Submitted to *Manufacturing & Service Operations Management*
manuscript

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

Zhankun Sun
Department of Decision Analytics and Operations, College of Business, City University of Hong Kong, Kowloon, Hong Kong

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

Huiyin Ouyang
Faculty of Business and Economics, The University of Hong Kong, Pok Fu Lam Road, Hong Kong
Corresponding author: oyhy@hku.hk

Problem definition: Physician productivity (measured by new patients seen per hour) in emergency departments (EDs) exhibits a distinct time-varying pattern. We examine the factors contributing to this phenomenon empirically and analytically, and investigate the implications of incorporating time-varying service rates into ED modeling and physician staffing. ***Methodology/results:*** Using data from a Canadian ED, we provide empirical evidence that the "shift hour" (time elapsed since the shift started) is the most significant predictor of physician productivity. We then model the new patient pickup decisions using an optimal control framework. Assuming physicians' objective is to maximize throughput while avoiding handoffs or overtime, the optimal policy implies adjusting pickup rates throughout a shift is a rational decision. 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 a multi-server queueing model with shift-hour-dependent service rates, despite its lower dimensionality, can accurately capture time-of-day-dependent patient waiting times, whereas models assuming constant service rates deviate significantly from observed data. Furthermore, a case study reveals that staffing plans optimized for time-varying rates outperform current practices, leading to substantial cost savings. ***Managerial implications:*** Our findings underscore the necessity of accounting for the time-varying nature of physician productivity. Hospital administrators and schedulers should incorporate this shift-hour dependency into staffing decisions to enhance resource allocation and ED operational efficiency.

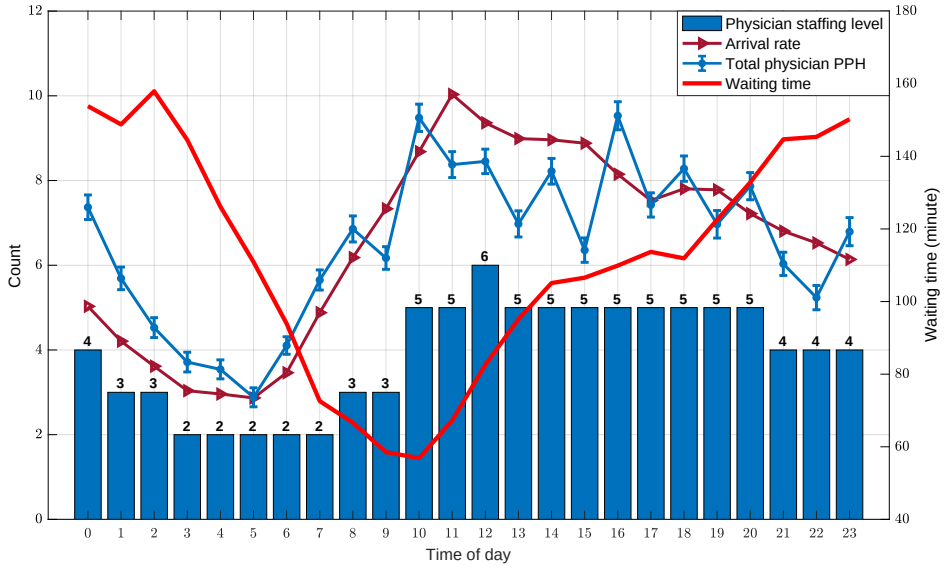
Key words: Healthcare, Emergency Department, Time-Varying Service Rate, Physician Staffing

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). As a result, extended patient waiting times have become extremely common in many healthcare systems around the world. To mitigate the resulting backlog and extended wait times, one must first understand the mechanics of ED flow. Since ED operations are naturally modeled as queueing systems, valid performance analysis requires a rigorous understanding of the stochastic nature of patient arrivals and service durations.

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 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 achieved by staggering shifts of different lengths; see a detailed description of the shift structures in Section 3. However, the outcome is less than satisfactory, as the average waiting time varies substantially over the course of the day and exceeds two hours at times.

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.**

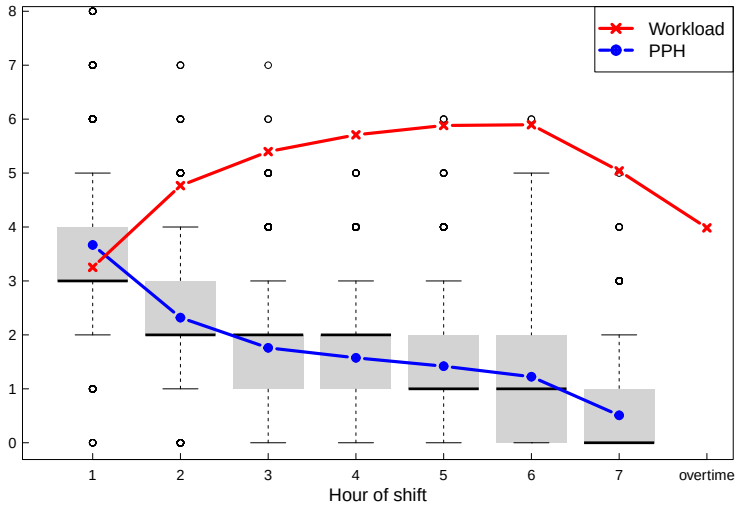


A key determinant of patient waiting time is physicians' speed in picking up new patients, measured by PPH—the number of new patients picked up by a physician per hour. Note that PPH is *not* the rate at which patients complete treatment in the ED, which is usually referred to as *throughput*; rather, 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, for example, Joseph et al. (2018, 2021), Zaerpour et al. 2022, and Niewoehner et al. (2023). Hereafter, we use physician productivity and 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 time of day to Figure 1. An intriguing observation is that the total physician PPH varies considerably 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 ranges from 5.9 to 10.0, a 69% difference. Notably, 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 any given time of day, as evidenced by the narrow 95% confidence intervals. This stability underscores the robustness of the PPH pattern. We also analyzed the total physician PPH separately for weekdays and weekends. Nevertheless, statistical tests show that the differences between the two 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 greater variability in service times increases 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, which is also called the multitasking level (KC 2013).**



Most shifts in our study ED are 7 or 8 hours long. Figure 2 shows the average PPH of individual physicians 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 A1 in Appendix A). Based on the time-varying structure of PPH observed from Figures 2 and A1, 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 pronounced for 7- and 8-hour shifts using half an hour as the time resolution; see Figure A2 in Appendix A. We observe

[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).

a similar pattern when we further plot the PPH for each individual physician or a specific type of shift. Comparable structures have also been reported in the emergency medicine literature using data from U.S. hospitals (Joseph et al. 2018, 2021).

Physician-level PPH dictates the aggregate service speed of the ED, directly influencing system-level performance metrics such as patient waiting times and throughput. Despite its importance, the time-varying nature of physician productivity has not been investigated in depth. Most prior studies analyze ED performance under the assumption of constant service rates, overlooking the dynamic nature of physician productivity throughout a shift; see, e.g., Ingolfsson et al. 2002, Savage et al. 2015, Wang et al. 2022. To address this gap, we aim to identify the factors influencing physician productivity and assess their operational implications. We specifically address two questions: (i) What factors effectively predict time-varying physician PPH? (ii) What is the impact of incorporating time-varying service rates into ED modeling and physician staffing?

To address the research questions, we conduct: (i) an empirical study to establish the timeÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœvarying pattern; (ii) an optimal control model to provide a behavioral explanation for this pattern; and (iii) a simulationÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœbased queueing model—parameterized using the empirical results from (i)—to demonstrate the value of our findings. We believe these complementary approaches together provide an integrated understanding of time-varying physician productivity and its implications for ED operations.

Our study makes the following contributions to the operations management literature on EDs. First, we provide empirical evidence that establishes the time-varying physician productivity. Using detailed operational data from Canadian EDs spanning seven months, we find that physician productivity follows a highly consistent pattern across shifts, physicians, and time periods. Our regression results in Section 3 further show that while physician workload, ED congestion level, and other operational factors all correlate with physician productivity, *shift hour* is the most important predictor of physician productivity.

To further investigate the observed time-varying pattern, we develop an optimal control framework. We model the physicianÃfÂ¢Ã¢â€šÂ¬Ã¢â€žÂ¢s objective as maximizing throughput (completed treatments) while minimizing patient handoffs (transfer of care) or physician overtime. By analyzing the decision to pick up new patients within a finite shift, we derive closed-form expressions for the optimal time-varying PPH. Our results imply that physicians acting to avoid handoffs or overtime would exhibit service rate adjustments similar to those observed in the data. The alignment between the theoretical PPH derived from our model and the empirical data suggests a plausible mechanism for the time-varying structure: (i) the exponential decay during the start-of-shift phase is consistent with the dynamics of physician multitasking, and (ii) the sharp drop in the end-of-shift phase aligns with behavior aimed at handoff and overtime avoidance. Furthermore, our analytical results indicate that allocating additional resources to test centers may not only enhance test

turnaround times but also potentially facilitate higher patient pickup rates by mitigating concerns regarding overtime and handoffs.

Finally, we investigate the practical implications of incorporating the time-varying physician productivity into ED modeling and physician staffing. We model the ED as a multiserver queue with nonstationary Poisson arrivals and exponential service times with time-varying rates. Simulation results demonstrate that this model accurately captures time-of-day-dependent performance metrics, closely matching data from two Canadian EDs. In contrast, simulations ignoring these dynamics (i.e., assuming constant rates) deviate significantly from observed reality. Furthermore, a case study reveals that staffing plans optimized under our time-varying framework outperform current schedules. Conversely, when the time-varying service rate is ignored and a constant service rate is assumed (which is the prevalent practice), the staffing plans generated using the same algorithm perform even worse than the status quo. These results highlight the practical risks of ignoring time-varying productivity in ED modeling and staffing decisions.

The remainder of this paper is organized as follows. We begin with a review of relevant literature in Section 2. Then, Section 3 describes ED operations and presents our exploratory analysis. In Section 4, we develop an optimal control model that provides a behavioral rationale for these observed patterns. We then investigate the impact of time-varying service rates on ED modeling in Section 5 and apply this framework to physician staffing optimization in Section 6. Finally, Section 7 concludes with insights and directions for future research. Proofs and supplementary results are provided 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, Ding et al. 2024), high level of multitasking (KC 2013), and overcrowded systems (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. 2020, Freeman et al. 2016), and routing decisions (Freeman et al. 2021) 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. 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. Chan (2018) finds 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. Shift hour is the most important predictor of PPH. 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 by Niewoehner et al. (2023) finds 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) leverage the time-varying PPH to assign physicians to predetermined shifts. In contrast, our study identifies shift hour as the most important predictor of the time-varying PPH and uses an optimal control framework to understand physicians' rationale. We then investigate the impact of incorporating time-varying PPH in ED modeling and physician staffing. Thus, both studies are relevant to ours but different in their research methodologies and objectives.

Methodology-wise, we use the optimal control framework to model physician decision making in EDs. Hence, works that use fluid models to study healthcare systems are relevant to ours. 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. 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 insights of 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, which can support physician scheduling (as shown by the case study in Section 6).

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.

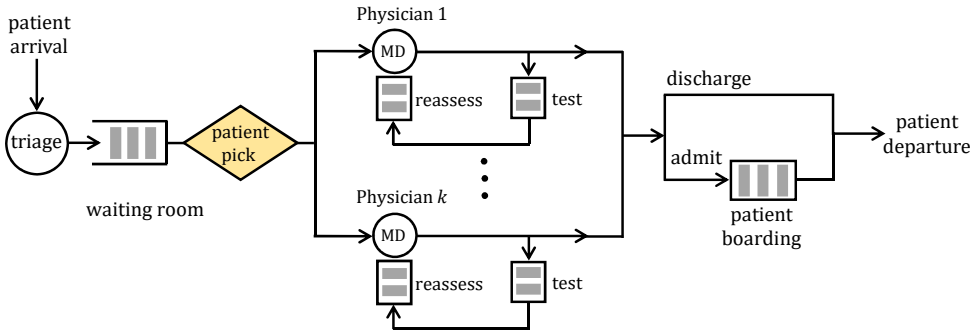
3. Exploratory Study of Physician Productivity

To investigate the factors associated with time-varying productivity, we first establish the context of the ED workflow. We then introduce the data and variables. Finally, we present our empirical results and conclude with a brief discussion on possible mechanisms found in the literature.

3.1. ED Operations and Patient Care Handoff

In this section, we describe the patient flow process in the main area of the study ED, excluding the fast-track area. Our description is based on ED operations in Alberta, Canada. While specific regional differences may exist, we believe the core features, such as the patient return-for-service process, are representative of most EDs. Patient flow in the main ED area is depicted in Figure 3.

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



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. 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 their ED stay. 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.

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. 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). It is suggested that physicians should "slack off" approaching the end of their shift, i.e., stop signing up new patients, to avoid handoff (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.

3.2. Data and Variables

Physicians' decisions to pick up new patients may be influenced by a variety of factors, such as their current active workload, the need to reassess existing patients, ED congestion, and anticipated handoffs. Building on prior descriptive evidence that links productivity to shift hour, this section empirically analyzes these factors to quantify their relative contributions to within-shift productivity variance. Our goal is to identify factors with the greatest explanatory and predictive power for observed variations in productivity patterns, rather than to establish causal relationships between PPH and these factors.

Our analysis is based on patient visit records and physician staffing data from the main area of our study ED, spanning January 3 to July 31, 2015. This period was selected for its stable physician staffing patterns. Figure 4 illustrates the daily physician shifts in our study ED during the study period. Each day includes 15 scheduled shifts: two fast-track shifts and the remainder in the main area. Shift lengths are 6, 7, or 8 hours. Shifts begin at staggered times to align physician capacity with patient demand and to prevent multiple physicians from ending shifts simultaneously, thereby facilitating smoother end-of-shift transitions.

We focus our analysis on the 7- and 8-hour shifts, excluding the less reliable records from 6-hour shifts (flexible shifts) and the two fast-track shifts, which operate in a different area under a different model. The final dataset consists of 2,475 shifts, including 1,654 7-hour and 821 8-hour shifts.

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). Figure 2 and other descriptive results show that productivity varies significantly across shift hours. Therefore, we include a categorical variable, *ShiftHour*, to capture potential heterogeneity across different hours within a shift. We then include a set of control variables to account for multitasking levels, system-wide congestion, physician activities beyond new patient intake, and unobserved heterogeneity. We define *Workload* as the time-averaged number of active patients under the care of the focal physician during the relevant shift hour.

Figure 4 The daily physician shifts in our study ED from January 3 to July 31, 2015.

Shift	0:00	1:00	2:00	3:00	4:00	5:00	6:00	7:00	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
S1							1	2	3	4	5	6	7											
S2								1	2	3	4	5	6	7										
S3									1	2	3	4	5	6	7	8								
S4											1	2	3	4	5	6								
S5											1	2	3	4	5	6	7	8						
S6					Fast-Track Shift →						1	2	3	4	5	6	7							
S7													1	2	3	4	5	6	7	8				
S8															1	2	3	4	5	6	7			
S9																	1	2	3	4	5	6	7	
S10																	1	2	3	4	5	6	7	8
S11													Fast-Track Shift →					1	2	3	4	5	6	7
S12	7																		1	2	3	4	5	6
S13	5	6	7																		1	2	3	4
S14	2	3	4	5	6	7																		1
S15	1	2	3	4	5	6	7																	

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.

To control for congestion levels, we follow the literature (Zaerpour et al. 2022, Li et al. 2023) and include *WaitRoomCensus* (the time-averaged number of patients in the waiting room during the relevant shift hour), *TreatRoomCensus* (the time-averaged number of patients in ED treatment rooms, excluding boarders during the relevant shift hour), and *Boarder* (the time-averaged number of admitted patients who remain in the ED while awaiting an inpatient bed during the relevant shift hour). The *Boarder* variable specifically captures the effect of ED blocking, which has been shown to influence physician behavior (Li et al. 2023).

Physicians may also devote time to activities other than seeing new patients, such as reassessing return patients, handing their patients over to other physicians, or taking handoff patients from others. We include *Reassess*, *Handover*, and *HandoverTaken*, which measure, respectively, the number of reassessments performed, the number of patients handed over to other physicians, and the number of patients received from other physicians during the focal hour.

To account for unobserved and temporal factors, we include several fixed effects. Physician fixed effects capture individual heterogeneity such as age, gender, experience, or baseline practice style, while shift fixed effects control for variation across shift types (e.g., the 06:00ÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Å“13:00 vs. 07:00ÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Å“14:00 shift). We also include a binary indicator, *Weekend*, which equals 1 if the shift occurs on a weekend and 0 otherwise.

Summary statistics for all variables, except physician and shift fixed effects, are provided in Table 1.

3.3. Models and Results

Having defined the variables of interest and described their summary statistics, we next examine how these factors are associated with physician productivity. Specifically, we estimate a series of regression models that quantify the marginal effects of key operational factors while controlling for physician- and shift-level heterogeneity. Our baseline empirical specification is given by

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.

$$PPH_{ist} = \beta_1 ShiftHour_{ist} + \beta_2 Workload_{ist} + \beta_3 Reassess_{ist} + \beta_4 HandoverTaken_{ist} + \beta_5 Handover_{ist} + \beta_6 WaitRoomCensus_{ist} + \beta_7 TreatRoomCensus_{ist} + \beta_8 Boarder_{ist} + \beta_9 Weekend_s + \gamma_i + \delta_s + \epsilon_{ist}, \quad (1)$$

where i indexes physicians, s denotes shifts, and t represents the hour within a shift. The term γ_i captures physician fixed effects and δ_s denotes shift-type fixed effects. The error term ϵ_{ist} is assumed to have mean zero and constant variance.

We begin our empirical analysis by estimating the baseline specification described in Equation (1), which includes all explanatory variables and fixed effects (Model 1). To examine the value of including different covariates in the model specification, we sequentially reÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœestimate the model by omitting, in turn: the physicianÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœlevel activity variables (*Reassess*, *Handover*, and *HandoverTaken*) together with the systemÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœlevel overcrowding measures (*WaitRoomCensus*, *TreatRoomCensus*, and *Boarder*) (Model 2); the workload measure (*Workload*) (Model 3); and the fixed effects associated with individual physicians (Model 4). The sequence of omitted variables follows a backward modelÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœselection approach, whereby variables are removed one at a time from the full specification. Although all these models are estimated, we report only a subset of representative results due to space constraints. Comparing the resulting R^2 values across models allows us to evaluate whether excluding these factors materially affects the explanatory power of the regression. Each set of specifications is estimated separately for 7ÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœhour and 8ÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœhour shifts, and the corresponding results are reported 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. Regarding model fit, removing the physician-level activity variables together with the system-level overcrowding measures results in only a marginal decrease

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

	7-hour-shift models				8-hour-shift models			
	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
(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.

in explanatory power, with the R^2 declining from 0.478 to 0.465 (a 2.7% drop). This indicates that although these ED census and physician activity variables are statistically significant, their combined contribution to explaining variation in *PPH* is relatively small. Likewise, excluding the *Workload* variable further reduces the R^2 only slightly, from 0.465 to 0.461 (less than 1% drop). In contrast, the inclusion of *ShiftHour* factor

and *Shift* fixed effects accounts for 41.4% of the variation in *PPH*. The observations for 8-hour shifts are qualitatively similar, although model fit is somewhat stronger than that of the 7-hour shifts.

The variable *Workload* appears to have a very marginal direct association with *PPH*. However, one may not infer this as evidence of negligible causal impact, since our analysis is correlational in nature. In fact, the correlation coefficient between *ShiftHour* and *Workload* is over 0.4, suggesting that physiciansÃfÂ¢Ã¢â€šÂ¬Ã¢â€žÂ¢ workload tends to accumulate as the shift progresses. Consequently, this intraÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœshift accumulation of workload may thus partially underlie the temporal productivity patterns captured by *ShiftHour*.

Turning to intraÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœshift productivity dynamics, the estimated coefficients on *ShiftHour* reveal systematic temporal variation in *PPH*. For example, for the 7ÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœhour shifts under Model 4, *PPH* declines sharply from Hour 1 to Hour 2 (ÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Å“1.319) and from Hour 2 to Hour 3 (ÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Å“0.553), remains relatively stable during the middle shifts (ÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Å“0.161 from Hour 3 to 4, ÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Å“0.140 from Hour 4 to 5, and ÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Å“0.208 from Hour 5 to 6), and then drops considerably in the final hour (ÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Å“0.693 from Hour 6 to 7).

To complement these inferential results, we further examine the relative importance of the explanatory factors from a prediction perspective (see Appendix B for details). In sum, the regression and prediction results underscore the dominant role of intraÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœshift temporal dynamics, captured by *ShiftHour*, in shaping physician productivity. Specifically, PPH decreases significantly with shift hour over the course of a shift. While our analysis establishes association rather than causality, the results indicate that the time-varying pattern is a robust characteristic of physician productivity.

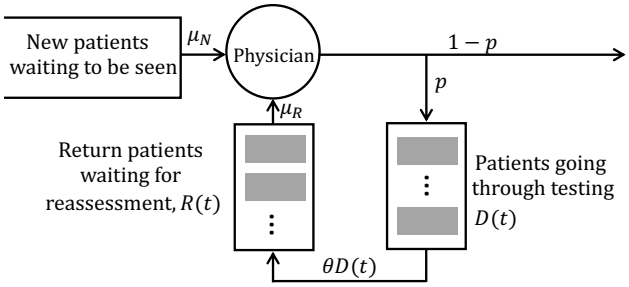
Note that the purpose of our exploratory study is neither to establish causality nor to identify the underlying mechanisms driving the observed timeÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœvarying productivity. In particular, the timeÃfÂ¢Ã¢â€šÂ¬Ã¢â‚¬Ëœvarying PPH pattern shown in Figure 2 likely reflects the joint effects of multiple interacting factors, which makes rigorous causal identification especially challenging. Nevertheless, we believe that a brief discussion of potential mechanisms, informed by the existing literature, may be of interest to readers and helpful in guiding future research. Hence, we include a short discussion in Appendix C.

4. Physician Behavior Behind Time-Varying Productivity

The exploratory analysis in Section 3 establishes a robust, non-linear pattern of physician productivity: a sharp decay at the start of the shift, a plateau in the middle, and a sharp drop near the end. While our exploratory study identifies this pattern and links it to shift hours, it does not explain why a rational physician would adopt this pacing. In this section, we develop an optimal control framework to investigate whether a physician maximizing throughput while subject to multitasking as workload accumulates and handoff/overtime avoidance would naturally generate a productivity trajectory consistent with the empirical

data. Rather than proving causality, this model serves to demonstrate that the observed PPH curve is consistent with rational behavior under ED operating constraints.

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, 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 7.

We further assume that return patients are prioritized over new patients, which generally aligns with the practices in our study hospitals. 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 active patients and the lower half shows the new patients waiting to be seen. The upper half is visible to this physician only, whereas the information on the lower half is available to all physicians. In general, a physician processes all of the reassessment requests before signing up a new patient so as to limit patients' length of stay. Physicians may also follow the shortest processing time rule because reassessment is generally faster than treating a new patient.

We 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:

$$\begin{aligned}
\max_{\alpha_N(t),\alpha_R(t)} \ & \left\{ \int_0^T (1-p)\left[\alpha_N(t)\mu_N + \alpha_R(t)\mu_R\right] \mathrm{d}t - h(D(T)+R(T)) \right\} \qquad (2)\\
\text{s.t.} \quad & D'(t) = p\left[\alpha_N(t)\mu_N + \alpha_R(t)\mu_R\right] - \theta D(t),\ D(t) \geq 0,\ D(0) = D_0 \geq 0,\\
& R'(t) = \theta D(t) - \alpha_R(t)\mu_R,\ R(t) \geq 0,\ R(0) = R_0 \geq 0,\\
& 0 \leq \alpha_N(t) + \alpha_R(t) \leq 1,\ \alpha_R(t) = \min\left\{1, \theta D(t)/\mu_R + \mathbf{1}_{\{R(t)>0\}}\right\},\ \alpha_N(t) \geq 0.
\end{aligned}$$

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. 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 (2) by applying the Pontryagin's maximum principle. We consider three cases: (i) $R_0 = 0, D_0 \le \mu_R/\theta$; (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).[2] 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 D.

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* (2) *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*

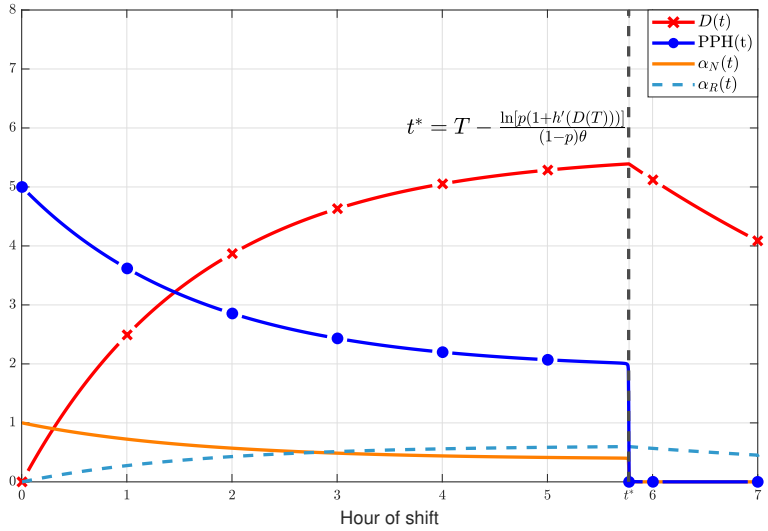
$$D(t) = \left(D_0 - \frac{p\mu_N}{\theta\left(1-p+p\mu_N/\mu_R\right)}\right) e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t} + \frac{p\mu_N}{\theta\left(1-p+p\mu_N/\mu_R\right)},\ t \in [0,t^*],\ \textit{and} \tag{3}$$

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

Theorem 1 completely characterizes the optimal policy for Problem (2) 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 (5) 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.

[2] 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 (2) 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 $t^* = 5.64$.**



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:*

$$t^* = \min\left(T, \max\left(0, T - \frac{\ln[p(1+h'(D(T)))]}{(1-p)\theta}\right)\right). \tag{5}$$

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

When the cost function is linear, i.e., $h'(\cdot)$ is a constant, the expression of t^* in (5) 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 spends 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* (2)*, we have*

$$PPH(t) = \frac{(1-p)\mu_N\mu_R}{p\mu_N + (1-p)\mu_R} + \mu_N\left(\frac{p\mu_N}{p\mu_N + (1-p)\mu_R} - \frac{\theta D_0}{\mu_R}\right)e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t},\ t \le t^*, \tag{6}$$

$$PPH(t) = 0,\ t^* < t \le T, \tag{7}$$

where $D(t^*)$ and t^* are given in (3) *and* (5)*, respectively.*

The expressions of $\text{PPH}(t)$ in (6) and (7) 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 this early-shift decay in pickup rates can arise from physician multitasking: as the shift progresses, physicians must devote increasing attention to reassessment requests and follow-up care for return patients, leaving less time available to pick up new patients. As the shift moves into its middle-of-shift phase, the pickup rate approaches a plateau, an effect reflected mathematically by the diminishing effect of the exponential term in $\text{PPH}(t)$ as t increases (see Figures 2 and 6). From a behavioral standpoint, the mid-shift plateau reflects a behavioral equilibrium in which physicians stabilize their pickup rate to balance the time spent assessing new arrivals against re-evaluating return patients, rather than merely a mechanical artifact of the exponential specification. During the end-of-shift phase, overtime and handoff avoidance discourage physicians from signing up any new patients, thereby contributing to the end-of-shift drop in productivity. 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 decrease with t. In fact, the expression of $\text{PPH}(t)$ in (6) 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. Impact of Time-Varying Service Rates on ED Modeling

The exploratory analysis in Section 3 establishes the pattern that physician productivity declines significantly over the course of a shift. Complementing this, the optimal control model in Section 4 provides a behavioral explanation for this pattern: if physicians are concerned about handoff or overtime avoidance, it is a rational decision to adjust their service rate during the shift. Taken together, these findings suggest that the time-varying pattern is an intrinsic and robust feature of physician productivity. In this section, we evaluate the impact of incorporating time-varying service rates into ED models compared to the standard assumption of constant rates.

5.1. 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, which makes the model analysis both theoretically and computationally challenging. 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. Our goal is to model the patient waiting process (time from triage to first physician assessment) and staffing decisions; we therefore propose and calibrate an effective pickup-rate process rather than explicitly tracking the network of queues with return of ongoing patients as in Yom-Tov and Mandelbaum (2014).

Motivated by the findings in Sections 3 and 4, we model ED operations as an $M(t)/M^{\text{PPH}}(t)/s(t)$ queueing system. This formulation captures time-varying arrivals, heterogeneous servers, and shift-hour-dependent service rates. 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). Let $\mathcal{S} = \{S_1, S_2, \cdots, S_k\}$ denote the daily schedule of k shifts. Due to shift-based scheduling, the number of active physicians (servers) at time t, denoted by $s(t)$, varies according to the overlap of these scheduled shifts (see Figure 1), where $s(t)$ is a nonnegative integer. 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.

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.

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 (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 A3 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 serves 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. 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.

5.2. 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.

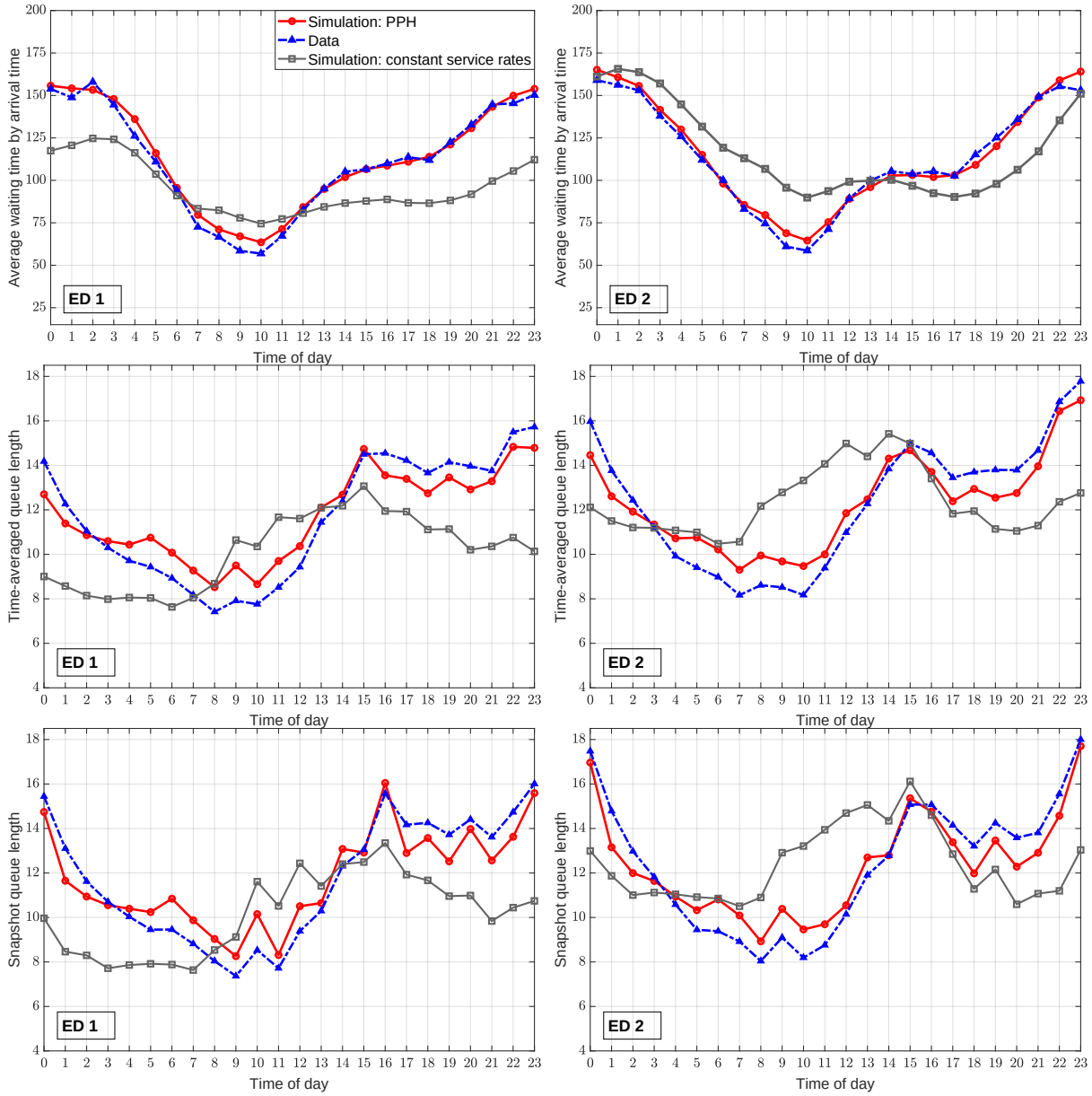
We ran the simulation for 5 replications, each with 500 weeks, and we identified the first 200 weeks as the warm-up period. We focus on waiting room dynamics and choose three time-of-day-dependent performance metrics: (i) the average waiting time of patients who arrived in the same hour of day; (ii)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

5.3. Simulation Results

The results of comparing the simulated performance measures with the data are shown in Figure 7. We observe that the average waiting times from the simulation 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 7. Furthermore, the aggregated average waiting time of all patients from ED 1 (ED 2) is 106.4 (109.5) minutes, whereas the simulated counterpart is 108.3 (110.9) minutes, which further shows the accuracy of our model.

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 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.

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



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 time-varying 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 $M(t)/M^{\mathrm{PPH}}(t)/s(t)$ 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%.

6. 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.

6.1. Improving Physician Staffing: A Case Study

In our study hospital, a scheduler first determines the start and end times of each shift every six months (more or less); then, physicians are allocated to each shift following required scheduling rules. Figure 4 shows the 15 physician shifts from January to July 2015 in our study ED. Among these, S6 and S11 are fast-track shifts, and all others are dedicated to serving patients in the main ED 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 with the uniformization method for the evaluation of each candidate schedule and this procedure takes less than two hours for each scenario. See more details in Appendix E.

6.2. 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 the waiting times obtained from 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 8. 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 8 shows that the reductions over the baseline schedule are statistically significant at the 5% level.

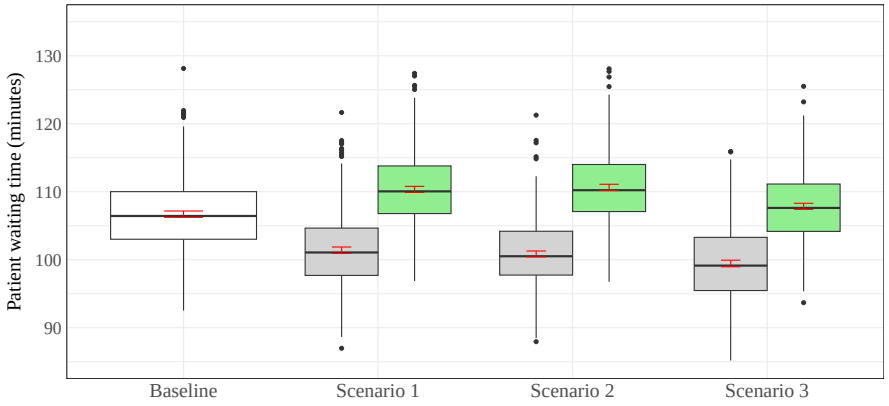
In contrast, when the time-varying service rates for physicians are not considered, more specifically, when we assume constant service rates over a shift, the optimized schedules perform worse than the baseline schedule. The average patient waiting time increases from 1.1% to 3.7% compared to the baseline schedule, which is equivalent to 3.1 to 9.4 hours of waiting for all patients in the ED per day. Figure 8 shows that the

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 8 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.



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,[3] 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), the implied annual cost savings for our study hospital range from CA$456,446 to CA$912,892. 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 incorporating time-varying service rates in physician staffing.

7. Conclusion and Future Research

Existing ED staffing models typically assume constant physician service rates, implying that ED productivity, measured by the number of new patients that can be seen per hour, depends solely on the number of physicians on duty. Motivated by empirical observations of time-varying physician productivity patterns during shifts, we challenge this simplifying assumption by examining the underlying factors driving these variations and their implications for ED modeling and staffing optimization. Our exploratory analysis reveals that shift hour is the predominant factor in explaining and predicting PPH fluctuations. Furthermore, using an optimal control framework, we demonstrate that this shift-hour-dependent productivity structure is not random but a rational behavioral response if physicians are concerned about overtime/handoff avoidance. Consequently, we argue that accurate ED modeling and staffing optimization must account for the time-varying nature of physician productivity.

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 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 incorporate 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

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

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 pickup 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) highlight that the true nature of ED service times is often obscured by physician multitasking. While modeling these interruptions explicitly is challenging, our results suggest that using PPH as an aggregate measure of service rate offers a practical and sufficiently robust alternative for ED modeling and staffing decisions. An interesting direction for future research would be to estimate physician PPH using the structured model in (6) and assess its performance based on the predicted PPH values.

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.

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

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

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.

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.

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.

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.

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, 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, 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. 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.

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.

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.