, the expression of t^* in (4) shows that it 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 and the test probability p only depends on patient clinical requirements. 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). Moreover, the monotonicity of t^* implies that the optimal switching time is greater for bigger θ and/or smaller p. In other words, faster test turnaround times or a smaller likelihood of requiring additional tests allow physicians to continue seeing new patients further along in their scheduled shift, rather than having to stop earlier. A key insight for hospital management is that allocating additional resources to test centers not only enhances test turnaround times but also encourages physicians to attend to more new patients due to reduced concerns about overtime and/or patient handoffs.

Proposition 2. *Assume that* $D_0 \le \frac{p\mu_N\mu_R}{\theta[p\mu_N+(1-p)\mu_R]}$ *and* $R_0 = 0$. *The optimal control* $\alpha_N^*(t)$ *is non-increasing in* μ_N *and non-decreasing in* μ_R *for* $0 < t < T$. *Moreover,* $\alpha_N^*(t)$ *decreases with* θ *for* $0 < t < t^*$.

Proposition 2 shows that when D_0 is sufficiently small, the optimal percentage of time the physician spent on new patients increases with μ_R and decreases with μ_N and θ. When the chance of requiring reassessment remains unchanged, the physician will spend less time on reassessment if the reassessment can be performed faster. As a result, the physician will spend more time treating new patients. On the other hand, when physicians can treat new patients faster or the test turnaround time is shorter, more time will be spent

reassessing return patients and less time treating new patients. However, this does not necessarily mean that the total physician time spent on new patients is less as the switching time t^* increases with θ (shown in Proposition 1); that is, with a shorter test turnaround time, physicians continue seeing new patients and stop at a later stage of their shift.

4.3. Time-Varying Physician Productivity

Let $\text{PPH}(t)$ denote a physician's productivity rate, i.e., the rate of assessing new patients. Then, we have $\text{PPH}(t) = \alpha_N(t)\mu_N$. Theorem 1 gives the following results immediately.

Proposition 3. *Assume that $D_0 \le \mu_R/\theta$ and $R_0 = 0$. Under the optimal policy for Problem* (1)*, we have*

$$PPH(t) = \frac{(1-p)\mu_N\mu_R}{p\mu_N + (1-p)\mu_R} + \mu_N\left(\frac{p\mu_N}{p\mu_N + (1-p)\mu_R} - \frac{\theta D_0}{\mu_R}\right)e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t},\ t \le t^*, \tag{5}$$

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

where $D(t^*)$ and t^* are given in (2) *and* (4)*, respectively.*

The expressions of $\text{PPH}(t)$ in (5) and (6) show that a physician's productivity within a shift is time-varying. A numerical illustration of $\text{PPH}(t)$ is shown in Figure 6. The fact that $\text{PPH}(t)$ 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 physician multitasking may be one main contributor to the dramatic reduction in physician productivity, 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 $\text{PPH}(t)$ 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, the overtime/handoff avoidance may incentivize physicians to not sign up any new patients; as a result, $\text{PPH}(t)$ drops to zero. This decision 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 worked overtime in 90% of these shifts. We should also note that $\text{PPH}(t)$ does not always decreases with t. In fact, the expression of $\text{PPH}(t)$ in (5) implies that when $D_0 > \frac{p\mu_N\mu_R}{\theta[p\mu_N+(1-p)\mu_R]}$, $\text{PPH}(t)$ increases with t. The intuition is that excessive handoff patients at the start of shift will take up physicians' time, and thus, fewer new patients will be attended to. However, our data shows that in over 75.6% of the shifts, the number of handoff patients taken by any physician in the first hour of their shifts is less than two, whereas a rough estimate shows that $\frac{p\mu_N\mu_R}{\theta[p\mu_N+(1-p)\mu_R]}$ is about 1.7 using our data.

5. Contributing Factors to Time-Varying Productivity

Our findings in the previous section suggest that physician productivity is a function of the shift hours. In fact, a descriptive analysis of our data reveals 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). Thus, shift hour appears to be an important predictor of PPH. In this section, we examine other potential factors that may contribute to the time-varying physician productivity. Note that our objective is *not* to establish a causal relationship between PPH and the factors of interest. Rather, we aim to identify the factors that have the greatest power in explaining the variations of PPH and predicting PPH.

5.1. Discussion on Existing Mechanisms

We observe the time-varying pattern of physician productivity, specifically physicians slowdown in treating new patients as their shifts progress. Our results suggest that physician multitasking (i.e., treating both new and return patients) and "slacking off" may contribute to time-varying physician productivity. However, there may be other factors that impact workers' productivity through various mechanisms. For example, fatigue may reduce a physician's speed and influence the behavior of hospital personnel (Dai et al. 2015), paramedics (Brachet et al. 2012, Bavafa and Jónasson 2024), and food inspectors (Ibanez and Toffel 2020); frequent tasking switching has a negative impact on physician productivity (Duan et al. 2020); 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); and greater average familiarity among physicians may increase the patient pickup rate and multitasking levels (Niewoehner et al. 2023); physicians may tend to batch admission requests approaching the end of shift, increasing physician productivity in a shift (Feizi et al. 2023). Furthermore, social loafing may dominate social pressure speedup. Specifically, nurses may work slower intentionally to avoid being assigned new patients due to a higher expected workload (Berry Jaeker and Tucker 2012). The impact may also be nonlinear. Physician multitasking levels have been found to have a nonlinear effect on throughput rates (KC 2013), and waiting room census has a nonlinear effect on ED treatment times (Batt and Terwiesch 2016). Physicians may first speed due to high workload and then slow down after long periods of increased load (KC and Terwiesch 2009).

Hence, the time-varying PPH patterns we observe in Figures 2 and 10 may be an aggregation of several lower-level mechanisms. Our goal is to examine other potential factors that contribute to the time-varying productivity. Next, we compare the directions and magnitudes of their respective impacts using our dataset.

5.2. Empirical Investigation

Our dataset contains patient visit records and physician staffing data from January 3 to July 31, 2015. We chose this period because the physician staffing remains the same during this period. Our data includes shifts with lengths of 6 hours, 7 hours, and 8 hours. We focus on the 7-hour and 8-hour shifts because the 6-hour

shifts are flexible shifts, and the data on 6-hour shifts are less reliable. After excluding the two fast-track shifts each day, our dataset contains 2475 shifts, of which 1654 are 7-hour shifts and 821 are 8-hour shifts. More details on the shifts are available in Section 3.2.

5.2.1. Choice of Variables We aim to identify the factors that contribute to physician productivity. Hence, the response variable is the number of new patients seen by a physician during a specific hour of a shift (that is, the PPH). As for the independent variables, both Figure 2 and our analytical results in Section 4 show that the hour of shift has a significant impact on PPH. Therefore, we include a categorical variable *ShiftHour* to control the hour of shift. To account for the heterogeneity in physician characteristics such as age, gender and experiences, we include the unique physician ID, denoted by *Physician*. Since shifts have different starting times and lengths, we include a categorical variable, *ShiftID*, to differentiate between the shifts.

As discussed in 5.1, the workload of the physician and the level of congestion of the system may impact the productivity of the physician. Hence, we define the variable *Workload* to measure the time-averaged number of patients under the care of the focal physician during the hour when PPH is measured. Following Zaerpour et al. (2022) and the literature we discussed earlier, we also define *WaitRoomCensus* and *TreatRoomCensus*, which respectively measure the time-averaged number of patients in the waiting room and ED treatment room (excluding boarding patients). Li et al. (2023) find that ED decision-makers may alter their patient pickup behavior due to many ED beds being occupied by boarding patients (this phenomenon is referred to as *ED blocking*). Hence, it is plausible that the number of boarding patients or the ED blocking level may impact physicians' decisions in attending new patients. We define *Boarder* as the time-averaged number of boarding patients, i.e., admitted patients waiting to be transferred to inpatient beds during the hour when PPH is measured. This variable corresponds to the second measure of the ED blocking level in Li et al. (2023). We also examine all results by using the first measure of the ED blocking level in Li et al. (2023). Although the results are not included in the paper, we find all our conclusions remain unchanged.

Physicians may engage in activities other than seeing new patients, such as reassessing returning patients, handing their patients over to receiving physicians, or taking handoff patients from other physicians. All of these activities consume time and may affect the productivity of the physician. Therefore, we define three variables: *Reassess*, *Handover*, and *HandoverTaken*. These variables measure the number of reassessments performed by the focal physician, the number of patients handed over to other physicians, and the number of patients taken from other physicians during the hour when PPH is measured. Additionally, we control the day of week that the shift is scheduled on using the binary variable *Weekend*, which equals 1 if it is on the weekend and 0 otherwise. Summary statistics for all variables, except physician and shift ID fixed effects, are provided in Table 1.

Table 1 Summary Statistics for Variables of Interest

	7-hour shifts					8-hour shifts				
	Mean	SD	Min	Max	95% CI	Mean	SD	Min	Max	95% CI
PPH	1.77	1.42	0.00	13.00	(1.74, 1.79)	1.52	1.31	0.00	8.00	(1.49, 1.55)
Workload	6.44	3.76	0.00	18.00	(6.37, 6.51)	5.71	3.37	0.00	17.00	(5.63, 5.80)
Reassess	0.96	1.16	0.00	7.00	(0.94, 0.98)	0.82	1.06	0.00	6.00	(0.79, 0.84)
HandoverTaken	0.62	1.52	0.00	10.00	(0.59, 0.65)	0.35	1.11	0.00	8.00	(0.32, 0.37)
Handover	0.05	0.34	0.00	5.00	(0.05, 0.06)	0.06	0.38	0.00	5.00	(0.05, 0.07)
WaitRoomCensus	13.46	6.37	0.00	37.14	(13.34, 13.57)	12.91	5.99	0.34	37.14	(12.77, 13.06)
TreatRoomCensus	37.41	7.39	10.64	65.39	(37.27, 37.54)	40.83	7.03	16.32	65.39	(40.66, 41.00)
Boarder	11.38	4.64	0.92	33.37	(11.29, 11.46)	11.93	4.74	1.29	33.37	(11.82, 12.05)
Weekend	0.29	0.45	0.00	1.00	(0.28, 0.29)	0.29	0.45	0.00	1.00	(0.28, 0.30)
Observations	11,548					6,568				
Shift Counts	1,654					821				

Notes. SD = standard deviation; CI = confidence interval.

5.2.2. Econometric Models and Results We start with a linear regression model with the following model specification:

$$PPH = \beta_0 + \beta_1 ShiftHour + \beta_2 ShiftID + \beta_3 Workload + \beta_4 \mathbf{X} + \beta_5 Weekend + \epsilon, \tag{7}$$

where the vector $\mathbf{X}$ include (i) the physician-level activities *Reassess, Handover, HandoverTaken*; and (ii) the system-level overcrowding measures *WaitRoomCensus, TreatRoomCensus,* and *Boarder*. The time of day has a strong collinearity issue and was not included in the model. The error term ϵ follows a standard normal distribution.

We estimate our first model by including all the variables described in Section 5.2.1. We then drop the variable $\mathbf{X}$, *Workload*, *ShiftID* one by one and check the goodness of fit with and without the variables of interest. We estimate four models for 7-hour and 8-hour shifts, respectively. The estimation results are presented in Table 2.

From Table 2, we first observe that across the models for 7-hour shifts, all variables of interest included in the model have statistically significant effects on physician productivity (PPH), except for *Weekend*. Notably, for the 8-hour shifts, *Weekend* is weakly significant. In terms of model fitness, we find that removing $\mathbf{X}$ from the model only marginally decreases the R^2 from 0.478 to 0.465, representing a 2.7% drop. This indicates that although the ED census, including the number of patients in the waiting room, patients in the treatment room, and boarding patients, has statistically significant effects on PPH, the magnitude of the impact is small. Similarly, the R^2 decreases from 0.465 to 0.461 (less than a 1% drop) if we remove *Workload* from the model. In contrast, the two factors *ShiftHour* and *ShiftID* explain 41.4% of variations in PPH. This significant explanatory power highlights the strong influence of these temporal and individual factors on physician productivity. The observations for 8-hour shifts are similar, except that the model fitness is better than that of the 7-hour shifts.

Table 2 Estimation results for the effect of various factors on physician productivity.

	7-hour-shift models				8-hour-shift models			
(Intercept)	4.334***	3.912***	3.777***	3.471***	4.265***	4.076***	3.969***	3.411***
	(0.122)	(0.11)	(0.109)	(0.038)	(0.125)	(0.104)	(0.104)	(0.04)
ShiftHour (base = ShiftHour1)								
ShiftHour2	-1.14***	-1.174***	-1.319***	-1.319***	-0.826***	-0.854***	-0.991***	-0.991***
	(0.04)	(0.04)	(0.036)	(0.038)	(0.049)	(0.049)	(0.047)	(0.048)
ShiftHour3	-1.597***	-1.631***	-1.872***	-1.872***	-1.381***	-1.465***	-1.69***	-1.691***
	(0.045)	(0.045)	(0.036)	(0.038)	(0.055)	(0.054)	(0.047)	(0.048)
ShiftHour4	-1.732***	-1.746***	-2.032***	-2.033***	-1.616***	-1.72***	-2.022***	-2.023***
	(0.048)	(0.048)	(0.036)	(0.038)	(0.061)	(0.059)	(0.047)	(0.048)
ShiftHour5	-1.894***	-1.87***	-2.174***	-2.173***	-1.808***	-1.853***	-2.209***	-2.21***
	(0.049)	(0.049)	(0.036)	(0.038)	(0.064)	(0.063)	(0.047)	(0.048)
ShiftHour6	-2.117***	-2.078***	-2.381***	-2.381***	-1.785***	-1.793***	-2.16***	-2.161***
	(0.049)	(0.049)	(0.037)	(0.038)	(0.065)	(0.064)	(0.047)	(0.048)
ShiftHour7	-2.827***	-2.784***	-3.074***	-3.074***	-2.073***	-2.05***	-2.415***	-2.415***
	(0.049)	(0.048)	(0.037)	(0.038)	(0.065)	(0.064)	(0.047)	(0.048)
ShiftHour8					-2.736***	-2.699***	-3.037***	-3.037***
					(0.064)	(0.061)	(0.047)	(0.048)
Workload	-0.036***	-0.038***			-0.044***	-0.049***		
	(0.004)	(0.004)			(0.006)	(0.006)		
Reassess	-0.049***				-0.066***			
	(0.009)				(0.012)			
HandoverTaken	-0.072***				-0.059***			
	(0.007)				(0.011)			
Handover	-0.08**				-0.023			
	(0.029)				(0.032)			
WaitRoomCensus	0.011***				0.015***			
	(0.002)				(0.002)			
TreatRoomCensus	-0.019***				-0.008***			
	(0.002)				(0.002)			
Boarder	0.016***				-0.002			
	(0.003)				(0.004)			
Weekend	0.035				0.076*			
	(0.023)				(0.03)			
Physician	Y	Y	Y	N	Y	Y	Y	N
ShiftID	Y	Y	Y	Y	Y	Y	Y	Y
R^2	0.478	0.465	0.461	0.414	0.506	0.495	0.49	0.45
Observations	11,548	11,548	11,548	11,548	6,568	6,568	6,568	6,568

Notes. Robust standard errors are shown in the parentheses. ***p<0.001; **p<0.01; *p<0.05
The coefficients for the physician and shift ID fixed effect are not shown due to space limitation.

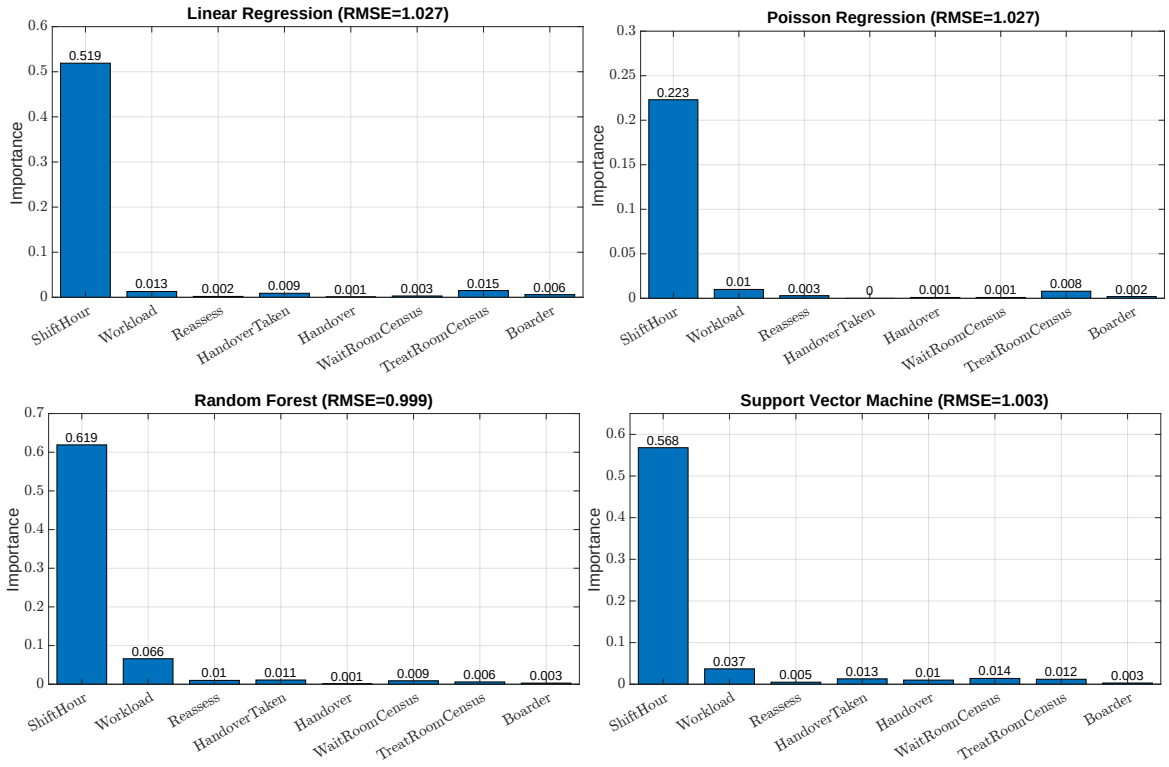
We conclude that the hour of shift is the most important factor in explaining the variations in PPH. The *ShiftHour* captures the temporal variations in physician productivity, which may be attributed to fatigue, task switching, and interruptions. The variable *Workload* seems to have a very marginal impact on PPH. However, one may not reach such a conclusion based on our result as this is not a causal analysis. In fact, the correlation coefficient between *ShiftHour* and *Workload* is over 0.4, suggesting that the physician's workload

accumulates over the course of the shift. As a result, the cognitive load of managing several patients increases, which may prevent the physician from admitting more new patients. Furthermore, physicians need to spend time reassessing their existing patients whose test results are ready, which also impacts PPH.

5.3. Importance of Factors in Predicting PPH

In this section, we explore the importance of these factors from a prediction perspective. By understanding which factors most significantly affect PPH, we can better predict physician productivity and potentially implement targeted interventions to improve efficiency. Hereafter, we use factor and feature interchangeably.

Figure 7 **The feature importance of variables using four models to predict the PPH for 7-hour shifts and their corresponding root mean squared error (RMSE). The physician, shift ID, and weekend are controlled. The feature importance is calculated by permuting the corresponding variable 50 times.**



We employed the permutation importance method (Fisher et al. 2019) to evaluate the significance of each factor in predicting PPH rates for different models. This method assesses feature importance by measuring the increase in the model's prediction error after permuting the feature values while keeping the other features unchanged. The resulting increase in the prediction error indicates the significance of the feature in predicting the target variable.

In our analysis, we use the rooted mean squared error (RMSE) as the model performance measure. We experiment with several prediction models, including linear regression, Poisson regression, random forest, and support vector machine (SVM). All variables described in Section 5.2.1 are included in the prediction

models. We randomly divide the dataset into training (60%) and testing (40%) datasets. For each model, we first train the model on the training dataset and then predict PPH using the testing dataset to obtain the RMSE. We then randomly permute the feature of interest multiple times on the testing dataset and re-evaluate the RMSE. The importance of the feature is determined by the average increase in the RMSE before and after permuting the feature values.

Figure 7 presents the increase in RMSE when each feature is randomly permuted 50 times while keeping others unchanged. Across all four regression models, it is evident that *ShiftHour* is the most important factor in predicting PPH, while *WaitRoomCensus*, *TreatRoomCensus*, and *Boarder* exhibit only marginal effects. *Workload* appears to be more important than the census measures but less important compared to *ShiftHour*.

6. Impact of Time-Varying Service Rates on ED Modeling

A main insight from Sections 4 and 5 is that a physician's productivity rate is time-varying, and decreases significantly over the course of a shift. Hence, it is important to account for the time-varying nature of physician productivity in the modeling of ED operations. In this section, we study the impact of considering the time-varying service rate in modeling ED operations by simulation.

6.1. A Queueing Model with Time-Varying Service Rates

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 four physicians, the dimension of the state space can easily exceed 35 million (see more details in Appendix C), which makes the model analysis both theoretically and computationally challenging. Hence, we seek dimension reduction techniques to simplify the problem. We aim to identify a model that can balance between details and tractability, model parameters that are easy to estimate, and system performances that match real data.

Motivated by the insights in Sections 4 and 5, 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 the 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. Note that the distribution of the service times is *not* a standard exponential distribution, and the cumulative distribution function of the service time for a patient picked up by physician n at time t is $F_{n,t}(x) = 1 - e^{-\int_t^{t+x} \mu_s(n) ds}$, where $\mu_t(n)$

is the service rate of physician n at t. We further assume that the arrival rate is periodic with a daily cycle, and 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 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 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. 2023). 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 (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 12 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 with the highest service rate at the moment, which usually is the physician who most recently started her shift. 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 is small. Our simulation results confirm this conjecture.

In the following, we refer to our queueing model as an $M(t)/M^{\text{PPH}}(t)/s(t)$ queue for simplicity. However, note that (i) the FCFS service rule when picking up a new patient, (ii) the exhaustive discipline when physicians become off duty, and (iii) the mechanism of routing a patient to the fastest physician when there is more than one idle physician, are all parts of the specifications of our queueing model.

6.2. Simulation Setup

Next, we simulate the $M(t)/M^{\text{PPH}}(t)/s(t)$ queue with parameters estimated using the data from our study ED (referred to as *ED 1*). 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 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.

Most existing studies on ED modeling and physician staffing explicitly or implicitly assume a single-stage physician service with a constant service rate (see, e.g., Ingolfsson et al. 2002, Savage et al. 2015, Wang et al. 2022). To examine the impact of considering the time-varying service rates in ED modeling, 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.

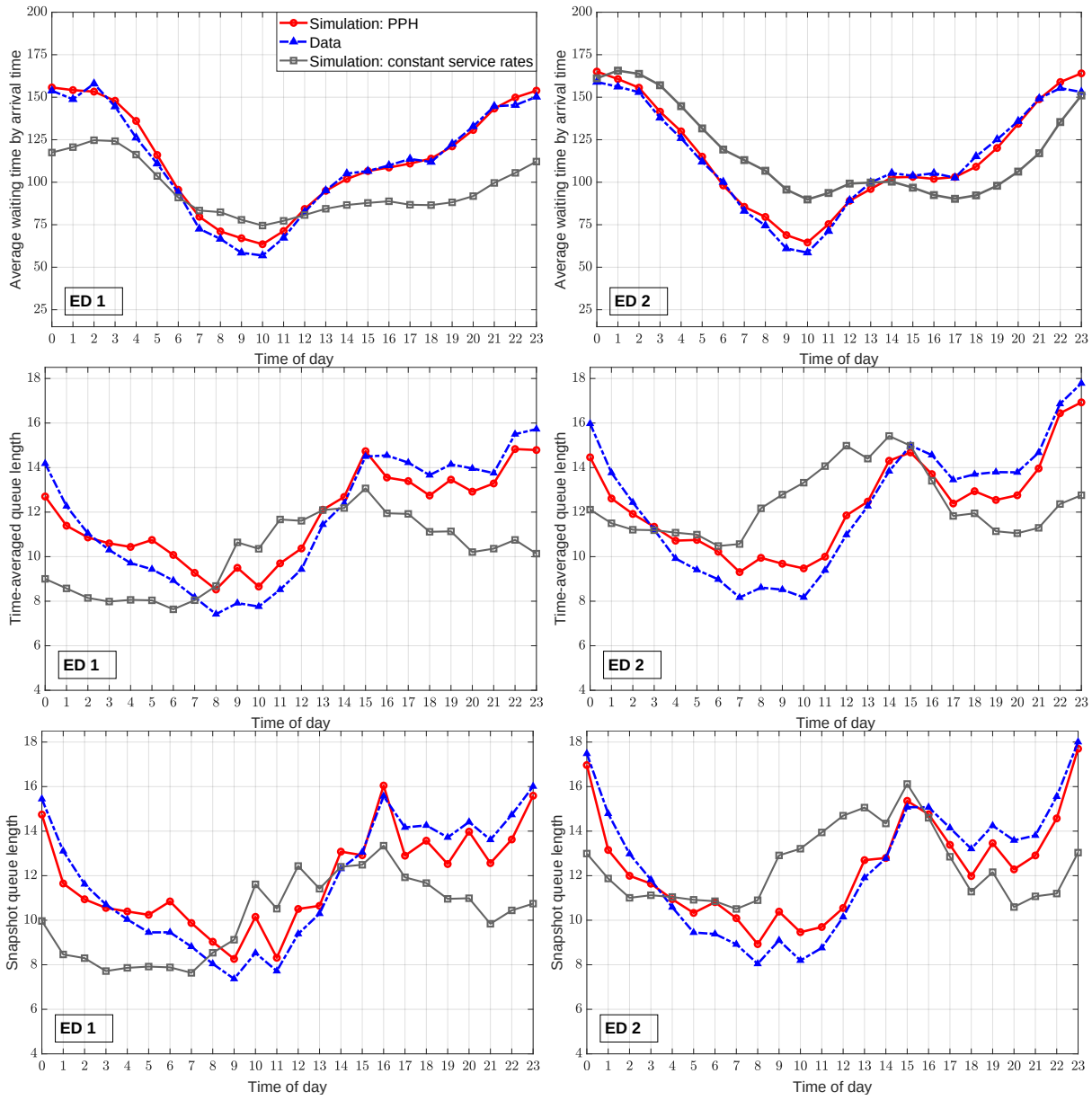
We ran the simulation for 5 replications, each with 500 weeks, and we identified the first 200 weeks as the warm-up period. We focus on waiting room dynamics and choose three time-of-day-dependent performance metrics: (i) the average waiting time of patients who arrived in the same hour of day; (ii) 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 compare the simulated time-of-day-dependent average waiting times and queue lengths with that from 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.

6.3. Simulation Results

The results of comparing the simulated performance measures with the data are shown in Figure 8. We observe that the average waiting times from the simulation when the time-varying service rates (i.e., PPH) are considered nicely match those from the data from both EDs—in terms of the patterns and the magnitudes. 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 8. 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.

In contrast, the simulated average waiting times and queue lengths under constant service rates deviate significantly from the data; see the gray line with squares in Figure 7. 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

Figure 8 **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).**



significantly over the day. However, the outcome is less than satisfactory, potentially due to the fact that the scheduler did not consider the time-varying physician service rates.

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.

Finally, we comment on the parameter estimation of the $M(t)/M^{\text{PPH}}(t)/s(t)$ model. In principle, one simply needs to count the number of arrivals per hour 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%.

7. Impact of Time-Varying Service Rates on Physician Staffing

In this section, we explore the impact of incorporating time-varying service rates in physician staffing decisions. Using data from January to July 2015 for the study ED, we optimize the physician staffing by adjusting the shift start times. First, we apply algorithms that consider the shift-hour-dependent service rates (i.e., PPH). Next, we apply the same algorithm but assume that the service rates are constant over the shift hours. Finally, we use our simulation model to compare the performance of the two approaches. This allows us to quantify the benefits, if any, of incorporating time-varying service rates into the physician staffing optimization.

7.1. 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 adjust the start times of the 13 shifts in the main ED area (referred to as the *baseline schedule* hereafter) to reduce the average patient waiting time.

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) and Zaerpour et al. (2022) 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 $\{0, 1, \cdots, 23\}$, which significantly reduces the computational complexity. 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 patient waiting time under each scenario because reducing waiting times achieves better health outcomes for patients 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) with the uniformization method (discussed below) for the evaluation of each candidate schedule, which takes less than two hours for each scenario.

7.2. Performance Evaluation Through Uniformization

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.

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 purposes, 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 4 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 $301 + 2^4 = 317$, whereas that of the model that considers patient returns explicitly exceeds 35 million, as we discussed at the beginning of Section 6.

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

$$\boldsymbol{\pi}(t_2) = \sum_{n=0}^{\infty} p_j(n) \boldsymbol{\pi}(t_1) P_{1j}^n, \tag{8}$$

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 (8). 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 D.

7.3. Results and Discussion

After solving the optimization problems under time-varying physician service rates and constant service rates, we evaluate the average patient waiting time under the baseline shift schedule and the optimized schedules for the three scenarios by simulation. Hence, a total of seven shift schedules are evaluated. We use simulation instead of uniformization so that we can construct confidence intervals. Moreover, the simulated waiting times fit the data better than that of uniformization.

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.

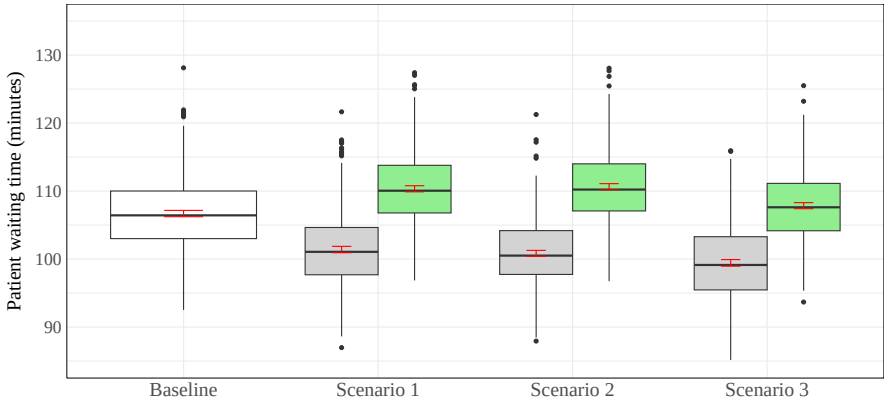
The results are shown in Table 3 and Figure 9. We first observe that when the time-varying service rates are considered, we can achieve a better match between patient demand and ED capacity by adjusting the shift start times. As a result, the average patient waiting time can be reduced by 5.0% to 6.8% 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 9 shows that the reductions over the baseline schedule are statistically significant at the 5% level.

Table 3 The start times of the optimized physician shift schedules. The last two columns show the average patient waiting time, the absolute reduction, and the percentage reduction of the optimized schedules over the baseline schedule for the time-varying service rates and the constant service rates, respectively.

	Shift Start Times													Waiting Time (mins)	
	S1	S2	S3	S4	S5	S7	S8	S9	S10	S12	S13	S14	S15	Average	Reduction (%)
Baseline (from data)	6	7	8	10	10	12	14	16	16	18	20	23	0	106.7	N/A
Time-varying service rates															
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	100.8	5.9 (5.5%)
Scenario 3	8	10	6	12	9	11	15	17	14	18	20	1	22	99.4	7.3 (6.8%)
Constant service rates															
Scenario 1	6	8	7	10	9	11	14	16	15	18	19	23	0	110.3	-3.6 (-3.4%)[†]
Scenario 2	7	8	7	10	9	11	15	16	15	17	20	23	0	110.6	-3.9 (-3.7%)
Scenario 3	8	6	7	10	9	11	14	15	16	20	17	22	0	107.9	-1.2 (-1.1%)

Note. The shift start times that are different from the baseline schedule are highlighted.
[†] The negative reduction represents an increase in waiting time.

Figure 9 Box plots and 95% confidence intervals for the simulated average patient waiting time based on 500 replications. For each scenario, the boxplots in gray represent the waiting time for the optimized staffing under time-varying service rates, whereas the ones in green represent the waiting time for the optimized staffing under constant service rates.



In contrast, when the time-varying service rates for physicians are not considered, more specifically, when we assume constant service rates over a shift, the optimized schedules perform worse than the baseline schedule. The average patient waiting time increases from 1.1% to 4.3% compared to the baseline schedule, which is equivalent to 3.1 to 9.4 hours of waiting for all patients in the ED per day. Figure 9 shows that the increase over the baseline schedule is statistically significant at the 5% level. Hence, we conclude that it is essential to consider the time-varying physician productivity in physician staffing decisions. Otherwise, the optimized schedules may lead to longer waiting times.

Next, we discuss the practical benefits of incorporating the time-varying service rates in physician staffing compared to using constant service rates. Take Scenario 1 for example. Table 3 shows that the average

waiting time is 110.3 minutes with constant service rates, increased by 8.9 minutes from 101.4 minutes, which is the average waiting time under the optimized schedule when the time-varying service rates are considered. Woodworth and Holmes (2020) find that EDs could save the total healthcare cost approximately 2% to 4% by reducing each patient's waiting time by 10 minutes. Based on public data from a government website,[4] the average cost per ED visit in Alberta, Canada was CA$449.2 in 2015–2016. With 57,086 visits to the main ED area per year (156.4 visits/day multiplied by 365 days), this monetary value for the difference in waiting time for our study hospital is from CA$456,446 to CA$912,892 annually. One can calculate the cost differences for Scenarios 2 and 3 in Table 3 in a similar fashion. Hence, we conclude that incorporating the time-varying service rates in physician staffing can generate significant cost savings over the schedules with constant service rates. 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 accounting for the time-varying service rates in physician staffing.

8. Conclusion and Future Research

Existing emergency department (ED) staffing models often assume that physician service rates remain constant over time. This simplifying assumption implies that the ED's productivity, measured by the number of new patients that can be seen per hour, is solely determined by the number of working physicians. Motivated by the intriguing observation of the time-varying pattern in physician productivity (measured by PPH), we challenge this assumption by studying the contributing factors to the time-varying productivity and its impact on ED modeling and physician staffing. Through an optimal control framework, we find that the shift-hour-dependent structure of physician productivity is intrinsic and may be attributed to physician multitasking and overtime/handoff avoidance. Our empirical analysis further confirms that shift hour is the most important factor in explaining the variations in PPH and predicting PPH. Therefore, it is essential to consider this dependency of physician productivity on shift hours when modeling ED operations and optimizing physician staffing.

By overlooking the time-varying nature of physician service rates, the standard ED staffing models may fail to accurately capture the nuances of physician productivity. This can lead to suboptimal physician staffing decisions that do not fully align with the dynamic patient demands and workflow patterns in the ED. Indeed, our study demonstrates that the ED model using a constant rate fails to accurately capture the ED's dynamics, which creates a discrepancy between the expected and actual performance of any staffing plan, undermining the effectiveness of the staffing strategy. Furthermore, our study quantifies the benefit of incorporating the time-varying service rates in physician staffing compared to using constant service rates. Our study hospital

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

can save close to one million dollars annually by considering the time-varying physician productivity in ED physician staffing decisions. Hence, our findings call for immediate attention from hospital management and healthcare planners to take the time-varying nature of physician productivity into the decision-making so as to better match healthcare resources with patient demand.

There are a number of opportunities for future research. First, it would be of interest to extend our approach to study the time-varying nurse productivity. Nurses in the ED play a vital role in monitoring patient status, administering medications, coordinating with other healthcare providers, etc. Hence, increasing nurse productivity is vital to improve the overall ED efficiency (Ding et al. 2024). Second, the decrease in the quality of care for patients admitted in the early stage of the shift due to physician fatigue or higher cognitive load is not captured in our model. It would be interesting to study the optimal patient pick-up strategy when the quality of care and the risk caused by patient handoffs are considered. Third, it is well known that the timing of patient departures in inpatient units has an impact on ED operations (Shi et al. 2015). Hence, the stochastic fluctuations beyond the ED may also contribute to the time-varying physician productivity. It would be interesting to extend our simulation model to account for the patient flow in inpatient units to capture the complete patient journey. 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. Our findings suggest that using PPH as the aggregated service rates, which can be easily derived from data, may be 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.

References

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

R. J. Batt and C. Terwiesch. 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.

H. Bavafa and J. O. Jónasson. The distributional impact of fatigue on performance. *Management Science*, 70(5): 3319–3337, 2024.

J. A. Berry Jaeker and A. L. Tucker. Hurry up and wait: Differential impacts of congestion, bottleneck pressure, and predictability on patient length of stay. *Harvard Business School working paper series# 13-052*, 2012.

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.

T. Brachet, G. David, and A. M. Drechsler. The effect of shift structure on performance. *American Economic Journal: Applied Economics*, 4(2):219–246, 2012.

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.

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

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*, 69(6):1936–1959, 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.

D. D. Cho, K. M. Bretthauer, K. D. Cattani, and A. F. Mills. Behavior aware service staffing. *Production and Operations Management*, 28(5):1285–1304, 2019.

H. Dai, K. L. Milkman, D. A. Hofmann, and B. R. Staats. The impact of time at work and time off from work on rule compliance: the case of hand hygiene in health care. *Journal of Applied Psychology*, 100(3):846, 2015.

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

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.

H. Ding, S. Tushe, D. S. KC, and D. Lee. Valuing nursing productivity in emergency departments. *Manufacturing & Service Operations Management*, 2024.

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.

A. Feizi, A. Carson, J. B. Jaeker, and W. E. Baker. To batch or not to batch? impact of admission batching on emergency department boarding time and physician productivity. *Operations Research*, 71(3):939–957, 2023.

A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177): 1–81, 2019.

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

M. Freeman, S. Robinson, and S. Scholtes. Gatekeeping, fast and slow: An empirical study of referral errors in the emergency department. *Management Science*, 67(7):4209–4232, 2021.

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.

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

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

M. R. Ibanez and M. W. Toffel. How scheduling can bias quality assessment: Evidence from food-safety inspections. *Management Science*, 66(6):2396–2416, 2020.

A. Ingolfsson, M. A. Haque, and A. Umnikov. Accounting for time-varying queueing effects in workforce scheduling. *European Journal of Operational Research*, 139(3):585–597, 2002.

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.

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*, 46: 254–259, 2021.

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

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

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

D. S. KC, S. Scholtes, and C. Terwiesch. Empirical research in healthcare operations: past research, present understanding, and future opportunities. *Manufacturing & Service Operations Management*, 22(1):73–83, 2020.

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

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

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

R. Liu and X. Xie. Weekly scheduling of emergency department physicians to cope with time-varying demand. *IISE Transactions*, 53(10):1109–1123, 2021.

L. X. Lu and S. F. Lu. Distance, quality, or relationship? interhospital transfer of heart attack patients. *Production and operations management*, 27(12):2251–2269, 2018.

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

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, 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.

Z. Wang, R. Liu, and Z. Sun. Physician scheduling for emergency departments under time-varying demand and patient return. *IEEE Transactions on Automation Science and Engineering*, 20(1):553–570, 2022.

W. Whitt and X. Zhang. A data-driven model of an emergency department. *Operations Research for Health Care*, 12: 1–15, 2017.

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*, 31(2):645–667, 2022.

Appendices

Appendix A. Further results on data analysis

Figure 10 **The average new patients seen per hour (PPH) for 8-hour shifts in the main ED area with a time resolution of 1 hour. The extra point on the curve for physician workload outside the shift duration is due to physician overtime for one hour.**

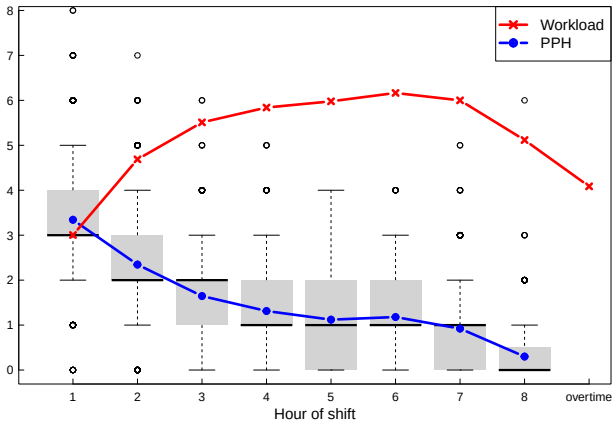
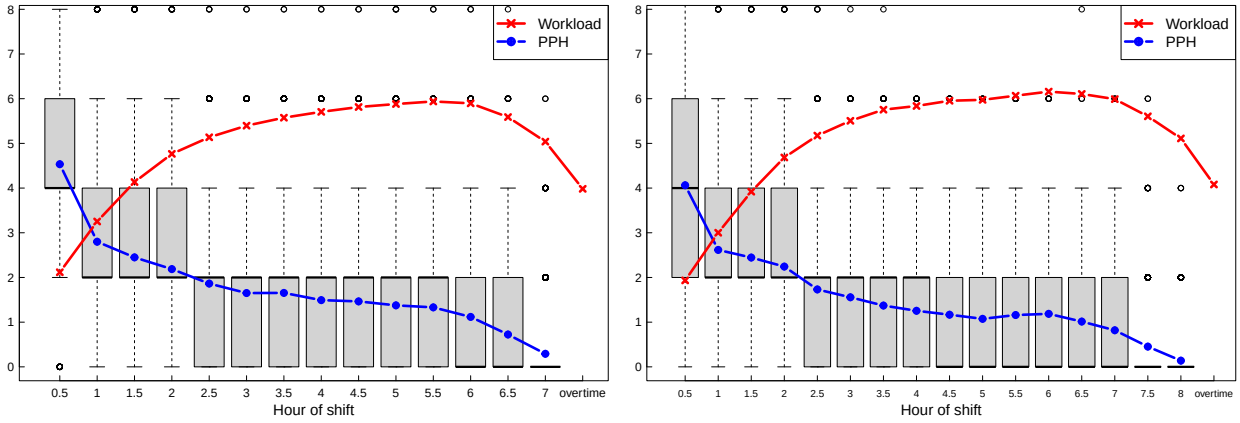


Figure 11 **The average new patients seen per hour (PPH) for 7- and 8-hour shifts with a time resolution of 30 minutes. The extra point for physician workload outside the shift duration is due to physician overtime for one hour.**



Appendix B. Solutions for the optimal control problem (1)

In this section, we study the optimal control problem defined in (1) by considering three cases: (i) $R(0) = R_0 = 0, D(0) = D_0, 0 \le D_0 \le \mu_R/\theta$; (ii) $R(0) = R_0 > 0$; (iii) $R(0) = R_0 = 0, D(0) = D_0 > \mu_R/\theta$. Theorem 1, Propositions 1, 2, and 3 present the results for Case (i) since it is the most relevant case. The corresponding

Figure 12 **The percentage of patients from each triage level in the main ED area by the time of day.**

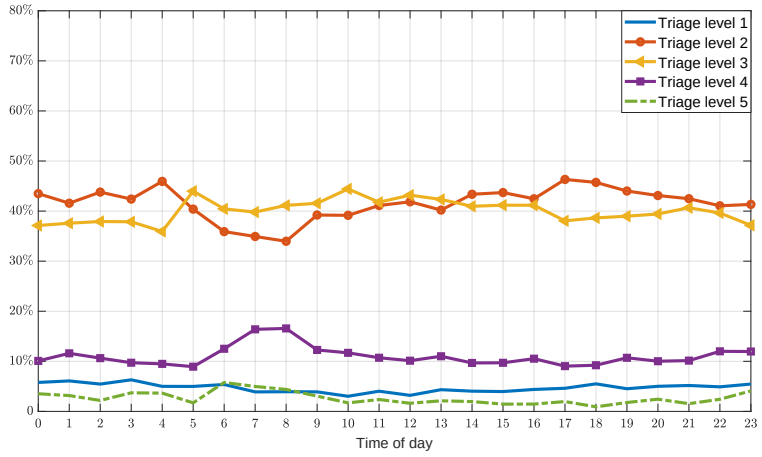
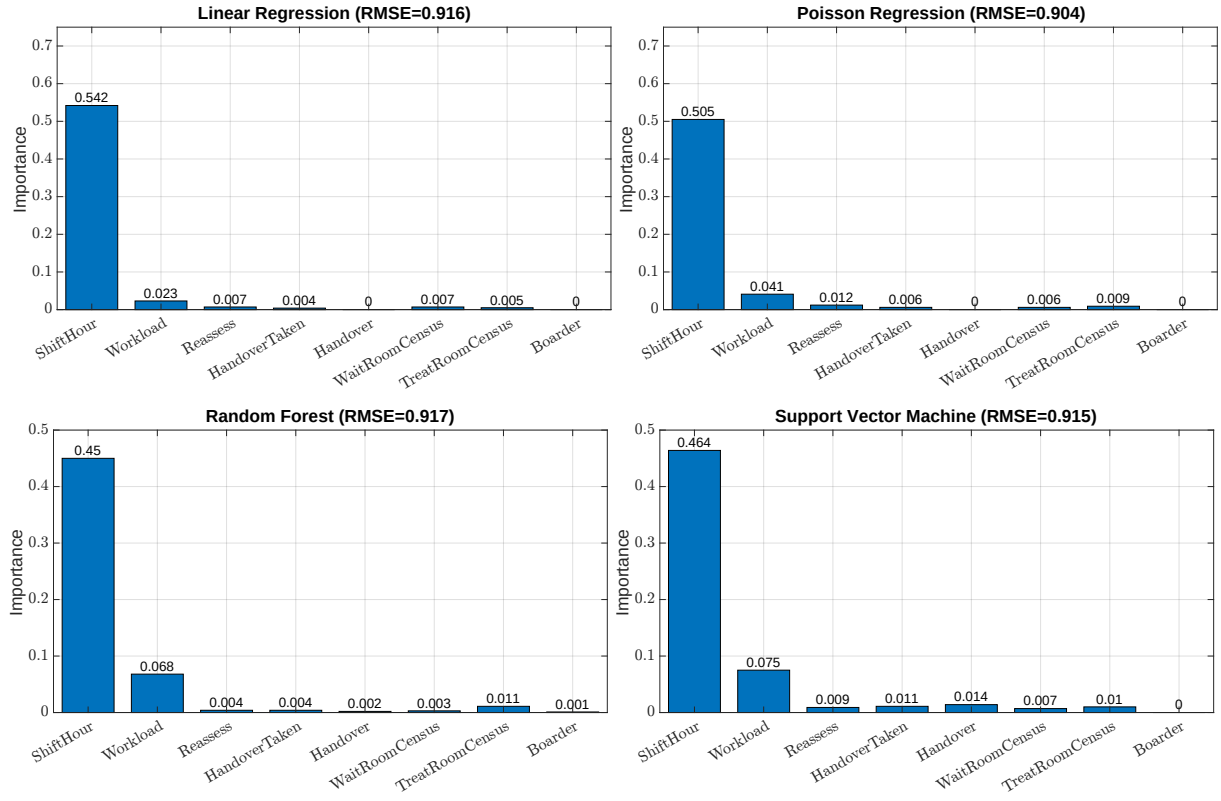


Figure 13 **The feature importance of variables using four models to predict the PPH for 8-hour shifts, and their corresponding root mean squared error (RMSE), The feature importance is calculated by permuting the corresponding variable 50 times.**



proofs are provided as follows. The optimal controls for Cases (ii)&(iii) are similar despite being more complicated, and their proofs are included in Appendix B.2 and Appendix B.3, respectively.

Appendix B.1. Proofs of Theorem 1 , Propositions 1, 2 , and 3 (Case (i)).

In this case, we assume $R(0) = R_0 = 0, D(0) = D_0, 0 \le D_0 \le \mu_R/\theta$. The implications of these conditions are that when the focal physician starts her shift at $t = 0$, (i) there is no patient waiting for reassessment; (ii) the number of patients who are handed off from other physicians to the focal physician and who are undergoing test is sufficiently low so that the focal physician has capacity to sign up new patients.

We first show that $R'(t) = 0,\ \forall t \in [0,T]$. When $t = 0$, because $D_0 \le \theta/\mu_R$, the constraint on $\alpha_R(t)$ in Problem (1) implies that $\alpha_R(0) = \theta D_0/\mu_R$. Hence, $R'(0) = \theta D_0 - \alpha_R\mu_R = 0$. For any $t > 0$ and $\theta D(t)/\mu_R < 1$, we have $\alpha_R(t) = \theta D(t)/\mu_R$ and hence $R'(t) = \theta D(t) - \alpha_R(t)\mu_R = \mu_R(\theta D(t)/\mu_R - \alpha_R(t)) = 0$. When $t > 0$ and $\theta D(t)/\mu_R = 1$, we have $\alpha_R(t) = 1$ and $\alpha_N(t) = 0$, which yields $R'(t) = \theta D(t) - \mu_R = 0$. Furthermore, $D'(t) = p\mu_R - \theta D(t) = -(1-p)\mu_R < 0$, which means that when $D(t)$ becomes sufficiently large such that $\theta D(t)/\mu_R = 1$, the derivative of $D(t)$ becomes negative and $D(t)$ starts to decrease. As a result, $\theta D(t+\delta t)/\mu_R < 1$ for any infinitesimally δt. Because $\theta D_0/\mu_R \le 1$, we conclude that $\theta D(t)/\mu_R \le 1$, $\alpha_R^*(t) = \theta D(t)/\mu_R$, and $R'(t) = 0,\ \forall t \in [0,T]$. Combining with $R(0) = 0$, we conclude that $R(t) = 0,\ \forall t \in [0,T]$.

Furthermore, whenever $D(t) = 0$, we have $D'(t) = p\alpha_N(t)\mu_N \ge 0$. Combining with $D(0) = D_0 \ge 0$, 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:

$$\begin{aligned} &\max_{\alpha_N(t)} \left\{ \int_0^T (1-p)[\alpha_N(t)\mu_N + \theta D(t)]\,\mathrm{d}t - h(D(T)) \right\} \\ &\text{s.t.} \quad D'(t) = p\alpha_N(t)\mu_N - (1-p)\theta D(t),\ D(0) = D_0 \ge 0,\ 0 \le \alpha_N(t) \le 1 - \theta D(t)/\mu_R. \end{aligned} \tag{9}$$

Next, we apply Pontryagin's maximum principle to Problem (9). Denote the co-state variable of $D(t)$ by $\lambda_D(t)$. The Hamiltonian is

$$\begin{aligned} H(D,\alpha_N,\lambda_D(t),t) &= (1-p)[\alpha_N(t)\mu_N + \theta D(t)] + \lambda_D(t)\left[p\mu_N\alpha_N(t) - (1-p)\theta D(t)\right] \\ &= (1-p)\theta(1-\lambda_D(t))D(t) + \mu_N\left(p\lambda_D(t) + 1 - p\right)\alpha_N(t). \end{aligned} \tag{10}$$

Note that the Hamiltonian (10) is linear in $\alpha_N(t)$. The Pontryagin's maximum principle requires the Hamiltonian to be maximized for all $t \in [0,T]$. Hence, the optimal policy to Problem (9) 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 constraints (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

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

The optimal policy to Problem (9) needs to satisfy the conditions below by Pontryagin's maximum principle:

(i) Maximum Conditions:

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

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

(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 (9) does not contain pure-state constraints, the co-state variable $\lambda_D(t)$ is continuous in t under optimality. Consider t at which $1 - \theta D(t)/\mu_R > 0$ and $\alpha_N(t) = 0$. Next, we prove that $\alpha_N(\hat{t}) = 0$ for any $\hat{t} \geq t$ under the optimal policy, which imply the optimal control $\alpha_N^*(t)$ is of threshold type.

Because of the complementary slackness and the constraint that $1 - \theta D(t)/\mu_R - \alpha_N(t) > 0$, we have $\mu_U = 0$. The adjoint equation for $\lambda_D'(t)$ becomes

$$\lambda_D'(t) = -(1-p)\theta(1-\lambda_D(t)) = (1-p)\theta\lambda_D(t) - (1-p)\theta. \tag{12}$$

Solving the ordinary differential equation in (12) yields

$$\lambda_D(t) = Ce^{(1-p)\theta t} + 1, \tag{13}$$

where C is a constant. If $C \geq 0$, then $\lambda_D(t) \geq 1$, and hence $p\lambda_D + 1 - p = p(\lambda_D - 1) + 1 > 0$. As a result, the maximum conditions imply that $\alpha_N(t) = 1 - \theta D(t)/\mu_R$, which contradicts with $\alpha_N(t) = 0$. Hence, $C < 0$. Plugging in the expression of $\lambda_D(t)$ in (13) into (12) yields

$$\lambda_D'(t) = (1-p)\theta Ce^{(1-p)\theta t} < 0,$$

whenever $\alpha_N(t) = 0$. This implies that once $\alpha_N(t) = 0$, then

$$p\lambda_D(\hat{t}) + 1 - p < p\lambda_D(t) + 1 - p < 0, \forall \hat{t} \in [t, T]$$

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.

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:

$$\frac{\mathrm{d}D(t)}{\mathrm{d}t} = p\left(\mu_N - \frac{\mu_N}{\mu_R}\theta D(t)\right) - (1-p)\theta D(t),\ D(0) = D_0,\ t \in [0, t^*],\ \textit{and} \tag{14}$$

$$\frac{\mathrm{d}D(t)}{\mathrm{d}t} = -(1-p)\theta D(t),\ t \in (t^*, T]. \tag{15}$$

Solving the ordinary differential equations in (14) and (15) yields:

$$D(t) = \left(D_0 - \frac{p\mu_N}{\theta\left(1 - p + p\mu_N/\mu_R\right)}\right)e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t} + \frac{p\mu_N}{\theta\left(1 - p + p\mu_N/\mu_R\right)},\ t \in [0, t^*],\ \textit{and} \tag{16}$$

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

which completes the proof for Theorem 1. □

Proof of Proposition 1. Solving (13) together with the boundary condition in the adjoint conditions yields

$$\lambda_D(t) = 1 - [1 + h'(D(T))]\, e^{(1-p)\theta(t-T)},\ t \in [0, T]. \tag{18}$$

Let t^* denote the optimal threshold under the optimal control policy. The maximum conditions imply that $p\lambda_D(t^*) + 1 - p = 0$. Plugging (18) into this equation and solving it yield

$$t^* = T - \frac{\ln[p(1 + h'(D(T)))]}{(1-p)\theta}. \tag{19}$$

Note that the right-hand side of (19) is not necessarily between 0 and T. Specifically, when $p \le [1 + h'(D(T))]^{-1}$, we have $\ln[p(1 + h'(D(T)))] \le 0$. Hence,

$$T - \frac{\ln[p(1 + h'(D(T)))]}{(1-p)\theta} \ge T.$$

On the other hand, when $h'(D(T) \ge e^{T(1-p)\theta}/p - 1$, we have

$$T - \frac{\ln[p(1 + h'(D(T)))]}{(1-p)\theta} \le 0.$$

Hence, we let $t^* = 0$ if the right-hand side of (19) is less than 0, and let $t^* = T$ if the right-hand side of (19) is greater than T. Hence, we get the expression of t^*. Next, we prove the monotonicity of t^* with respect to θ and p.

When $t^* = 0$ or $t^* = T$, it is obvious that t^* does not change with θ or p. Next, we consider the case $0 < t^* < T$, which implies that $p > [1 + h'(D(T))]^{-1}$. Take the derivatives of t^* with respect to θ and p, respectively, we get

$$\frac{dt^*}{d\theta} = \frac{\ln[p(1 + h'(D(T)))]}{(1-p)\theta^2} > 0,$$
$$\frac{dt^*}{dp} = -\frac{(1-p)/p + \ln[p(1 + h'(D(T)))]}{(1-p)^2\theta} < 0.$$

which completes the proof for Proposition 1. □

Proof of Proposition 2. We first take the derivatives of α_N with respect to μ_N and μ_N:

$$\frac{d\alpha_N}{d\mu_N} = -\frac{p(1-p)\mu_R}{[p\mu_N + (1-p)\mu_R]^2}\left(1 - e^{-\theta(1-p+p\mu_N/\mu_R)t}\right) - \frac{\theta pt}{\mu_R}\left(\frac{p\mu_N\mu_R}{p\mu_N + (1-p)\mu_R} - \theta D_0\right)e^{-\theta(1-p+p\mu_N/\mu_R)t} < 0,$$
$$\frac{d\alpha_N}{d\mu_R} = \frac{\theta}{\mu_R}\left(\frac{D(t)}{\mu_R} - \frac{dD(t)}{d\mu_R}\right) = \frac{D_0}{\mu_R}e^{-\theta(1-p+p\mu_N/\mu_R)t} + \frac{p(1-p)\mu_N\mu_R}{\theta[p\mu_N + (1-p)\mu_R]^2}\left(1 - e^{-\theta(1-p+p\mu_N/\mu_R)t}\right)$$
$$+ \frac{p\mu_N t}{\mu_R^2}\left(\frac{p\mu_N\mu_R}{p\mu_N + (1-p)\mu_R} - \theta D_0\right)e^{-\theta(1-p+p\mu_N/\mu_R)t} > 0,$$

and both inequalities hold because of $D_0 \le \frac{p\mu_N\mu_R}{\theta[p\mu_N+(1-p)\mu_R]}$. Next, taking the derivatives of α_N with respect to θ yields

$$\begin{aligned}\frac{d\alpha_N}{d\theta} &= -\frac{D_0}{\mu_R}e^{-\theta(1-p+p\mu_N/\mu_R)t} + \left(\frac{\theta D_0}{\mu_R} - \frac{p\mu_N\mu_R}{p\mu_N+(1-p)\mu_R}\right)(1-p+p\mu_N/\mu_R)te^{-\theta(1-p+p\mu_N/\mu_R)t} \\ &= -\frac{1}{\mu_R}e^{-\theta(1-p+p\mu_N/\mu_R)t}\left[D_0 + \frac{p\mu_N+(1-p)\mu_R}{\mu_R}\left(\frac{p\mu_N\mu_R}{p\mu_N+(1-p)\mu_R} - \theta D_0\right)t\right] < 0,\end{aligned}$$

and the inequality holds because of $D_0 \le \frac{p\mu_N\mu_R}{\theta[p\mu_N+(1-p)\mu_R]}$. □

Proof of Proposition 3. Because $\mathrm{PPH}(t) = \alpha_N(t)\mu_N$, we get

$$\mathrm{PPH}(t) = \frac{(1-p)\mu_N\mu_R}{p\mu_N+(1-p)\mu_R} + \mu_N\left(\frac{p\mu_N}{p\mu_N+(1-p)\mu_R} - \frac{\theta D_0}{\mu_R}\right)e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t},\ \forall t\in[0,t^*].$$

It is obvious that $\mathrm{PPH}(t) = 0,\ \forall t \in (t^*, T]$, which completes the proof. □

Appendix B.2. The structure of the optimal policy under Case (ii).

In this case, we assume $R(0) = R_0 > 0, D(0) = D_0 \ge 0$. The implication is that when the focal physician starts her shift at $t = 0$, there are return patients waiting for reassessment. The system dynamics can be described by the ordinary differential equation as follows:

$$R'(t) = -\alpha_R\mu_R + \theta D(t),\ R(0) = R_0 > 0, \tag{20}$$

$$D'(t) = p\left[\alpha_N\mu_N + \alpha_R\mu_R\right] - \theta D(t),\ D(0) = D_0. \tag{21}$$

Since we assume that return patients are prioritized over new patients, the physician will spend 100% of her efforts to return patients, i.e., $\alpha_R(t) = 1$ and $\alpha_N(t) = 0$, until $D(t) = \mu_R/\theta$. Solving (20) and (21) jointly, we obtain the following:

$$D(t) = \left(D_0 - \frac{p\mu_R}{\theta}\right)e^{-\theta t} + \frac{p\mu_R}{\theta},\ D(t) \ge 0, \tag{22}$$

$$R(t) = R_0 - (1-p)\mu_R t + \left(D_0 - \frac{p\mu_R}{\theta}\right)\left(1-e^{-\theta t}\right),\ R(t) \ge 0. \tag{23}$$

The next lemma shows that $R(t) = 0$ has a unique solution.

Lemma 1. *There exists $t_R > 0$ such that $R(t_R) = 0$ and $D(t_R) \le \mu_R/\theta$.*

Proof of Lemma 1. We first re-write $R(t)$ as

$$R(t) = D_0 + R_0 - (1-p)\mu_R t - \frac{p\mu_R}{\theta} - \left(D_0 - \frac{p\mu_R}{\theta}\right)e^{-\theta t} \equiv R_1(t) - D(t), \tag{24}$$

where $R_1(t) = D_0 + R_0 - (1-p)\mu_R t$. It is clear that (i) $R_1(t)$ is a decreasing function in t; (ii) $R_1(0) = D_0 + R_0 > D(0) = D_0$; and (iii) $\lim_{t\to\infty} R_1(t) = -\infty < \lim_{t\to\infty} D(t) = p\mu_R/\theta$. Both $R_1(t)$ and $D(t)$ are monotonic functions of t. Hence, they have a unique intersection, i.e., $R(t) = 0$ has a unique solution.

Let t_R denote the solution to $R(t) = 0$, i.e., $R_1(t_R) = D(t_R)$. Next, we consider the following three cases separately to prove that $D(t_R) < \mu_R/\theta$: (i) $D_0 < p\mu_R/\theta$, (ii) $p\mu_R/\theta \le D_0 < \mu_R/\theta$, and (iii) $D_0 \ge \mu_R/\theta$. When $D_0 < p\mu_R/\theta$, it is obvious that $D(t) < p\mu_R/\theta \le \mu_R/\theta$ holds for any $t > 0$. Hence, $D(t_R) < \mu_R/\theta$. When $p\mu_R/\theta \le D_0 < \mu_R/\theta$, it is clear that $D(t)$ decreases in t and $D(t) \le D(0) = D_0 < \mu_R/\theta$. At last, we consider the case $D_0 \ge \mu_R/\theta$. Since $D(t)$ is a continuous function on the closed interval $[0, t_R]$ and differentiable on $(0, t_R)$, the Lagrange mean value theorem implies that there exists $t_c \in (0, t_R)$ such that

$$D'(t_c) = \frac{D(t_R) - D(0)}{t_R - 0} = \frac{D(t_R) - D_0}{t_R}. \tag{25}$$

Hence, we have

$$R_1'(t_R) = \frac{R_1(t_R) - R_1(0)}{t_R - 0} = \frac{D(t_R) - D_0 - R_0}{t_R} \le D'(t_c) \le D'(t_R), \tag{26}$$

where the first equality holds because $R_1(t)$ is a linear function of t, the second equality holds because $R_1(t_R) = D(t_R)$, the first inequality holds because $R_0 > 0$, and the second inequality holds because $D'(t)$ increases in t due to $D''(t) = (D_0 - p\mu_R/\theta)\theta^2 e^{-\theta t} \ge 0$. As a result, $R'(t_R) = R_1'(t_R) - D'(t_R) \le 0$. We know from (22) that $R'(t) = -\mu_R + \theta D(t)$. Hence, $-\mu_R + \theta D(t_R) \le 0$, and we have $D(t_R) \le \mu_R/\theta$, which completes the proof. □

Lemma 1 implies that when the return patients handed off to the focal physician is exhausted at t_R, the number of patients in the test queue, i.e., $D(t_R)$, is less than μ_R/θ. Hence, the system dynamics after t_R is again the same as that under Case (i) in Appendix B.1. Therefore, the structure of the optimal policy is again of threshold type. Since the analysis is the same as that in Case (i), the details are omitted.

Appendix B.3. The structure of the optimal policy under Case (iii).

In this case, we assume $R(0) = R_0 = 0, D(0) = D_0 > \mu_R/\theta$. The implications of these conditions are that when the focal physician starts her shift at $t = 0$, (i) there is no patient waiting for reassessment; (ii) the number of patients who are handed off from other physicians to the focal physician and who are undergoing test is sufficiently high so that the focal physician has no capacity to sign up new patients.

Because $R(0) = 0$ and $D_0 > \mu_R/\theta$, we have $\alpha_R(0) = \min\{1, \theta D_0/\mu_R\} = 1$ and $\alpha_N(0) = 0$. In fact, as long as $D(t) > \mu_R/\theta$, we have $\alpha_R(t) = 1$ and $\alpha_N(t) = 0$. Hence, the system dynamics when $D(t) > \mu_R/\theta$ can be described by the ordinary differential equation as follows:

$$D'(t) = p\mu_R - \theta D(t) = \mu_R(p - \theta D(t)/\mu_R) < 0, \tag{27}$$

$$R'(t) = -\mu_R + \theta D(t),\ R(0) = 0, \tag{28}$$

which implies that $D(t)$ decreases until $D(t) = \mu_R/\theta$. Solving (27) together with the initial condition $D(0) = D_0$, we have

$$D(t) = \left(D_0 - \frac{p\mu_R}{\theta}\right) e^{-\theta t} + \frac{p\mu_R}{\theta}. \tag{29}$$

Let $D(t) = \mu_R/\theta$, we have

$$t_0 = \frac{1}{\theta} \ln \frac{\theta D_0 - p\mu_R}{(1-p)\mu_R}, \tag{30}$$

where t_0 is the time it takes for $D(t)$ to decrease to μ_R/θ. Because $D(t) \geq \mu_R/\theta$ for $t \in [0, t_0]$, we know from (28) that $R(t_0) > 0$. The dynamic of the system after t_0 will be the same as that under Case (ii) with the initial system state $D_0 = \mu_R/\theta,\ R_0 = R(t_0) > 0$. Therefore, the structure of the optimal policy is again of threshold type. Since the analysis is the same as that in Case (ii), the details are omitted.

Appendix C. Dimensionality of a Markovian ED Model

In this section, we use a CTMC to model the system dynamics of ED patient flow, and the system state is represented by a vector $(x_w, x_T^{(1)}, x_R^{(1)}, x_T^{(2)}, x_R^{(2)}, \ldots, x_T^{(k)}, x_R^{(k)})$, where k is the number of physicians on duty, x_w is the number of patients waiting to be seen in the waiting room, $x_T^{(i)}$ is the number of patients going through tests, and $x_R^{(i)}$ is the number of patients waiting for reassessment for Physician i, $i = 1, \cdots, k$. Assume that there are at most N_w patients in the waiting room, and at most $M^{(i)}$ patients in the test and reassess queues of Physician i. Hence, the state space is

$$\left\{(x_w, x_T^{(1)}, x_R^{(1)}, x_T^{(2)}, x_R^{(2)}, \ldots, x_T^{(k)}, x_R^{(k)}) : 0 \leq x_w \leq N_w, x_T^{(i)} \geq 0, x_R^{(i)} \geq 0, x_T^{(i)} + x_R^{(i)} \leq M^{(i)}, i = 1, \cdots, k\right\}.$$

Hence, the state space has a dimension $(N_w + 1)\Pi_{i=1}^{k}[(M^{(i)} + 1)(M^{(i)} + 2)/2]$. Assuming $k = 4, N_w = 20$, and $M^{(i)} = 7$ for all i, then the dimension of the state space is $21 \times [(7+1)(7+2)/2]^4 = 35,271,936$. The modeling and computation are similar to that of model $\mathcal{S}$ in Campello et al. (2016) despite the differences in the patient pick-up mechanism.

Appendix D. 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:

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

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 identity matrix of the same dimension as $\boldsymbol{\pi}^T\boldsymbol{\pi}$; 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 $x_1 < \eta_j$, i.e., the number of waiting patients is less than the number of physicians who begin their shifts, we assign the patients in waiting to newly-arrived physicians with the highest service rates, following the same rule described in Section 6.1. This is achieved by ranking the newly-arrived physicians by their current service rates. Specifically, in $\mathbf{y}_2 = (\mathbf{y}_1 Q, \mathbf{1}_{x_1}, \mathbf{0}_{\eta_j - x_1})$, $\mathbf{1}_{x_1}$ represents the x_1 physicians with the highest service rates among the η_j newly-arrived physicians and $\mathbf{0}_{\eta_j - x_1}$ represents the remaining $\eta_j - x_1$ slower physicians. 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 follow Liu and Xie (2021) to compute the average patient waiting time. The total expected patient waiting time in the jth period $((j-1)l, jl]$, denoted by W_j, can be expressed 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 $\sum_{j=1}^{24/l} W_j / \sum_{j=1}^{24/l} \lambda_j$. Interested readers are referred to the appendix of Liu and Xie (2021) for detailed derivation and explanation.