Submitted to *Manufacturing & Service Operations Management* manuscript

The Cost of Task Switching: Evidence from Emergency Departments

(Authors' names blinded for peer review)

Problem Definition: Emergency department (ED) physicians treat patients with different symptoms and constantly switch between tasks. Utilizing three years of comprehensive data on patient visits and lab tests from two large EDs, we investigate the impact of task switching on physician productivity, quality of care, and patient routing. We subsequently provide operational solutions based on the identification findings.

Methodology: To address estimation bias due to measurement errors and endogenous patient selection, we refine the sample period and construct an instrumental variable called switching likelihood, which exploits the exogenous composition of waiting patients. By exploring the heterogeneous impact on physician productivity among different patient type pairs from data, we leverage a max bisection algorithm to partition patients into two clusters to minimize the negative impact of task switching.

Results: Our estimates indicate that, at different EDs, a 10% increase in the switching frequency of physicians reduces the number of patients treated per hour by 8.65% - 11.53% on average. However, we find no significant influence on treatment quality. We propose a data-driven queue management method to optimally partition patients into two queues. Based on the simulation of implementing the proposed two-queue system in our collaborating EDs, we find that the average waiting time is reduced by up to 40%.

Managerial Implications: Task switching negatively impacts ED physician productivity, and this impact is more prominent for certain patient type pairs. Being aware of the switch cost, we propose measures to mitigate switch costs, which can considerably reduce ED congestion and patient waiting times.

Key words: Task switching, emergency department, behavioral queueing, data-driven, queue management, empirical, max bisection, graph.

1. Introduction

Recent technological advancements and economic growth have given rise to an increase in skilled and complex jobs, replacing routine and monotonous positions (Autor et al. 2003, Acemoglu and Autor 2011, Acemoglu and Restrepo 2018). As a result, workers are now required to frequently perform and transition between various tasks (Lindbeck and Snower 2000). However, this trend poses challenges, as psychological studies have demonstrated that task switching imposes additional cognitive burdens, prolongs task completion times, and increases performance errors (Jersild 1927, Allport et al. 1994, Rogers and Monsell 1995, Rubinstein et al. 2001). Consequently, the cost of task switching (also known as switch cost) has become an increasingly crucial determinant of worker productivity.

Despite the psychological experiments conducted, there remains limited evidence concerning the impact of task switching on workplace productivity. Studies investigating this phenomenon face challenges in measuring task switching, as detailed diary data on tasks performed during everyday work is often unavailable or too expensive to gather. Furthermore, task switching frequently occurs endogenously (Ibanez et al. 2018), resulting in biased estimates. These complications hinder the evaluation of task switching consequences, particularly from a causal perspective.

In this study, we investigate the influence of task switching on the productivity of emergency department (ED) physicians. As jacks of all trades, ED physicians possess a wide array of medical knowledge, practice various clinical skills, and treat patients presenting diverse medical symptoms. Due to the variability in patients' symptoms and treatment approaches, physicians are compelled to switch tasks regularly. Additionally, ED physicians work individually and face a high volume of incoming patients their shifts. These factors establish EDs as an ideal setting to examine task switching and its effects on physician productivity.

We leverage the administrative data of patient visits and lab tests from two large EDs on the west coast of Canada. To identify task switching, we categorize patients into different *types* based on their chief complaint systems (CCSs), which are recorded by a triage nurse upon the patient's arrival at the ED. For each physician, we measure task switching by calculating the *switching frequency*, defined as the number of patients with a differing type from the preceding picked patient relative to the total number of patients treated by the physician within a specific time frame. We then quantify the switch cost by examining how switching frequency impacts service speed and quality. Additionally, we explore the effects of task switching on physicians' routing decisions.

Our analysis, however, encounters two primary challenges. The first challenge pertains to the measurement of physician service speed. Since our data does not capture every detailed activity of physicians, there is no direct measure of the actual time a doctor spends on each patient. Specifically, ED physicians frequently face interruptions from reentrant patients and multitasking (KC 2013). The second challenge lies in addressing the endogeneity issue during the estimation process. In our collaborating EDs, physicians select patients to treat when they become available, rather than being assigned patients by others. This grants physicians considerable discretion in patient selection. Consequently, the observed sequence of patient types is subject to selection bias, as physicians may strategically choose patients. If physicians aim to minimize switch cost by prioritizing patients of the same type, ordinary least squares (OLS) estimates will likely underestimate the actual impact of task switching (Heckman 2010).

To address the first challenge, we refine the sample period to a *clean period* and employ the number of patients treated per hour (PPH) as a measure of physician service speed. Leveraging a comprehensive and unique lab test dataset, we obtain the *earliest lab result time* for each

physician's shift by identifying the minimum lab result release time among two patient groups: (1) patients selected by the focal physician earlier in the day, and (2) patients picked by other ED physicians earlier in the day whose shifts have concluded by the lab result release time. The second group encompasses all potential hand-over patients. We then define the clean period as the duration from the start of the shift to the earliest lab result time for each physician. During the clean period, physicians primarily perform initial assessments (IA) for new patients, with reentrant interruptions being virtually absent as no lab results have been returned yet. Meanwhile, the PPH measures a physician's average efficiency over a time period, effectively averaging out patient-specific measurement errors and noises. Notably, the PPH is unaffected by physicians' batch-picking and multitasking behavior. We divide the obtained clean period into several 30-minute blocks, for which we compute the PPH.

We address the second challenge by proposing an instrumental variable (IV) as we utilize the *switching likelihood* (SL) as the IV. SL is the ratio of unique types to the queue length at the beginning of each 30-minute block. By design, SL is positively correlated with the physician's actual switching frequency and independent from the focal patient's unobserved physical attributes. Importantly, it capitalizes on the type composition of waiting patients determined prior to patient selection in the subsequent stage. This allows for exogenous variation in patient-type switching, enabling us to distinguish the switch cost from the endogenous patient selection process.

Our analysis reveals that across different EDs, a higher switching frequency leads to decreased physician productivity. Specifically, holding others constant, a 10% increase in the switching frequency results in an average reduction of 0.41 to 0.47 patients treated per hour, or an 8.65% to 11.53% decrease in PPH at the two EDs. The estimated cost is statistically significant and robust across alternative specifications. Additionally, we find that physicians exhibit switch-aversion, meaning they strategically select patients of the same type to circumvent task switching. However, task switching has minimal impact on treatment quality, as measured by 7- and 30-day revisit-and-readmission (RAD) rates. These findings suggest that task switching primarily influences the efficiency of ED operations, while physicians manage to reduce task switching (possibly subconsciously rather than intentionally) without significantly affecting service quality.

As an operational solution, we propose a practical, data-driven queue management method that leverages the identified switch cost heterogeneity among different CCS pairs to optimally partition patients into two queues. Our queue redesign minimizes intra-queue switching costs and maximizes inter-queue switching costs to mitigate the negative impact of task switching on ED efficiency. We demonstrate that designing such a two-queue system is equivalent to solving a max bisection problem, which can be addressed using semi-definite programming (SDP) relaxation (Goemans and Williamson 1995, Frieze and Jerrum 1997). Through simulation, we further show

that physicians' productivity improves under the proposed two-queue system, resulting in a more substantial reduction in patient waiting times and ED congestion compared to current practices. Our findings indicate that the average PPH at both EDs increases by 3.25% to 6.19% using this new design, translating to a 38.16% to 39.66% reduction in average patient waiting time.

Our study makes several contributions. First, we illustrate that task switching is costly for emergency care delivery by impairing physician productivity. Our estimates indicate that task switching reduces ED physician service speed. Due to the queueing system's externality, task switching's negative impact can be more significant at the ED system level compared to the patient level. We also find evidence that ED physicians tend to prioritize patients without task switching in the routing stage. Although this may partially alleviate the switch cost, the estimated impact of task switching remains sizeable and robust.

Second, we propose an implementable data-driven queue management method in ED based on heterogeneous switch cost estimates. To mitigate the negative impact of task switching, we partition patients into two queues so that patients with inexpensive switching costs are clustered into the same queue. This method is general and can easily be extended to incorporate other patient demographics. Through a simulation study, we find that implementing such a two-queue system can reduce average patient waiting time by up to 40%. This will yield a substantial improvement if implemented in our collaborating ED.

Third, the identification strategy developed in this paper may appeal to other empirical researchers interested in ED operations. By exploiting the lab test data, we propose a new approach to identify a clean period in which physicians are almost free from interruptions from reentrant patients, a primary challenge in estimating the ED physician's productivity. We also propose an IV to address endogeneity due to physicians' strategic patient selection. This IV can be easily constructed in commonly used ED data sets and generalized to other ED-related studies.

The remainder of our paper is organized as follows. The next section reviews the literature related to our work. Section 3 describes the ED background information and the data we use. Section 4 formulates the empirical models. The main results are presented in Section 5. Then Section 6 proposes a data-driven queue management method and present simulation results. In Section 7, we explore the influence of task switching on the quality of care and patient routing decisions. Finally, Section 8 discusses the managerial implications. Section 9 concludes.

2. Literature Review

Our paper is related to five strands of literature. First, we contribute to healthcare management studies on physician productivity. From a queueing theoretic perspective, EDs can be modeled as a system with state-dependent service rate (Mandelbaum and Pats 1998, Abouee-Mehrizi and

Baron 2016). As such, existing research has investigated various factors affecting productivity and efficiency in ED, including queue length (KC and Terwiesch 2009), shift schedules (Chan 2018, Batt et al. 2019), triage (Batt and Terwiesch 2016) and peer pressure (Chan 2016, Silver 2020). They also propose alternative designs of the care delivery system regarding physician shift scheduling (Liu and Xie 2018, Zaerpour et al. 2022), new triage policies (Saghafian et al. 2014), and waiting time prediction and announcement (Ang et al. 2016, Dong et al. 2019, Ding et al. 2020). KC (2019) provides an overview of studies on worker productivity from different angles. We complement the above literature by identifying the switch cost as a novel factor influencing physician productivity and system performance.

Second, our paper is related to research on multitasking (KC 2013, Goes et al. 2018) and interruptions (Cai et al. 2017, Gurvich et al. 2019). KC (2013) shows that a physician's productivity has an inverted U-shape response to her on-hand patients indicating the multitasking level. Our paper complements this finding by showing that the patient mix and the service sequence also matter—a physician becomes more productive when he consecutively sees patients sharing the same type. Since we are concerned with the IA time the physician spends on each individual patient, we use a different measure of physician productivity from KC (2013). Regarding interruptions, note that it is different psychological notion than task switching as interruptions represent distinct cognitive demands and disruptions. Task switching, as an executive function, involves voluntarily or involuntarily shifting attention and cognitive resources between tasks (Jersild 1927, Allport and Wylie 1999, Monsell 2003). This cognitive flexibility enables individuals to adapt to new task demands. Interruptions, however, entail unexpected breaks that introduce new tasks, often involuntarily (Miyata and Norman 1986). These disruptions, such as phone calls or messages, necessitate a temporary shift in attention before resuming the interrupted task (Miyata and Norman 1986, González and Mark 2004, Mark et al. 2005). While both task switching and interruptions demand attentional and cognitive shifts, the primary distinction lies in the voluntary or involuntary nature of these shifts. Our paper complements this thread of existing literature by examining the impact of task switching on focal task service speed.

Our research is also related to task specialization (Ong and Png 2021, Gong and Png 2022) and task variety and their effects on productivity (Staats and Gino 2012, Avgerinos and Gokpinar 2018, Narayanan et al. 2009). Conceptually, task specialization and task variety are cumulative, and their effects on productivity dampen or exacerbate over time. However, task switching is an instantaneous event, and its effect is relatively transitory. Note that task variety also differs from our "frequency of task switching." To illustrate, consider two job sequences consisting of two job types: 1-2-1-2-1-2-1-2-1-2 and 1-1-1-1-1-2-2-2-2-2. Both sequences have the same "variety of tasks" but exhibit distinct task switching frequencies. Existing research also shows that task performance

only benefits from experience on related tasks (Schilling et al. 2003, Boh et al. 2007, KC and Staats 2012). Some researchers explore task pooling and its impacts on healthcare service efficiency and quality (Song et al. 2015, 2020). Our subsequent queue design also connects to this thread.

Fourth, our paper connects to a growing body of research on discretionary task routing. Ding et al. (2019) show that ED physicians route patients according to their waiting time and triage acuity codes, but Bayati et al. (2017) find that the arrival of low acuity patients may delay the treatment of high acuity patients. Related to our paper, KC et al. (2020) examine how varying workload conditions lead individuals to self-select tasks from a larger set of available tasks in the ED context, focusing on the short-term and long-term effects of such behavior. Their study centers on physicians' inclination to choose easier patients when workloads are heavier. This definition of "task switching" differs from ours, as easier patients may have the same or different types as previous patients. Furthermore, they find that this behavior improves throughput for the current shift but hurt long-term productivity in the future. However, we focus on the the influence of task switching on the focal task. Ibanez et al. (2018) examine how radiological physicians sequence tasks given a preassigned workload and how that will impact productivity. They find that endogenously circumventing task switching by grouping similar tasks impedes productivity. Although they posit that exogenously grouping similar tasks could be beneficial. In our paper, we find that most ED physicians have a tendency to prioritize tasks of similar types, resulting in task switching arising from a combination of endogenous and exogenous behavior. In this context, we show that task switching still negatively impacts productivity.

Finally, we contribute to the literature on the efficiency-quality trade-off in service queues. Allon and Kremer (2019) review a broad scope of papers in behavioral queuing research. They quantitatively identify the total system welfare as the product of customers' net utility and system throughput. These two components represent the system operation efficiency and service quality, respectively. Song and Veeraraghavan (2018) review and show that a vital triptych of quality measures is structure, process, and outcome in healthcare analytics. Roth et al. (2019) study the trifecta among efficiency, quality ,and patient experience in hospitals. In addition, Batt and Tong (2020) investigate when and how server-level quality metrics differ from customer-experienced metrics and the effect of such judgment. Our paper investigates the potential trade-off between the time and quality of care delivery due to task switching and finds that switch cost materializes in the dimension of time mostly.

3. Clinical Setting and Data

Our study takes place at two major EDs in Vancouver, Canada. Both EDs provide emergency care for thousands of patients each month. ED A operates from 8 AM to 8 PM seven days a week, while

ED B functions on a 24/7 basis. ED physicians work independently, treating patients with a wide range of demographics and medical conditions. A typical physician's shift lasts for seven to eight hours.

3.1. The Care Delivery Process

Figure 1 illustrates the treatment process when a patient visits an ED. Upon arrival, a triage nurse attends to the patient, collecting their demographic and clinical information, evaluating the medical situation, and assigning a triage acuity code ranging from 1 (resuscitation) to 5 (nonurgent). The nurse also classifies the patient's symptoms into a chief complaint according to the a specified list (Grafstein et al. 2008). These information are recorded in an online system that physicians can easily access. Patients with a triage code of 1 are sent directly to physicians for urgent treatment (which account for less than 0.5% of all visits), while others typically wait in the waiting room until an available physician selects and meets them. When available, physicians scan the online patient information and pick the next patient from the waiting room. A physician then meets with the selected patient in the treatment room for an initial assessment (IA), which is the focus of our study. Although there is no explicit selection rule, it has been shown that physicians consider both the patient's waiting time and triage code. Specifically, physicians attempt to adhere to a fractile response objective, which outlines a triage-specific target waiting time and a target percentage of patients to be seen within that time frame (Beveridge 1998, Ding et al. 2019). However, the high volume of patients often makes it challenging for ED physicians to comply with this objective.

In order to acquire a more precise measure of physician productivity, we examine the detailed activities ED physicians engage in during a patient's length of stay at the EDs. As depicted in Figure 1, a physician's activities can be categorized into two stages.

1. Initial Assessment (IA): A physician selects a patient by scanning information in the patient care information system. The physician needs to physically find the selected patient, talk to the patient, make diagnostic decisions, order lab tests, and charting. After the first meeting, the patient is put under observation, during which the physician can chart for the patient and work on other patients.
2. Follow-up Services (FS): A physician provides follow-up assessment and consultation when a patient's lab test results are available or when the patient's observation period ends. In the former case, the physician reviews the lab results with the patient. Then the physician decides whether to discharge or admit the patient to an inpatient unit.

The specific sequence of the aforementioned events may differ from case to case. For instance, lab tests are occasionally ordered before the diagnosis based on the physician's review of the patient's information from the patient portal prior to talking to the patient.

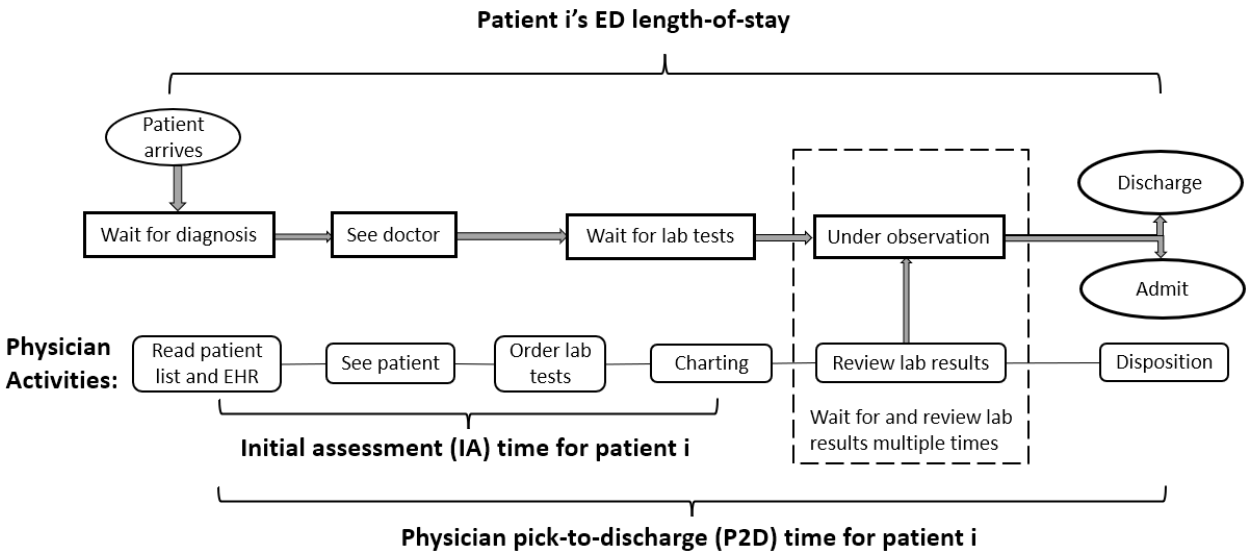


Figure 1 Typical ED Care Delivery Process

In summary, physicians closely attend to patients during the IA period, which serves as a suitable indicator of ED physicians' service speed. Conversely, the FS period is subject to various disturbances, such as reentrant interruptions from previous patients or requests to review lab test results. As a result, our study primarily investigates the impact of task switching on the duration of the IA period.

Note that in the existing literature (e.g., Kuntz and Sülz 2013, Batt and Terwiesch 2016, Chaou et al. 2018), the service time may refer to the pick-to-discharge (P2D) time, i.e., the interval between when a physician selects a patient and when that patient is discharged home or admitted to an inpatient unit. However, the P2D duration of an individual patient is far from an accurate measure of the actual time a physician spends on the patient. This is because the majority of a patient's P2D duration (on average, more than three hours) is spent waiting for lab results, during which the physician may spend most of their time treating other patients (KC 2013).

To gather firsthand evidence supporting the use of IA in our study, we conducted a time-and-motion study by shadowing physician shifts at one of the two EDs. As external observers, we recorded a physician's time spent on each individual activity using a stopwatch and interviewed several healthcare workers to gather additional information. In total, we shadowed two shifts for four hours each and observed the treatments of 22 patients. The average IA time per patient in our shadowing study is 13.3 minutes, and the average time for each FS is 3.6 minutes.

The time-and-motion study provided valuable insights into the care delivery process. First, using the P2D time to measure a physician's actual time spent on each patient would result in a significant overestimation. This is because, during most of the P2D duration, patients are waiting for lab

results or disposition decisions, which require minimal input from the physician. Second, the time-and-motion study demonstrates that the service process can be best described by a single-server reentrant queue (Yom-Tov and Mandelbaum 2014, Huang et al. 2015, He et al. 2019). After being selected, a patient will progress through

$$\text{IA} \to \text{waiting} \to \text{FS} \to \text{waiting} \to \text{FS} \to \dots$$

until being discharged or admitted. In this process, waiting and FS may repeat multiple times. Third, most of the actual service time is spent in the IA stage. The majority of the physician's work, including selecting and locating the patient, diagnosing the problem, ordering lab tests, and charting, is completed during IA, and physicians need time to familiarize themselves with the patient's clinical condition. In contrast, the FS time is much shorter because the physician has already gained an understanding of the patient's situation.

3.2. Identifying the Clean Period

In this study, we focus on how task switching impacts physician productivity during the IA period, for which we adopt a series of cleaning and identification strategies. To distinguish IA from the overall P2D time, we exploit a unique lab test data set with timestamps of the request, performance, and release of all lab tests (blood, urine, stool, ECG, X-ray, CT scan, etc.) Within each physician's shift, we define the *earliest lab result time* as the minimum lab result release time among two patient groups: (1) patients picked by the focal ED physician at an earlier time of the day and (2) patients picked by other ED physicians earlier in the day whose shifts have ended by the lab result release time. The second group encompasses all patients who could potentially be *handed over* to the focal physician by other physicians who have completed their shifts. By defining the earliest lab result time in this manner, we ensure that the focal physician will not receive any lab test results for potential reentrant patients, including those they have seen earlier and those who might be handed over to them by other physicians who have completed their shifts. We refer to the period between the shift start time (i.e., the time to pick the first patient in a shift) and the earliest lab result time as the *clean period*. In Figure 2, we present a diagram to depict how we obtain the clean period.

Based on our discussions with ED physician collaborators and our own extensive experiences as patients visiting the ED, a physician generally will not meet a patient again before receiving updates about the patient's lab test results after ordering lab tests. In fact, even when a patient's lab results are out, the physician may process other jobs first before responding to the lab results. For example, according to Figure 2, the physician may first conduct IA for patient $n+1$ before responding to the lab results of patient 2. The above definition of the earliest lab result time guarantees that

during the clean period, they are unlikely to spend time on reentrant patients (patients who have already met the focal physician or other physicians). While there may be exceptional cases, such as a physician needing to speak to a reentrant patient again due to forgetting something or a reentrant patient's health suddenly deteriorating, these instances are rare in daily operations. Interruptions or multitasking could also be caused by the arrival of triage-1 patients with preemptive. However, such interruptions can be captured in our empirical analysis using the timestamps of treatment for the triage-1 patients. Furthermore, triage-1 patients only account for less than 0.5% of the total patient volume, so including or excluding these cases would not change our results.

We also assess the representativeness of the refined sample. We explore the distribution of patient demographics in the refined sample and the entire sample in Section EC.1.2. The results indicate that the patients used in our study are representative of ED operations.

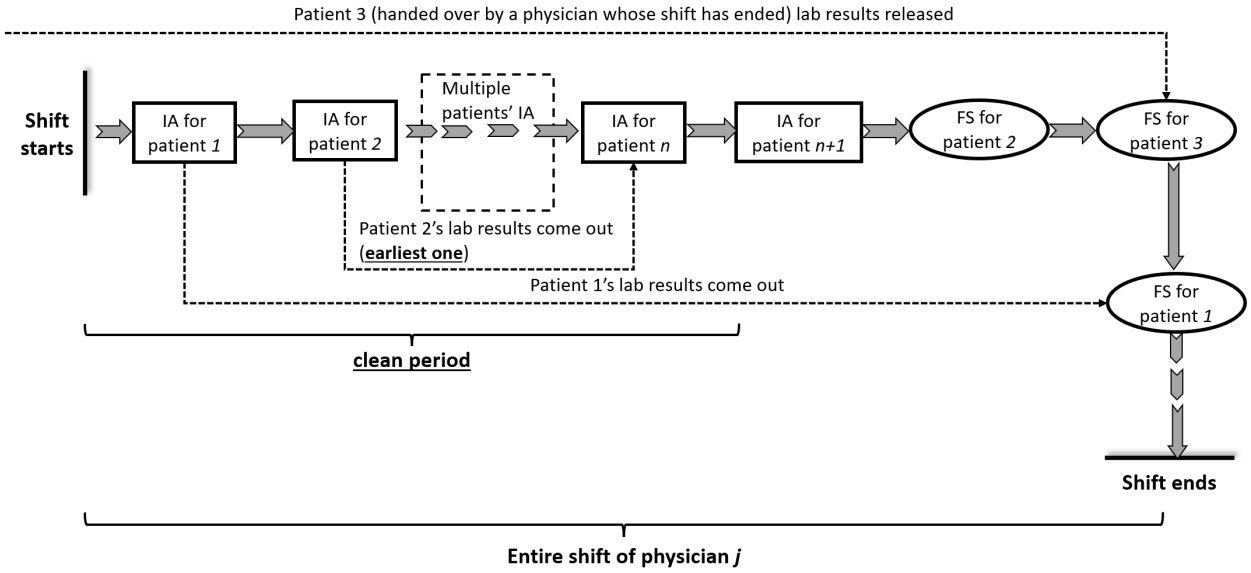


Figure 2 ED Physician Shift Structure and Clean Period

3.3. Data

Our primary dataset encompasses all patient visits to the two EDs from April 2013 to March 2016. ED B has a distinct regular/fast track system, while ED A operates a single track where patients of all types are pooled together. Following Ding et al. (2019), we focus on ED A and the regular track for ED B in the main model and include the results of the ED B fast track as a robustness check in Section EC.1.

The data set contains rich information about patient characteristics such as gender, age, homelessness, arriving transportation, triage acuity code, and chief complaints. A total of 163 chief complaint categories (CCC) are identified under 15 chief complaint systems (CCS) (for example,

cardiovascular or gastrointestinal). In the baseline analysis, we define *task switching* as whether the focal and preceding picked patients of the same physician have different CCSs. The focal patient is named *switcher*. In section EC.1, we also consider alternative definitions of task switching. Additionally, we define the waiting time of a patient as the difference between the arrival time and the calling time (Dong et al. 2019, Ding et al. 2019), and we use the revisit-and-readmission (RAD) rates in 7 or 30 days as measures of service quality (Calder et al. 2015, Wang et al. 2019).

We measure physician productivity using patients treated per hour (PPH) (Zaerpour et al. 2022, Ouyang et al. 2021) during the clean period, in which potential noises from reentrant patients are eliminated. As a variable at a more aggregate level, PPH is robust to physician multitasking and batch-picking behavior at the beginning of physician shifts. It also averages out random errors. To calculate PPH, we partition the refined clean period into multiple 30-minute blocks, with the last block rounded upward (to include the portion beyond the earliest lab result time). The PPH of each 30-minute block represents the number of service cycles (pick-to-pick period) contained in that block divided by 30 minutes. Service cycles that are censored by the 30-minute cutoffs are counted as half of a normal cycle, based on the assumption that censoring times are independent of patient pick times. We also remove blocks during which no patient is waiting in the ED. This ensures that PPH does not overestimate physician productivity due to potential idling times.

Table 1 summarizes the key variables in the refined sample. The sample we use for analysis includes 3,559 and 4,879 shifts for the two EDs, and the sample period ranges from 1.08 to 1.19 hours on average. Overall, we obtain 7,845 30-minute blocks as observations for ED A and 10,521 blocks for ED B. The average numbers of patients served are 2.391 and 2.031 per 30 minutes, equivalent to 4.783 and 4.062 patients per hour. The average 7 and 30-day RAD probabilities are 0.02 and 0.045 for ED A and 0.033 and 0.083 for ED B. Among all patients at the two EDs, 13.3% and 26.8% of patients are admitted to an inpatient unit, respectively.

To measure task switching, we compute the *switching frequency* by taking the ratio of the number of switchers among the picked patients in each block, which serves as a proxy for switching intensity in a certain block. The average switching frequencies at the two EDs are 0.865 and 0.839, respectively, suggesting that task switching is common in ED services. The average switching likelihoods (which is the instrumental variable we use, refer to Section 4) at the two EDs are 0.809 and 0.758, respectively.

Regarding other ED- and patient-level variables, the average numbers of waiting patients (queue length) at the two EDs are 4.52 and 5.15, and the average numbers of patients being served (ED load) are 10.76 and 25.9, respectively. We use this variable to control for ED bed constraints. The average age of patients is around 50, and the gender distribution is relatively even. Moreover, the two EDs differ significantly in the proportion of patients arriving by ambulance. This is because

ED B is larger and serves as a major ambulance destination. We also include the proportion of five groups of patients with different waiting times. ED physicians see most patients within 60 minutes of waiting time, while only about 17% of all patients are seen in 15 minutes. Most patients have intermediate triage levels (triage codes 2, 3, and 4).

Table 1 Summary Statistics

Variables	ED A Mean (S.D.)	ED B Mean (S.D.)
Patients treated in each block	2.391 (1.106)	2.031 (0.95)
PPH	4.783 (2.212)	4.062 (1.9)
7-day RAD	0.020(0.103)	0.033 (0.141)
30-day RAD	0.045 (0.152)	0.083 (0.216)
Inpatient admission	0.133 (0.256)	0.268 (0.35)
Switching frequency	0.865 (0.255)	0.839 (0.299)
Switching likelihood	0.809 (0.196)	0.758 (0.216)
Average queue length	4.520 (2.803)	5.152 (3.444)
ED load	10.76 (5.083)	25.94 (7.119)
Patient age	53.48 (15.58)	49.74 (15.14)
Female patient	0.538 (0.357)	0.447 (0.387)
Arrival by ambulance	0.125 (0.244)	0.389 (0.38)
Triage code	3.309 (0.505)	2.899 (0.522)
<15 minutes	0.170 (0.300)	0.160 (0.306)
15–30 minutes	0.369 (0.389)	0.326 (0.393)
30–60 minutes	0.333 (0.382)	0.306 (0.384)
60–120 minutes	0.156 (0.314)	0.195 (0.339)
>120 minutes	0.021 (0.121)	0.052 (0.193)
Physicians	57	60
Shifts	3,559	4,879
30-minute blocks	7,845	10,521
Average time since shift starts to earliest lab result	1.078	1.192

4. Empirical Methods

4.1. Model

Consider ED physician shifts indexed by $i = 1, 2, \cdots, I$ where I is the total number of shifts (3,559 for ED A and 4,879 for ED B). Each 30-min block in shift i is indexed by $t = 1, 2, \cdots, N_i$, until the earliest release of lab test results. In each block, we assume the following linear model:

$$\text{PPH}_{it} = \delta \cdot \text{SF}_{it} + \mathbf{x}_{it}^{\top}\beta + \mathbf{F}_i + \mathbf{D}_{it} + u_{it}. \tag{1}$$

The dependent variable PPH_{it} is the extrapolated number of patients seen per hour. The key explanatory variable SF_{it} is the switching frequency. The parameter of interest, δ, reflects the effect of task switching on PPH. Covariates in $\mathbf{x}_{it}$ include the average queue length and ED load within the block; the average age and triage level of treated patients; the proportion of treated patients in

each waiting time range; the share of female, homeless, and arriving by ambulance patients; and the share of patients in each CCS. $\mathbf{F}_i$ denotes physician fixed effects and day-of-week fixed effects that do not vary within each shift i. $\mathbf{D}_{it}$ stands for the clock hour fixed effects of the start of block. Lastly, u_{it} is the unobserved error term.

Estimating Equation (1) using OLS yields a biased estimate of δ because the switching frequency can be correlated with the error u_{it}. One cause of such correlation is the physician's endogenous selection of patients (Heckman 2010), where the likelihood of selecting a specific patient depends on the patient's attributes (Ibanez et al. 2018, KC et al. 2020). In Section 7, we show that physicians are likely aware of the switch cost and tend to avoid selecting patients with a different type.

4.2. Instrumental Variable

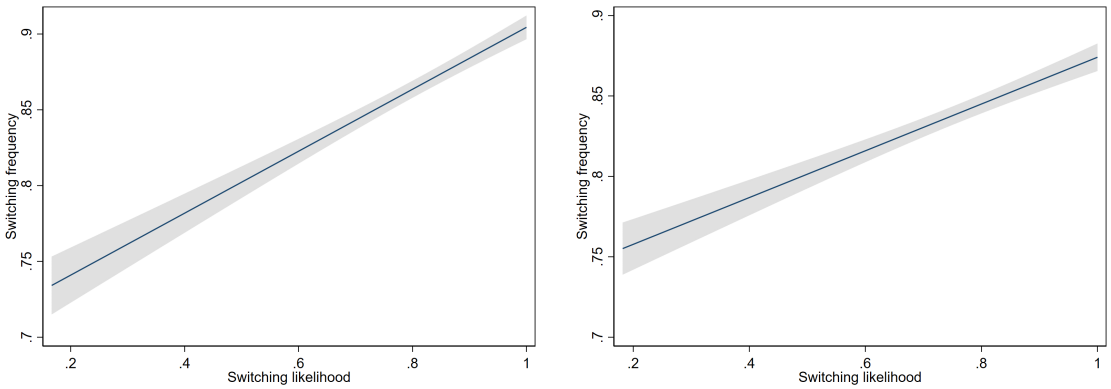
To correct the sample selection bias, we exploit the exogenous composition of waiting patients to construct a variable called *switching likelihood* (SL) as the IV for switching frequency. Specifically, let $\mathcal{J}(i,t)$ be the set of waiting patients at the starting time of block t of shift i. We define SL as the number of unique types in $\mathcal{J}(i,t)$ divided by the total number of waiting patients, $|\mathcal{J}(i,t)|$. As shown in Figure 3, SL is positively correlated with switching frequency: The more unique types among waiting patients at the beginning of the shift, the more likely the physician will switch between types in subsequent treatment. For example, suppose $\mathcal{J}(i,t)$ consists of three patients and each has a distinct type. Then the SL equals one and the physician will likely switch tasks between every two patients. In contrast, if all patients had the same type, the switching likelihood would be zero and the physician would hardly switch tasks.

Moreover, we argue that our IV satisfies the exclusion restriction, i.e. it affects the PPH only through task switching. By construction, SL only depends on pre-determined composition of waiting patients who arrive on average 40 minutes before the block begins. Thus, it is uncorrelated with unobserved factors during the block that contemporaneously affects PPH. On the other hand, waiting patients who are not selected for treatment should not affect physicians' productivity with respect to the patients being treated. In particular, since physicians already prioritize the most acute patients at the beginning of the shift, we do not expect further interruptions from other waiting patients. If anything, interruptions arise from previously treated patients or patients with time-varying medical conditions, but both are uncorrelated with the type composition of waiting patients at the beginning of the shift.

With the IV, we estimate the following equations using limited information maximum likelihood (LIML), which has better finite sample properties than two-stage least squares (Hansen 2022):

$$\mathrm{PPH}_{it} = \delta \cdot \mathrm{SF}_{it} + \mathbf{x}_{it}^{\top}\beta + \mathbf{F}_i + \mathbf{D}_t + u_{it}, \tag{2}$$

$$\mathrm{SF}_{it} = \alpha \cdot \mathrm{SL}_{it} + \mathbf{x}_{it}^{\top}\gamma + \mathbf{F}_i + \mathbf{D}_t + \epsilon_{it}. \tag{3}$$



(a) ED A (b) ED B

Figure 3 Relationship between SL and switching frequency: OLS Fitting Line and 95% Confidence Interval

It is worth mentioning that our IV approach shares a similar spirit with Ibanez et al. (2018), who examine discretionary task routing and estimate the impact of deviation from first-come-first-serve (FCFS) on the image-reading speed of radiologists. Their IV for deviation is based on an indicator of whether there is a chance to deviate from a prescribed sequence of images. Analogously, our IV can be interpreted as the extent to which a physician may avoid task switching. If the SL is high and most patients have distinct types, the physician has little chance to avoid task switching.

5. Results

5.1. The Effect of Task Switching on Physician Productivity

Table 2 presents estimates of Equation (2). The first row displays estimates of δ, representing the cost of task switching on the PPH of ED physicians. For the two EDs, the estimates are -4.139 and -4.685, respectively, both of which are statistically significant. Consequently, a 10% increase in switching frequency results in an expected PPH reduction of $4.139 \times 0.1 = 0.41$ and $4.685 \times 0.1 = 0.47$ for the respective EDs. These reductions correspond to 8.65% and 11.53% decreases in PPH relative to the sample means (see Table 1). This indicates that task switching leads to substantial efficiency losses in ED operations.

In comparison, Table EC.1 in Appendix EC.1.1 presents the OLS estimates of Equation (1). Compared with the LIML estimates, the OLS estimates of δ are much smaller and statistically insignificant. These results indicate that physicians' discretionary selection of patients would bias the OLS estimates and therefore justify our use of IV. Besides, the bottom of Table 2 shows the Wald F statistics. At both EDs, we reject the null hypothesis that the IV is only weakly correlated with switching frequency. This further enhances the validity of our IV.

We perform several robustness checks for the switch cost estimates. The results are reported in Appendix EC.1.2. In particular, our estimates are robust to alternative definitions of patient types

Table 2 Effect of Task Switching on PPH

	ED A	ED B
Switching frequency	-4.139***	-4.685***
	(1.041)	(0.993)
Average queue length	0.092***	0.031***
	(0.015)	(0.011)
Average ED load	0.015*	0.022***
	(0.008)	(0.005)
Average age	-0.003	-0.001
	(0.002)	(0.002)
Average triage	0.423***	0.046
	(0.073)	(0.055)
Female proportion	-0.009	-0.046
	(0.066)	(0.054)
Homeless proportion	0.023	-0.099
	(0.369)	(0.145)
Ambulance arrival proportion	-0.024	0.021
	(0.097)	(0.075)
Waiting time 15-30 min	0.596***	0.711***
	(0.101)	(0.061)
Waiting time 30-60 min	0.760***	1.052***
	(0.135)	(0.077)
Waiting time 60-120 min	0.746***	1.156***
	(0.158)	(0.096)
Waiting time $>$ 120 min	0.602*	1.375***
	(0.322)	(0.192)
Fixed effects:		
Clock hour	Included	Included
Physician	Included	Included
Day-of-week	Included	Included
Observations	7,845	10,521
Kleibergen-Paap rk Wald F statistic	177.8	45.29

Notes: Estimated by LIML. All estimates control for physician fixed effects, day-of-the-week fixed effects and clock-hour fixed effects. The results for the share of patients in each CCS are omitted from the table due to size constraint. Robust standard errors clustered by physicians in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

(CCS $\times$ triage codes and CCC) and alternative levels of aggregation (by shift versus 30-minute block). In Appendix EC.1.3, we estimate the switch cost at the fast track of ED B. Although the patient mix differs between regular and fast tracks, both samples exhibit sizeable and statistically significant switch cost.

5.2. Effect Heterogeneity

Besides the main effect, we also estimate heterogeneous switch costs across patients and other environmental factors. Specifically, we examine the moderating effects of the following factors: patient age, triage level, arrival by ambulance, and day-of-week. We modify Equation (1) with an

additional term interacting with the switching frequency with one of the moderators once at a time. The new regression model is given by

$$\mathrm{PPH}_{it} = \delta \cdot \mathrm{SF}_{it} + \delta_w \cdot (\mathrm{SF}_{it} \times W_{it}) + \mathbf{x}_{it}^{\top}\beta + \mathbf{F}_i + \mathbf{D}_t + u_{it}, \tag{4}$$

where W_{it} is the moderator of interest. Other covariates and fixed effects remain the same.

Table 3 Effect Heterogeneity

Variables	ED A	ED B
Panel A		
Switching frequency	-4.157***	-4.674***
	(1.050)	(0.989)
Switching frequency× average age	0.025	0.011
	(0.061)	(0.061)
Kleibergen-Paap rk Wald F statistic	8.081	17.14
Panel B		
Switching frequency	-3.967***	-4.636***
	(1.034)	(1.020)
Switching frequency× average triage	-2.578	-0.747
	(2.093)	(1.718)
Kleibergen-Paap rk Wald F statistic	16.97	10.98
Panel C		
Switching frequency	-4.364***	-4.348**
	(1.145)	(1.360)
Switching frequency× ambulance arrival	2.350	-0.848
	(2.706)	(2.360)
Kleibergen-Paap rk Wald F statistic	8.680	11.88
Panel D		
Switching frequency	-4.176***	-4.750***
	(1.136)	(1.107)
Switching frequency × visit on weekend	0.127	0.260
	(1.219)	(1.534)
Kleibergen-Paap rk Wald F statistic	65.45	21.28
Fixed effects:		
Clock hour	Included	Included
Physician	Included	Included
Day-of-week	Included	Included
Observations	7,845	10,521

Notes: Estimated by LIML. All estimates control for explanatory variables $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ and fixed effects $\mathbf{F}_i$ and $\mathbf{D}_t$. The omitted ones from the table are due to size constraints. In panels A and B, we demean average age and triage level to zero mean. In panels C and D, we explore how switch cost varies with ambulance arrival proportions and with weekdays/weekends. Robust standard errors clustered by physician in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

We estimate Equation (4) by LIML using SL_{it} and $\mathrm{SL}_{it} \times W_{it}$ as IVs for SF_{it} and $\mathrm{SF}_{it} \times W_{it}$, respectively. The estimated δ and δ_w are reported in Table 3. It is worth noting that we control

for the main effects of other moderators in each regression. For example, when we estimate the heterogeneous switch costs in age, other factors are controlled as unchanged. As such, the effect heterogeneity along different moderators does not confound each other. We find that almost all interactive terms are small and statistically insignificant at both EDs. Therefore, the switch cost exhibits little variation across patient characteristics and the ED environment.

Furthermore, we explore the heterogeneity of switch costs across individual physicians. For this analysis, we concentrate on the top five physicians at each ED, ranked by the total number of patients seen. Together, these top five physicians account for 33.1% and 20.9% of the samples at the two EDs, respectively. Figure EC.2 in Appendix EC.1.4 displays estimates of δ from Equation (2), using the clean periods of each top five physician. The productivity of most physicians is adversely affected by task switching, with the magnitude of this impact varying considerably among these physicians.

5.3. Estimation by CCS Pairs: Identifying Costly Switches

Thus far, we have assumed that all switches have the same impact on PPH, and we estimate the aggregated switch cost on physicians' PPH in Equation (1). However, the magnitude of switch costs may vary for different task combinations. This is in line with our observations from physician collaborators, who have noted that certain CCS pairs are more similar than others in terms of treatment approaches and required resources. For instance, the switch cost between Respiratory and OB-GYN (Obstetrics/Gynecology) is larger than that between Gastrointestinal and Genitourinary, as the latter pair requires similar equipment and treatment methods compared to the former.

To estimate such heterogeneous switch costs, we utilize the CCS of patients before and after each switch to construct *switch pairs*. We allow task switching to be asymmetric, i.e., switching from one CCS to another can be more or less costly than switching in the reverse direction. In total, we obtain $13 \times 12 = 156$ distinct switch pairs (permutations of 13 CCSs) at the two EDs, respectively. Here we cluster three most scarce CCS types into one category. We then estimate the following model:

$$\mathrm{PPH}_{it} = \sum_{m,n} \delta_{m,n} \cdot \mathrm{SF}_{it}^{m,n} + \mathbf{x}_{it}^{\top}\beta + \mathbf{F}_i + \mathbf{D}_t + u_{it}, \tag{5}$$

where m and n denote the CCS of the preceding and the focal patients, respectively. The variable $\mathrm{SF}_{it}^{m,n}$ represents the frequency of switching between pair (m,n) in block t of physician-shift i. The parameter of interest, $\delta_{m,n}$, represents the switch cost associated with CCS pair (m,n). A negative $\delta_{m,n}$ represents a reduction in PPH due to task switching, and we assume $\delta_{m,m} = 0$.

Our estimation leads to three primary conclusions. First, among the statistically significant pairs $\delta_{m,n}$ (5% significance level), most are negative and only one to two pairs are positively significant. Second, a considerable number of switch pairs exhibit small switch costs, as the estimates are

not statistically distinguishable from zero. Third, among pairs that do exhibit significant switch costs, the magnitudes vary considerably across different pairs. In Appendix EC.1.5, we report the estimates of all significant pairs at both EDs.

6. A Data-driven Approach for Queue Management in ED

6.1. Queue Management and Max Bisection

The empirical results presented in Section 5 indicate that task switching negatively impacts physician productivity (PPH), and that this impact is heterogeneous, with the magnitude varying based on the CCS pairs before and after each switch. These findings inspire us to redesign the queueing system in the ED to mitigate the impact of task switching. Specifically, we explore the idea of dividing patients in the ED into two queues and assigning each ED physician to one of the queues. Physicians will only see patients from a different queue when their current queue is empty. Utilizing a data-driven method, we search for the optimal partition of patients into the two queues to minimize the total cost from intra-queue switching and maximize the total cost from inter-queue switching. This approach will substantially reduce the expected switch cost because inter-queue task switching occurs much less frequently than intra-queue task switching, according to our priority rule.

Several EDs have adopted a two-queue system. For instance, Song et al. (2015) studied the ED of a Kaiser Permanente Medical Center and demonstrated that patients' length-of-stay in the ED was significantly reduced in a two-queue system compared to a single-queue system. Our collaborating ED B also divides the ED into a fast track and a regular track, routing patients primarily by their triage acuity scores. Each physician is assigned to one of the tracks and prioritizes patients within the same track. The administrator of that ED is optimistic to incorporating additional patient attributes, such as CCS, into the criteria that are used to partition the patients into two tracks. Consequently, the method we study here has the potential for implementation in our collaborating ED.

We propose a data-driven method to partition patients into two queues based on their CCS, in accordance with the findings from our empirical results in Section 5.3, which indicate that CCS pairs are predictors for the magnitude of switch costs. This method can be easily customized to incorporate other patient attributes, such as the combination of triage acuity scores, age, and chief complaint categories. Furthermore, we stipulate that after partitioning, the two queues should have equal visit volumes, aligning with the current practice wherein the total visits in the fast track and regular track are comparable (i.e., 48% vs. 52%). The method can also be extended to partitioning patients into subsets of varying sizes, though the procedures will become more complex.

The main idea of our method is to construct a weighted graph, where a vertex represents each patient visit, and the switch cost from one visit to another is represented by the cost (weight) on

the associated directed arc. In this setup, maximizing the total inter-queue switch cost is equivalent to finding a max-bisection on the weighted graph.

We formulate a weighted undirected graph $G = (V, E, \omega(\cdot,\cdot))$, where each vertex $i \in V$ represents a unit mass of patients with the same CCS (i.e., 300 patient visits with a CCS in our case). Let $m(i)$ denote the index of the CCS associated with vertex i. We first calculate the following value for any unordered pair of vertices i, j based on the average switch cost for both directions:

$$\omega'(i,j) = \frac{1}{2}(-\widehat{\delta}_{m(i),m(j)} - \widehat{\delta}_{m(j),m(i)}), \tag{6}$$

where $\widehat{\delta}_{m(i),m(j)}$ represents the reduction in PPH led by a task switching from CCS $m(i)$ to $m(j)$; see Equation (5) for the interpretation of the coefficient $\widehat{\delta}_{m,n}$. Note that when $m(i) = m(j)$, that is, two vertices belong to the same CCS group, it follows from $\widehat{\delta}_{m,m}$ that $\omega'(i,j) = 0$. Also, if the switch cost in both directions is not significantly different from zero, we set that $\omega'(i,j) = 0$. Since the significant estimates include a few positive ones (see Table EC.4) which may result in a negative $\omega'(i,j)$, we devise the weight (cost) of the graph G, $\omega(i,j)$, such that all pairs of vertices are equipped with non-negative costs. Concretely, we obtain $\omega(i,j)$ by shifting $\omega'(i,j)$ upward:

$$\omega(i,j) = \omega'(i,j) + \left\lceil \max\{|\text{all negative } \omega'(i,j)|\} \right\rceil \tag{7}$$

We further note that shifting all weights by a constant does not change the optimal solution when solving the problem. We then define the edge set as $E := \{\{i,j\} |\, \omega(i,j)\}$. As an illustrative diagram, Figure 4 presents a part of the graph with four CCS types. The dotted lines represent the edges connecting with other CCS types.

We aim to search for a bisection of the patient population into two queues so that the expected intra-queue switch cost is minimized. To formulate the intra-queue switch cost, we need to calculate the frequency for each pair of patients consecutively picked by the same doctor in a shift. However, we lack data on how patients with different CCSs would be sequenced under the newly proposed two-queue system in the ED. For tractability, we assume that in a doctor's shift, a sequence of patients (with random length) will be uniformly sampled from the population, regardless of their CCS. In other words, each patient visit is added to a sequence with an equal probability. Consequently, any patient pair will appear in a sequence with equal probability. This enables us to formulate the problem of minimizing the expected intra-queue switch cost (or equivalently, maximizing the inter-queue switch cost) as an integer programming (IP). See the subsequent Proposition. The proof is attached in Section EC.1.6.

Proposition 1 *Suppose the sequences of patients seen in each shift are uniformly sampled from the population. Then the bisection that maximizes the total inter-queue switch cost is a solution to the following IP:*

$$\begin{aligned} \max_{(i,j)\in\tilde{E}} \quad & \frac{1}{2}\sum (1-y_i y_j)\,\omega(i,j) \\ \text{subject to} \quad & \sum_{i<j} y_i y_j \le -n/2, \\ & y_i \in \{-1,1\}, \quad \forall i. \end{aligned} \tag{8}$$

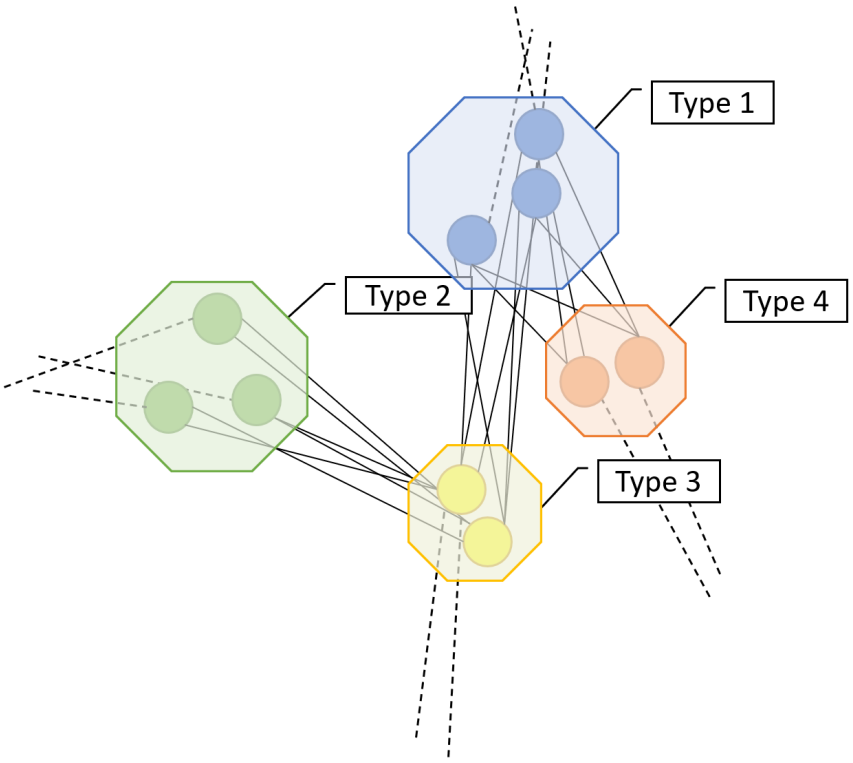


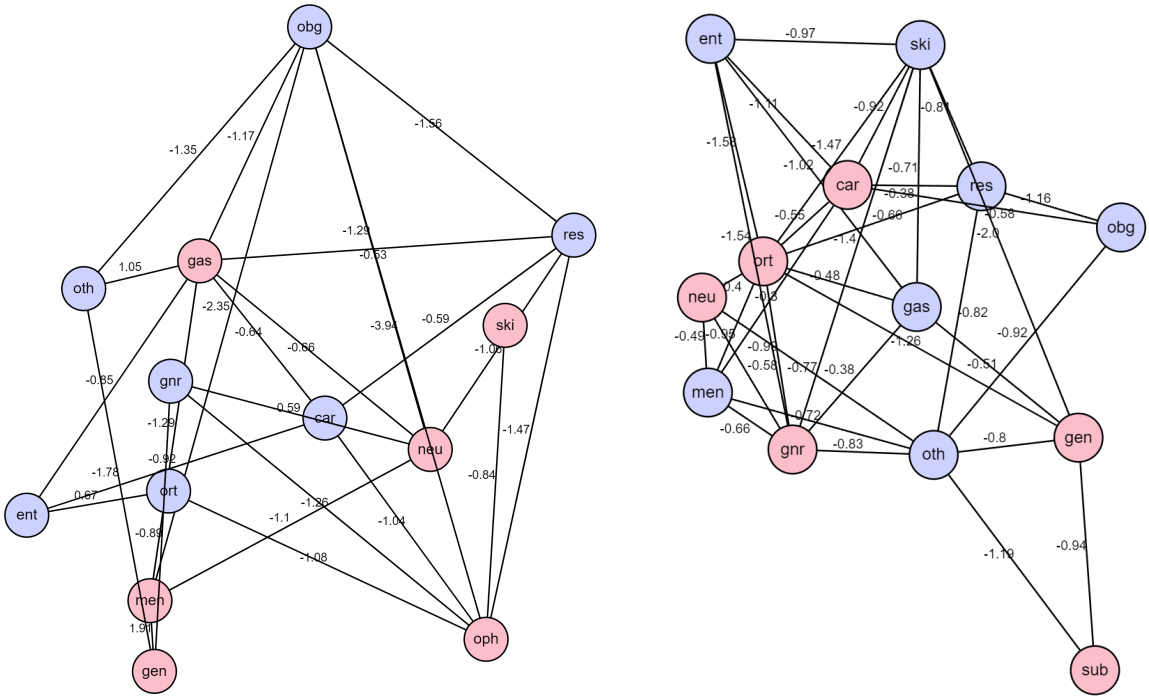
Figure 4 Graph of Patient Types

We apply the Frize and Jerrum approach (Frieze and Jerrum 1997) to solve the max-bisection, which is built on the SDP relaxation method (Goemans and Williamson 1995). The optimization computation is performed through Julia (Garstka et al. 2021). The SDP relaxation of Equation (8) is as follows.

$$\begin{aligned} \max \quad & \frac{1}{2}\sum_{i<j} w_{ij}\,(1-v_i\cdot v_j) \\ \text{subject to} \quad & \sum_{i<j} v_i\cdot v_j \le -n/2 \\ & v_j \in \mathbf{S}_n, \quad \forall j \in V \end{aligned} \tag{9}$$

where V is the set of all vertices in the graph G and $\mathbf{S}_n = \{x \in \mathbb{R}^n : |x| = 1\}$ stands for the unit sphere in n dimensions.

We present the obtained bisection results using different colors in Figure 5 for ED A and B, respectively. The values associated with each edge represent the eq. (6) from original estimates. For



(a) ED A (b) ED B

Figure 5 CCS Graph Max-Bisection Results

ED A, we assign six CCS types to one queue, including mental (MEN), skin (SKI), genitourinary (GEN), neurology (NEU), ophthalmology (OPH), gastrointestinal (GAS), and assign all else to the other one. For ED B, we obtain six CCS types in one queue using similar method, which are: cardiovascular (CAR), orthopedics (ORT), neurology (NEU), general and minor (GNR), genitourinary (GEN), substance misuse (SUB), and all other CCS types belong to the other queue.

6.2. Simulation under Two-Queue ED System

We conduct a simulation study to demonstrate how our proposed two-queue system enhances ED efficiency. In the counterfactual scenario, we assume that doctors in each shift will be dedicated to serving one queue and will only serve another queue when the assigned queue is empty. This priority rule is also used in our collaborating ED B with a regular and fast track.

To mimic the actual system, we need to define the patient selection rule in the simulated scenario. At a high level, we aim to preserve the order of patients being selected as in the original scenario so that the difference between the simulation and the original scenarios can be solely attributed to the patient clustering practice. This can be accomplished in most cases where physicians face the original choice set. In the counterfactual scenario, we allow the physician to pick the same patient as they did in the original scenario. However, due to the introduction of the two-queue system and the modification of the PPH, the choice sets seen by the doctor may differ from the original ones.

Consequently, we determine the choice using the revealed preference as follows. Suppose we observe a physician selecting patient A from a choice set $A, B, C, \cdots$. We can infer that patient A has a higher priority over others, i.e., $A \succ B$, $A \succ C$... We then obtain the closure of this partial order by transitivity, e.g., if patient B is later selected from a new choice set B, C, D, we infer $B \succ D$, and thus $A \succ D$. For the remaining patients, their relative order is imputed by first-come-first-serve to minimize modifications to the choice set.

Next, we reconstruct the initial assessment time for each patient according to the original data, while factoring in the impact of task switching under the simulated patient service sequences; see Section EC.2 for the details of the reconstruction procedure. The simulation analysis is performed for the busy periods of ED operations, specifically from 10 am to 6 pm for ED A and from 10 am to 10 pm for ED B. We simulate 58,531 patient visits in 4,347 shifts in ED A, and 83,149 patient visits in 6,938 shifts in ED B.

Table 4 Two-Queue Simulation Results

Variables	ED A	ED B
Panel A. Actual Scenario		
PPH	3.381	2.898
Waiting time	44.10	43.78
Waiting census	4.682	4.471
Panel B. Simulated Scenario		
PPH	3.491	3.077
Waiting time	27.27	26.42
Waiting census	3.275	3.075
Panel C. Differences		
Δ PPH	0.110***	0.179***
Δ Waiting time	-16.83***	-17.36***
Δ Waiting census	-1.407***	-1.397***
Patient visits	58,531	83,149
Physician shifts	4,347	6,938

Notes: Panel A reports the average PPH, waiting time, and waiting census in the actual scenario, and Panel B reports the counterfactual scenario values. Panel C summarizes the mean differences and reports the two-sample t-tests significance (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

Table 4 presents the results of the simulation study. As the baseline, panel A displays the average PPH, patient waiting time, and waiting census (number of patients counted every 10 minutes) in the actual scenario. Panel B shows the same metrics in the simulated scenario, and panel C reports

the mean differences between panels A and B, along with the significance under two-sample t-tests. We find significant efficiency gains from clustering patients into two queues. Across the two EDs, the average PPH is increased by 0.11 and 0.179, which equal 3.25% and 6.19% of the original levels and are statistically significant. Furthermore, the average waiting time is reduced by 16.83 and 17.36 minutes (38.16% and 39.66%), and the average waiting census by 1.407 and 1.397 (30.04% and 31.23%), respectively. All these differences are statistically significant. Additionally, we also find that the proportion of patients affected by task switching will further decrease from 85% to 76.7% - 78.7%. As a result, we conclude that the impact of task switching on the ED system, as measured by waiting time and the waiting census, is roughly seven times greater than that on the physician level measured by PPH. This is because any faster (or slower) treatment will have a ripple effect on multiple subsequent patients waiting to be treated.

In summary, our simulation study demonstrates the performance improvement achieved by implementing a two-queue system in the studied EDs. Under our data-driven queue management method, we are able to alleviate the switch cost on physician productivity. Although the improvement in individual physicians' productivity is moderate, the ripple effect on the ED system is seven times larger. Therefore, mitigating the impact of task switches will significantly reduce patient waiting time and ED congestion.

7. Effects of Task Switching on Quality and Routing

7.1. Quality of Care

We have uncovered a significant switch cost on the efficiency of ED physicians. Consequently, it is natural to consider that task switching could also impact the quality of their work. We utilize 7/30-day RAD rates and inpatient admissions as proxies for this. Therefore, we consider the following linear model

$$r_{it} = \delta \cdot s_{it} + \mathbf{x}_{it}^{\top}\beta_1 + \mathbf{z}_{it}^{\top}\beta_2 + \mathbf{F}_i + \mathbf{D}_t + u_{it}, \tag{10}$$

where r_{it} represents the share of revisit and readmission in 7 and 30 days, and the share of inpatient admission upon the focal visit among patients selected in block (i,t). We still include the explanatory variables $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$, and fixed effects $\mathbf{F}_i$ and $\mathbf{D}_t$, as in Equation (1).

As demonstrated in Panels A and B of Table 5, task switching has minimal and statistically insignificant impacts on both RAD shares. Furthermore, we find little connection between inpatient admission and switcher proportion (Panel C).

Nonetheless, we must exercise caution when interpreting the aforementioned estimates. If physicians are aware of the switch cost, they may counteract the adverse effects on quality by exerting additional effort and allocating extra time for diagnosis and other remedial procedures. As such, our estimates are better understood as the outcome of physicians' trade-off between efficiency and quality in the presence of the switch cost.

Table 5 Effect of Task Switching on Quality of Care

	ED A	ED B
Panel A. RAD in 7 days		
Switching frequency	-0.0011	-0.0031
	(0.0060)	(0.0044)
Panel B. RAD in 30 days		
Switching frequency	-0.0095	0.0075
	(0.0075)	(0.0071)
Panel C. Inpatient admission		
Switching frequency	0.0079	-0.0087
	(0.0112)	(0.0119)
Fixed effects:		
Clock hour	Included	Included
Physician	Included	Included
Day-of-week	Included	Included
Observations	7,845	10,521

Notes: Estimated by fixed effects OLS. Estimates of other regressors in Equation (10) are omitted from the table due to size constraint. Robust standard errors clustered by physicians reported in parentheses (* $p<0.1$, ** $p<0.05$, *** $p<0.01$).

7.2. Patient Routing

We explore whether physicians consider the switch cost when selecting patients. Ding et al. (2019) have demonstrated that physicians' routing decisions exhibit an effort to minimize the average waiting cost of patients, specifically by prioritizing patients with more severe clinical conditions and longer waiting times. Furthermore, KC et al. (2020) find that physicians tend to choose patients with shorter expected processing times. Given that task switching reduces PPH and negatively impacts physician productivity, physicians might also prioritize patients of the same type as their preceding patients.

To explore physicians' choice patterns, we adopt a conditional logistic model. Instead of using 30-min blocks, we focus on the individual patient selection process. Suppose the physician's utility from selecting patient j is given by $U_j = Z_j^\top \zeta + D_j \pi + v_j$, which depends on the patient's characteristics, Z_j, whether task switching is required, D_j, and an unobserved error v_j following the type-I extreme value distribution (McFadden 1973). As such, the probability of choosing patient j from the choice set $\mathcal{J}(i,t)$ is given by

$$\Pr\{\text{selecting } j\} = \frac{\exp\left(Z_j^\top \zeta + D_j \pi\right)}{\sum_{k \in \mathcal{J}(i,t)} \exp\left(Z_k^\top \zeta + D_k \pi\right)}. \tag{11}$$

The coefficient of interest, π captures the impact of task switching on the choice probability. If $\pi < 0$, patients with a switched type will be selected with lower priority. Following Ding et al.

(2019), we let Z_j include a triage-specific, piecewise linear function of waiting time. The intuition is that the marginal waiting cost differs by triage levels and whether the waiting time has exceeded the target physicians want to achieve.

Table 6 Effect of Task Switching on Patient Routing

Variables	ED A	ED B
Different type	0.871***	0.974*
	(0.018)	(0.015)
Triage 2	0.031***	0.009***
	(0.018)	(0.001)
Triage 3	0.024***	0.007***
	(0.014)	(0.001)
Triage 4	0.054***	0.018***
	(0.028)	(0.002)
Triage 5	0.140***	0.090***
	(0.076)	(0.015)
Waiting time × triage 1	1.127**	0.997
	(0.061)	(0.002)
Waiting time × triage 2	1.398***	1.352***
	(0.014)	(0.010)
Waiting time × triage 3	1.148***	1.143***
	(0.003)	(0.002)
Waiting time × triage 4	1.077***	1.062***
	(0.002)	(0.002)
Waiting time × triage 5	1.039***	1.023***
	(0.002)	(0.002)
Observations	340,258	406,521
Pseudo R^2	0.25	0.23

Notes: Conditional logit models estimated by maximum likelihood. Dependent variable: indicator of selected patients. The estimates of age, gender, homeless, arrival mode, and patient CCS are omitted from the table due to size constraint. Coefficients converted to odds ratios. Robust standard errors clustered by physicians in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

We estimate coefficients in Equation (11) using maximum likelihood and cluster standard errors by physicians. Table 6 presents the estimates converted to odds ratios (exponential of the original estimates). We discover that the effects of task switching on choice probability are negative and statistically significant at both EDs, corresponding to odds ratios between 0.871 and 0.974. Holding everything else constant, the odds of selecting a patient of a different type are 87.1% – 97.4% of the odds of selecting a patient of the same type. The results suggest that physicians exhibit

switch-aversion when choosing patients. During the selection process, they prefer patients with the same characteristics as the previously picked patient. This finding also supports our use of IV, as physicians' strategic patient selection contributes to sample selection bias.

8. Managerial Implications

We summarize the managerial insights from our study to assist ED managers and healthcare practitioners in enhancing operational efficiency.

First, our study examines task switching as a prevalent phenomenon in emergency departments and identifies a substantial switch cost on the efficiency of ED physicians. We find that at different EDs, a 10% increase in switching frequency across patient CCS types reduces the average PPH by 8.65% - 11.53%. However, we do not find evidence that task switching affects the quality of care, which might be due to the limited methods of measuring care quality. While efforts to reduce task switching may offer significant value, researchers and ED managers should remain aware of the potential risk of compromised care quality. Further research is necessary to explore the impact of task switching on care quality.

In reality, it is impossible to eliminate task switching due to the complexity of ED activities and the diversity of ED patients. However, several strategies can be employed to mitigate the impact of task switching. In addition to the proposed data-driven method of partitioning patients into different queues, we can introduce the pod system to ED workers (Valentine and Edmondson 2015, Gavin and Peterson 2017). This arrangement can enhance inter-professional communication efficiency and alleviate potential delays caused by task switching. For instance, if physicians encounter a new patient type and require information, they will have quick access to nearby information resources. Furthermore, the computer system can highlight waiting patients of the same type as the focal patient to help physicians avoid task switching. Ultimately, physicians must carefully balance the switch cost against other factors (e.g., acuteness, waiting time) when making patient prioritization decisions.

Moreover, our empirical evidence reveals that ED physicians exhibit switch aversion. Holding everything else constant, the odds of selecting a same-type patient can be twice as high as that of selecting a different-type patient. According to our discussions with ED physicians, such switch-averse behavior is typically subconscious. If ED managers could increase physicians' awareness of the task switching problem, physicians could make better trade-offs in patient routing and enhance their service efficiency.

9. Conclusions and Future Research

Although experimental research on task switching is abundant, less is known about the impact of task switching in real workplaces. In this paper, we investigate the task switching behavior of

ED physicians to bridge this gap. Emergency departments are well-known for being extremely busy and heavily loaded most of the time. ED patients also present with heterogeneous clinical conditions. Consequently, physicians must provide a variety of medical care at a fast pace and suffer from productivity loss due to task switching. Our analysis reveals that a 10% increase in switching frequency leads to an 8.65% and 11.53% decrease in PPH on average, and physicians try to avoid task switching when selecting patients. On the flip side, task switching does not have significant impact on treatment quality.

By investigating the heterogeneity among different CCS pairs, we propose a data-driven queue management method to mitigate the switch costs in EDs. We use the classical max-bisection algorithm to partition patients into two queues based on their CCSs, minimizing the total intra-queue switch costs. The subsequent simulation shows that under the two-queue system, ED physician productivity can be improved, and as a result, the average waiting time of ED patients can be reduced by about 40%. This improvement can be achieved without adding substantial extra resources.

Our study has broad implications for occupations and industries beyond healthcare. Task switching is prevalent in both the service and production sectors of the economy, and it is ubiquitous in human lives. It is anticipated that people's work will become more complex, more flexible, and less specialized in specific tasks in the future. However, human brains may still not be prepared for intensive task switching, multitasking, or interrupted working environments (Autor et al. 2003, Acemoglu and Autor 2011, Acemoglu and Restrepo 2018). As such, the cost of task switching may become an increasingly important factor impeding workplace productivity. The negative impact of task switching also underscores the importance of managerial practices to counteract it, including alternative work arrangements, organizational structures, and technological assistance to enhance the productivity of future jobs (Bloom and Van Reenen 2011).

We conclude our paper by suggesting several directions for future research. Our paper focuses on ED physicians' task switches across patient types and the consequences on productivity. Future studies can investigate task switching within a patient's treatment but across different activities, develop alternative measures for task switching, or examine the switch cost on physicians' mental stress, work sustainability, patient satisfaction, and treatment quality. Additionally, since we do not observe physician characteristics in our data, future research could study determinants of the switch cost at the physician level. As task switching leads to fewer patients treated and reduced fee-for-service earnings in a time unit, it is crucial to quantify the monetary value of the switch cost. Future research may compare task switching, multitasking, and work interruptions in a unified setting and investigate how workers respond. Since task switching is costly, jobs requiring frequent task switching may have to pay a wage premium to compensate for the resulting disutility. As such, a hedonic price framework (Rosen 1974) will help researchers recover the monetary equivalent cost of task switching.

References

Abouee-Mehrizi H, Baron O (2016) State-Dependent M/G/1 Queueing Systems. *Queueing Systems* 82(1):121–148.

Acemoglu D, Autor D (2011) Skills, Tasks and Technologies: Implications for Employment and Earnings. *Handbook of Labor Economics*, volume 4, 1043–1171.

Acemoglu D, Restrepo P (2018) The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. *American Economic Review* 108(6):1488–1542.

Allon G, Kremer M (2019) Behavioral Foundations of Queueing Systems. Donohue K, Katok E, Leider S, eds., *The Handbook of Behavioral Operations*, 325–366 (Hoboken, NJ: John Wiley & Sons).

Allport A, Styles EA, Hsieh S (1994) Shifting Intentional Set: Exploring the Dynamic Control of Tasks. Umiltà C, Moscovitch M, eds., *Attention and Performance XV: Conscious and Nonconscious Information Processing*, 421–452 (Cambridge, MA: MIT Press).

Allport A, Wylie G (1999) Task-Switching: Positive and Negative Priming of Task-Set. Humphreys GW, Duncan JE, Treisman AE, eds., *Attention, Space, and Action: Studies in Cognitive Neuroscience*, 273–296 (New York, NY: Oxford University Press).

Ang E, Kwasnick S, Bayati M, Plambeck EL, Aratow M (2016) Accurate Emergency Department Wait Time Prediction. *Manufacturing & Service Operations Management* 18(1):141–156.

Autor DH, Levy F, Murnane RJ (2003) The Skill Content of Recent Technological Change: An Empirical Exploration. *Quarterly Journal of Economics* 118(4):1279–1333.

Avgerinos E, Gokpinar B (2018) Task Variety in Professional Service Work: When It Helps and When It Hurts. *Production and Operations Management* 27(7):1368–1389.

Batt RJ, KC DS, Staats BR, Patterson BW (2019) The Effects of Discrete Work Shifts on a Nonterminating Service System. *Production and Operations Management* 28(6):1528–1544.

Batt RJ, Terwiesch C (2016) Early Task Initiation and Other Load-Adaptive Mechanisms in the Emergency Department. *Management Science* 63(11):3531–3551.

Batt RJ, Tong JD (2020) Mean Service Metrics: Biased Quality Judgment and the Customer–Server Quality Gap. *Manufacturing & Service Operations Management* 22(5):975–995.

Bayati M, Kwasnick S, Luo D, Plambeck EL (2017) Low-Acuity Patients Delay High-Acuity Patients in an Emergency Department. *SSRN Working Paper No. 3095039* .

Beveridge R (1998) CAEP Issues. The Canadian Triage and Acuity Scale: A New and Critical Element in Health Care Reform. Canadian Association of Emergency Physicians. *Journal of Emergancy Medicine* 16(3):507–511.

Bloom N, Van Reenen J (2011) Human Resource Management and Productivity. Card D, Orley A, eds., *Handbook of Labor Economics*, volume 4, 1697–1767 (Elsevier).

Boh W, Slaughter SA, Espinosa AJ (2007) Learning from Experience in Software Development: A Multilevel Analysis. *Management Science* 53(8):1315–1331.

Cai X, Gong J, Lu Y, Zhong S (2017) Recover Overnight? Work Interruption and Worker Productivity. *Management Science* 64(8):3489–3500.

Calder L, Pozgay A, Riff S, Rothwell D, Youngson E, Mojaverian N, Cwinn A, Forster A (2015) Adverse Events in Patients with Return Emergency Department Visits. *BMJ Quality & Safety in Health Care* 24(2):142–148.

Chan DC (2016) Teamwork and Moral Hazard: Evidence from the Emergency Department. *Journal of Political Economy* 124(3):734–770.

Chan DC (2018) The Efficiency of Slacking Off: Evidence from the Emergency Department. *Econometrica* 86(3):997–1030.

Chaou CH, Chen HH, Tang P, Yen AMF, Wu KH, Hsiao CT, Chiu TF (2018) Traffic Intensity of Patients and Physicians in the Emergency Department: A Queueing Approach for Physician Utilization. *Journal of Emergency Medicine* 55(5):718–725.

Ding Y, Nagarajan M, Zhang G (2020) Parallel Queues with Discrete-Choice Arrival Pattern: Empirical Evidence and Asymptotic Characterization. *SSRN Working Paper No. 3584880* .

Ding Y, Park E, Nagarajan M, Grafstein E (2019) Patient Prioritization in Emergency Department Triage Systems: An Empirical Study of the Canadian Triage and Acuity Scale (CTAS). *Manufacturing & Service Operations Management* 21(4):713–948.

Dong J, Yom-Tov E, Yom-Tov GB (2019) The Impact of Delay Announcements on Hospital Network Coordination and Waiting Times. *Management Science* 65(5):1969–1994.

Frieze A, Jerrum M (1997) Improved approximation algorithms for maxk-cut and max bisection. *Algorithmica* 18(1):67–81.

Garstka M, Cannon M, Goulart P (2021) COSMO: A conic operator splitting method for convex conic problems. *Journal of Optimization Theory and Applications* 190(3):779–810, URL http://dx.doi.org/10.1007/s10957-021-01896-x.

Gavin N, Peterson K (2017) Team-based Pod System Reduces Lengths of Stay for Treat-and-Release Patients. *ED Management: the Monthly Update on Emergency Department Management* 29(6):67–69.

Goemans MX, Williamson DP (1995) Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)* 42(6):1115–1145.

Goes PB, Ilk N, Lin M, Zhao JL (2018) When More is Less: Field Evidence on Unintended Consequences of Multitasking. *Management Science* 64(7):3033–3054.

Gong J, Png IP (2022) Automation, specialization, and productivity: Field evidence. *Available at SSRN 3597725* .

González VM, Mark G (2004) "constant, constant, multi-tasking craziness" managing multiple working spheres. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 113–120.

Grafstein E, Bullard MJ, Warren D, Unger B, Group CNW (2008) Revision of the Canadian Emergency Department Information System (CEDIS) Presenting Complaint List Version 1.1. *Canadian Journal of Emergency Medicine* 10(2):151–161.

Gurvich I, O'Leary KJ, Wang L, Van Mieghem JA (2019) Collaboration, Interruptions, and Changeover Times: Workflow Model and Empirical Study of Hospitalist Charting. *Manufacturing & Service Operations Management* 22(4):754–774.

Hansen B (2022) *Econometrics* (Princeton University Press).

He S, Sim M, Zhang M (2019) Data-Driven Patient Scheduling in Emergency Departments: A Hybrid Robust-Stochastic Approach. *Management Science* 65(9):4123–4140.

Heckman JJ (2010) Selection Bias and Self-Selection. *Microeconometrics*, 242–266 (Springer).

Huang J, Carmeli B, Mandelbaum A (2015) Control of Patient Flow in Emergency Departments, or Multiclass Queues with Deadlines and Feedback. *Operations Research* 63(4):892–908.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary Task Ordering: Queue Management in Radiological Services. *Management Science* 64(9):4389–4407.

Jersild AT (1927) Mental Set and Shift. *Archives of Psychology* 89:5–82.

KC DS (2013) Does Multitasking Improve Performance? Evidence from the Emergency Department. *Manufacturing & Service Operations Management* 16(2):168–183.

KC DS (2019) Worker productivity in operations management. *Available at SSRN 3466947* .

KC DS, Staats BR (2012) Accumulating A Portfolio of Experience: The Effect of Focal and Related Experience on Surgeon Performance. *Manufacturing & Service Operations Management* 14(4):618–633.

KC DS, Staats BR, Kouchaki M, Gino F (2020) Task Selection and Workload: A Focus on Completing Easy Tasks Hurts Performance. *Management Science* 66(10):4397–4416.

KC DS, Terwiesch C (2009) Impact of Workload on Service Time and Patient Safety: An Econometric Analysis of Hospital Operations. *Management Science* 55(9):1486–1498.

Kuntz L, Sülz S (2013) Treatment Speed and High Load in the Emergency Department—Does Staff Quality Matter? *Health Care Management Science* 16(4):366–376.

Lindbeck A, Snower DJ (2000) Multitask Learning and the Reorganization of Work: From Tayloristic to Holistic Organization. *Journal of Labor Economics* 18(3):353–376.

Liu R, Xie X (2018) Physician Staffing for Emergency Departments with Time-Varying Demand. *INFORMS Journal on Computing* 30(3):588–607.

Mandelbaum A, Pats G (1998) State-Dependent Stochastic Networks. Part I. Approximations and Applications with Continuous Diffusion Limits. *The Annals of Applied Probability* 8(2):569–646.

Mark G, Gonzalez VM, Harris J (2005) No task left behind? examining the nature of fragmented work. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 321–330.

McFadden D (1973) Conditional Logit Analysis of Qualitative Choice Behavior. Zarembka P, ed., *Frontiers in Econometrics*, 105–142 (New York: Academic Press).

Miyata Y, Norman DA (1986) Psychological issues in support of multiple activities. *User centered system design: New perspectives on human-computer interaction* 265–284.

Monsell S (2003) Task switching. *Trends in cognitive sciences* 7(3):134–140.

Narayanan S, Balasubramanian S, Swaminathan JM (2009) A Matter of Balance: Specialization, Task Variety, and Individual Learning in a Software Maintenance Environment. *Management Science* 55(11):1861–1876.

Ong P, Png IP (2021) Automation, deskilling, and labor supply: Empirical evidence. *Available at SSRN 3452464* .

Ouyang H, Liu R, Sun Z (2021) Emergency Department Modeling and Staffing: Time-Varying Physician Productivity. *SSRN Working Paper No. 3963226* .

Rogers RD, Monsell S (1995) Costs of a Predictable Switch between Simple Cognitive Tasks. *Journal of Experimental Psychology: General* 124(2):207–231.

Rosen S (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy* 82(1):34–55.

Roth A, Tucker AL, Venkataraman S, Chilingerian J (2019) Being on the Productivity Frontier: Identifying "Triple Aim Performance" Hospitals. *Production and Operations Management* 28(9):2165–2183.

Rubinstein JS, Meyer DE, Evans JE (2001) Executive Control of Cognitive Processes in Task Switching. *Journal of Experimental Psychology: Human Perception and Performance* 27(4):763–797.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2014) Complexity-Augmented Triage: A Tool for Improving Patient Safety and Operational Efficiency. *Manufacturing & Service Operations Management* 16(3):329–345.

Schilling MA, Vidal P, Ployhart RE, Marangoni A (2003) Learning by Doing Something Else: Variation, Relatedness, and the Learning Curve. *Management Science* 49(1):39–56.

Silver D (2020) Haste or Waste? Peer Pressure and the Distribution of Marginal Returns to Health Care. *Review of Economic Studies* forthcoming.

Song H, Tucker AL, Graue R, Moravick S, Yang JJ (2020) Capacity Pooling in Hospitals: The Hidden Consequences of Off-Service Placement. *Management Science* 66(9):3825–3842.

Song H, Tucker AL, Murrell KL (2015) The Diseconomies of Queue Pooling: An Empirical Investigation of Emergency Department Length of Stay. *Management Science* 61(12):3032–3053.

Song H, Veeraraghavan S (2018) Quality of Care. *Handbook of Healthcare Analytics: Theoretical Minimum for Conducting 21st Century Research on Healthcare Operations* 79–108.

Staats BR, Gino F (2012) Specialization and Variety in Repetitive Tasks: Evidence from a Japanese Bank. *Management Science* 58(6):1141–1159.

Valentine MA, Edmondson AC (2015) Team Scaffolds: How Mesolevel Structures Enable Role-Based Coordination in Temporary Groups. *Organization Science* 26(2):405–422.

Wang Y, Ding Y, Park E, Hunte G (2019) Do Financial Incentives Change Length-of-stay Performance in Emergency Departments? A Retrospective Study of the Pay-for-performance Program in Metro Vancouver. *Academic Emergency Medicine* 26(8):856–866.

Yom-Tov GB, Mandelbaum A (2014) Erlang-R: A Time-Varying Queue with Reentrant Customers, in Support of Healthcare Staffing. *Manufacturing & Service Operations Management* 16(2):283–299.

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

Online Appendix

EC.1. Additional Tables and Figures

EC.1.1. OLS Estimates

Table EC.1 reports the OLS estimates of Equation (1). We find that the OLS estimates from two EDs both underestimate the true switch cost.

Table EC.1 Effect of Task Switching on PPH: OLS Estimates

	ED A	ED B
Switching frequency	-0.162	0.073
	(0.102)	(0.056)
Average queue length	0.088***	0.032***
	(0.015)	(0.011)
Average ED load	0.013*	0.026***
	(0.007)	(0.004)
Average age	-0.002	-0.002
	(0.001)	(0.001)
Average triage	0.435***	0.096**
	(0.063)	(0.041)
Female proportion	0.028	-0.021
	(0.060)	(0.043)
Homeless proportion	0.216	-0.020
	(0.290)	(0.118)
Ambulance arrival proportion	-0.032	-0.013
	(0.088)	(0.052)
Waiting time 15-30 min	0.583***	0.661***
	(0.082)	(0.044)
Waiting time 30-60 min	0.844***	0.972***
	(0.117)	(0.063)
Waiting time 60-120 min	0.846***	1.114***
	(0.165)	(0.083)
Waiting time $>$ 120 min	0.859**	1.216***
	(0.323)	(0.167)
Fixed effects:		
Clock hour	Included	Included
Physician	Included	Included
Day-of-week	Included	Included
Observations	7,845	10,521
R^2	0.196	0.142

Notes: Estimated by OLS. All estimates control for physician fixed effects, day-of-the-week fixed effects, clock-hour fixed effects. We also control for the share of patients in each CCS which is omitted from the table due to size constraint. Robust standard errors clustered by physicians in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

EC.1.2. Robustness Checks

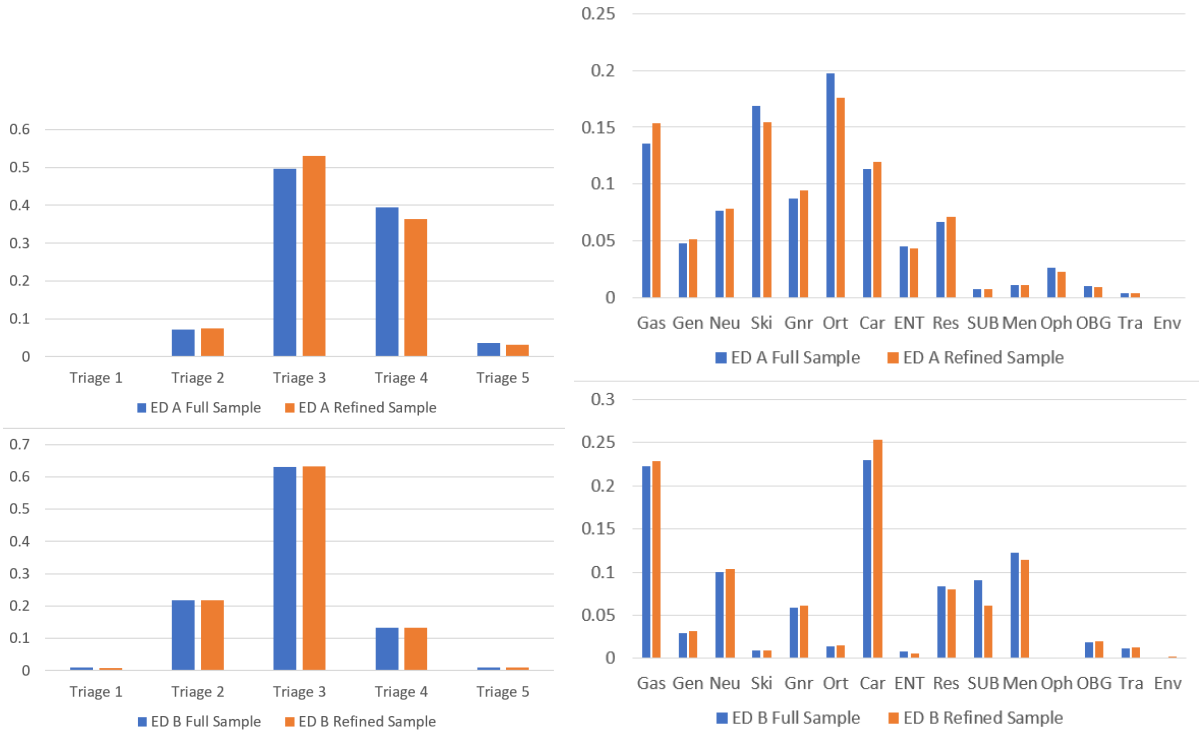
We first examine the distribution of triage codes, CCS, gender and age in the final sample and the entire sample (for ED B it is the regular track), respectively. We find that the distribution is similar to that of the entire sample as reflected in Figure EC.1. There is a slight difference in the triage distribution, as physicians tend to pick more acute patients during the first hours. However, since our patient type is defined based on CCS, the distribution among the patients picked in the clean period is almost identical to that of the entire patient population. Consequently, the patients used in our study are representative of ED operations.

To check the representativeness of our refined sample for ED operations, we examine the distribution of patient demographics, including triage, CCS, gender, and age, in both the refined sample and the entire sample. We present their distribution histograms in Figure EC.1. Our analysis reveals that the distribution of patient demographics is similar between the two samples. There is a slight difference in the triage distribution, as physicians tend to select more acute patients during the first hours. However, since our patient type is now defined based on CCS, the distribution among the patients picked during the clean period is almost identical to that of the entire patient population. As a result, the patients used in our study are representative of ED operations.

Next, we examine the robustness of the switch cost to alternative definitions of patient types and an integrated sample period format (instead of 30-min blocks). The results are presented in Table EC.2.

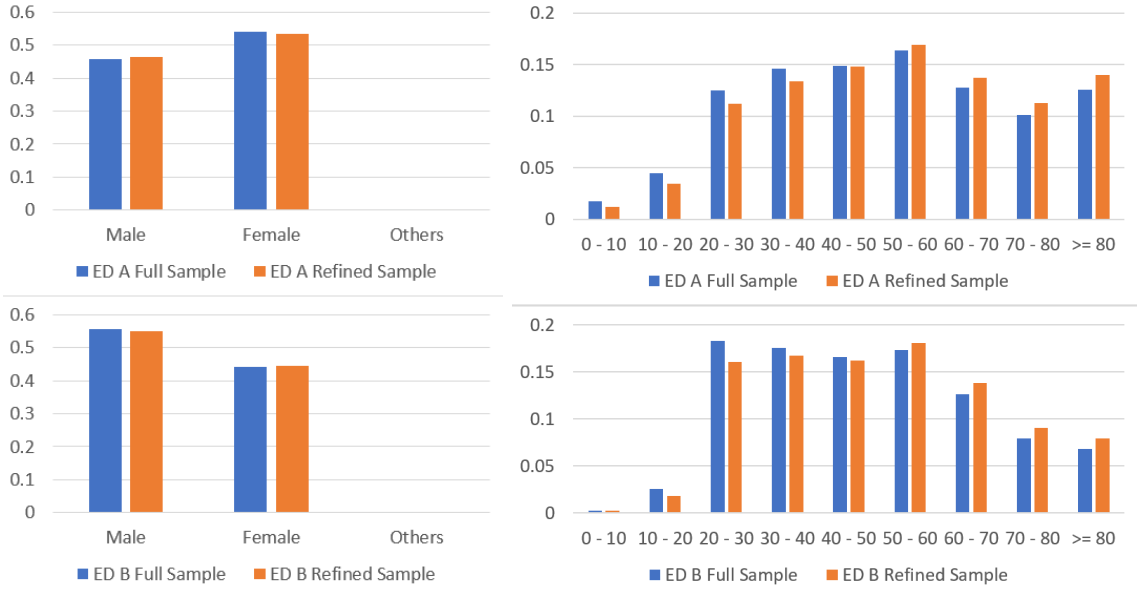
In the preferred specification, patient types are defined based on their CCS. Alternatively, we define patient types as (i) both distinct CCSs and triage levels, and (ii) distinct chief complaint categories (CCC). Both definitions are more granular, as there are 15 unique CC systems, 68 unique CCS-triage combinations, and 163 unique CC categories. With more granular classification, the switching frequency and switch likelihood of physician service also increases. In panel A and panel B of Table 3, we find that the LIML estimates of the switch cost δ remain negatively significant. The effect magnitudes are very similar to those from the main model. The results show the robustness of the switch cost under different patient type definitions.

In addition, we calculate PPH for each physician-shift over the entire clean period, i.e., from the beginning of the shift until when the first lab test comes out. As before, we construct the switch likelihood at the beginning of the shift. The sample size is then reduced to 2,754 for ED A and 4,141 for ED B. In Table EC.2, panel C, we find that switching frequency is still negative and significant. The results support our strategy of restricting the sample period using the earliest lab result time.



(a) Triage Distribution

(b) CCS Distribution



(c) Gender Distribution

(d) Age Distribution

Figure EC.1 Distribution of Patient Demographics in the Refined Sample vs. Entire Sample

Table EC.2 Alternative Patient Type Definition and Integrated Sample Method

Variables	ED A	ED B
Panel A. Patient type by CCS × triage		
Switching frequency	-3.050***	-2.443***
	(0.841)	(0.670)
Average queue length	0.087***	0.033**
	(0.013)	(0.010)
Average ED load	0.013	0.025***
	(0.007)	(0.004)
Average age	-0.002	-0.002
	(0.002)	(0.001)
Average triage	0.399***	0.097*
	(0.066)	(0.041)
Observation	7,845	10,521
Kleibergen-Paap rk Wald F statistic	145.5	83.06
Panel B. Patient type by CC category		
Switching frequency	-3.387***	-3.814***
	(0.922)	(0.785)
Average queue length	0.103***	0.048***
	(0.013)	(0.008)
Average ED load	0.027***	0.024***
	(0.007)	(0.004)
Average age	-0.002	-0.001
	(0.001)	(0.001)
Average triage	0.469***	0.175***
	(0.066)	(0.036)
Observation	7,845	10,521
Kleibergen-Paap rk Wald F statistic	174.4	102.6
Panel C. Integrated sample period		
Switching frequency	-3.074***	-7.442***
	(0.819)	(2.127)
Average queue length	0.114***	0.024
	(0.020)	(0.016)
Average ED load	-0.003	0.023**
	(0.010)	(0.007)
Average age	-0.009**	0.002
	(0.003)	(0.004)
Average triage	0.820***	-0.047
	(0.137)	(0.129)
Observation	2,754	4,141
Kleibergen-Paap rk Wald F statistic	121.4	20.07
Fixed effects:		
Clock hour	Included	Included
Physician	Included	Included
Day-of-week	Included	Included

Notes: Estimated by LIML. All estimates control for explanatory variables $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ and fixed effects $\mathbf{F}_i$ and $\mathbf{D}_t$. Other regressors are omitted from the table due to size constraints. Robust standard errors clustered by physician in parentheses (* $p<0.1$, ** $p<0.05$, *** $p<0.01$).

EC.1.3. Switch Cost in the Fast Track

For ED B operating both regular and fast tracks, we also explore the effect of task switching in the fast track using the same empirical strategy. As before, we focus on the clean period obtained by the same process of each fast-track shift and exclude potential idle periods. Consistent with the previous findings, the estimated switch cost at the fast track remains negative and significant. The results are presented in Table EC.3.

Table EC.3 Effect of Task Switching on PPH in Fast-track of ED B

Variable	Estimates
Switching frequency	-10.15***
	(1.691)
Average queue length	0.083***
	(0.014)
Average ED load	-0.009
	(0.020)
Average age	-0.003
	(0.003)
Average triage	0.313***
	(0.094)
Female proportion	0.034
	(0.133)
Homeless proportion	-0.309
	(0.293)
Ambulance arrival proportion	0.114
	(0.214)
Waiting time 15-30 min	0.344
	(0.283)
Waiting time 30-60 min	0.413
	(0.272)
Waiting time 60-120 min	0.178
	(0.281)
Waiting time $>$ 120 min	0.252
	(0.371)
Fixed effects:	
Clock hour	Included
Physician	Included
Day-of-week	Included
Observations	5,604
Kleibergen-Paap rk Wald F statistic	46.84

Notes: Estimated by LIML. All estimates control for physician fixed effects, day-of-the-week fixed effects and clock-hour fixed effects. The results for the share of patients in each CCS are omitted from the table due to size constraint. Robust standard errors clustered by physicians in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

EC.1.4. Effect Heterogeneity across Physicians

In this appendix, we investigate the switch cost heterogeneity across individual physicians. To obtain physician-specific estimates of the switch cost δ, we split the sample by physician ID and estimate Equation (2) separately on each subsample. We focus on top 10 physicians at each ED to ensure there are enough observations to identify the regression parameters.

Figure EC.2 plots the resulting estimates at both EDs, with physicians ranked by the number treated of patients. Almost all physicians' productivity is negatively impacted by task switching, and the effect magnitudes differ significantly across physicians.

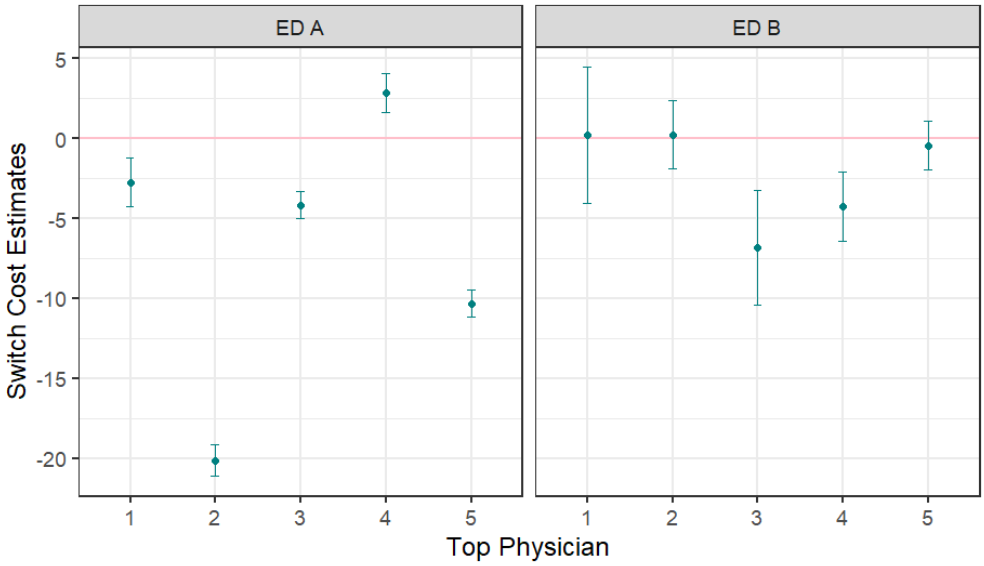


Figure EC.2 Switch Cost Coefficients Across Top Five ED Physicians

EC.1.5. Effect Heterogeneity across CCS Pairs

As discussed in Section 5.3, we estimate the coefficients $\delta_{m,n}$ in Equation (5) to explore the heterogeneity among 156 different switch pairs. Based on these estimates we further compute eq. (7) and obtain $\omega(i,j)$ as the weight of edges. Here we present the significant (at 5% level) estimates.

Table EC.4 Significant CCS Pairs Estimates

Pair	δ	Pair	δ	Pair	δ	Pair	δ
Panel A. ED A Significant CCS Pair Estimates							
gas - neu	-0.657**	gen - gnr	-0.893*	gnr - neu	-0.591**	gas - car	-0.637**
	(0.218)		(0.375)		(0.214)		(0.217)
ent - gas	-0.849***	ort - ent	0.672*	ent - car	-0.924**	gas - res	-0.528*
	(0.227)		(0.305)		(0.344)		(0.215)
neu - res	-1.056***	res - car	-0.592**	oth - gas	1.051*	gen - oth	-1.785***
	(0.244)		(0.16)		(0.48)		(0.424)
gas - men	-1.291**	gen - men	1.908*	men - neu	-1.1**	oph - ski	-0.841*
	(0.394)		(0.717)		(0.375)		(0.418)
oph - gnr	-1.261***	oph - ort	-1.077**	oph - car	-1.038*	oph - res	-1.475**
	(0.309)		(0.369)		(0.42)		(0.47)
gas - obg	-1.167**	obg - gas	1.542**	obg - neu	-1.294*	res - obg	-1.557***
	(0.352)		(0.56)		(0.64)		(0.341)
obg - oth	-1.354**	men - obg	-2.355**	oph - obg	-3.94***		
	(0.402)		(0.809)		(1.004)		
Panel B. ED B Significant CCS Pair Estimates							
neu - gnr	-0.576*	gas - gnr	-0.379**	gas - ent	-1.022**	ent - gnr	-1.538**
	(0.253)		(0.125)		(0.278)		(0.573)
car - ent	-1.109***	neu - ort	-0.889**	ort - neu	1.691*	gas - ort	-0.479*
	(0.287)		(0.308)		(0.7)		(0.231)
gnr - ort	-0.987***	car - ort	-0.551*	ent - ort	-1.576*	oth - neu	-0.775*
	(0.226)		(0.253)		(0.761)		(0.339)
gnr - oth	-0.946**	oth - gnr	-0.724*	gas - ski	-0.715*	gnr - ski	-1.636***
	(0.286)		(0.326)		(0.279)		(0.301)
ski - gnr	-1.158**	car - ski	-0.92***	ent - ski	-0.965***	ort - ski	-0.901*
	(0.392)		(0.227)		(0.098)		(0.415)
ski - ort	-2.036***	neu - men	-0.488**	men - gnr	-0.659**	car - men	-0.299*
	(0.39)		(0.151)		(0.194)		(0.139)
men - ort	-0.948**	men - oth	-0.717**	sub - oth	-1.189***	car - res	-0.385**
	(0.281)		(0.244)		(0.308)		(0.136)
res - ort	-0.659*	oth - res	-0.816**	res - ski	-0.812*	obg - car	-0.579*
	(0.309)		(0.244)		(0.332)		(0.251)
oth - obg	-0.916*	obg - res	-1.162***	gas - gen	-0.512*	gen - ort	-1.261**
	(0.42)		(0.294)		(0.196)		(0.38)
oth - gen	-0.797*	gen - ski	-1.995***	sub - gen	-0.941**		
	(0.331)		(0.3)		(0.321)		

Notes: ED A sample comprises 7,845 observations and R^2 is equal to 0.375. ED B sample comprises 10,521 observations and R^2 is equal to 0.398. Robust standard errors clustered by physicians in parentheses (* $p<0.05$, ** $p<0.01$, *** $p<0.001$).

EC.1.6. Proof of Proposition 1

Proof. Suppose all patient visits have been partitioned into two queues S and S^C, with $|S| = |S^C|$. If two patients i and j are in the same queue, then they will be picked consecutively by the same doctor with probability C. This probability C does not depend on i and j by our assumption that each sequence is uniformly sampled from the population. Thus, the total intra-queue switch cost is given by

$$C \cdot \sum_{i \neq j,\ i,j \in S} \omega(i,j) + C \cdot \sum_{i \neq j,\ i,j \in S^C} \omega(i,j)$$

which equals

$$C \cdot \sum_{i \neq j, i,j \in V} \omega(i,j) - C \cdot \sum_{i \in S, j \in S^C} \omega(i,j).$$

Since $C \cdot \sum_{i \neq j, i,j \in V} \omega(i,j)$ is a constant, minimizing the above difference is equivalent to maximizing the second term

$$C \cdot \sum_{i \in S, j \in S^C} \omega(i,j)$$

which is equivalent to

$$\max_{(i,j) \in \widetilde{E}} \quad \frac{1}{2} \sum (1 - y_i y_j)\, \omega(i,j)$$

by letting $y_i = 1$ if $i \in S$ and $y_i = -1$ if $i \in S^C$.

The additional constraint $\sum_{i<j} y_i y_j \leq -n/2$ is equivalent to $\sum_m y_m = 0$, which requires the sought-for partition to be a bisection. *Q.E.D.*

EC.2. Computation Steps in Simulation

In the simulation study, we simulate each patient visit to the ED and reconstruct the activity path accordingly. Specifically, we compute the waiting time and pick-to-pick (P2P) duration of each patient under the simulated scenario. Moreover, we need to consider the impact of switch cost on physician productivity under the new patient clustering and physician shift scheme. To achieve this, we consider the following approximation procedures regarding the translation between PPH and P2P.

Consider the picking moment of a patient i in a physician j's shift. We want to obtain the associated modified P2P duration until the next available time to select another patient. As a heuristic, we consider the time period starting from 40 minutes prior to the selection of patient i to the selection moment of the next patient after i by the focal physician as the PPH "surrounding" patient i. For the first 30-minute interval of a physician's shift, we compute the PPH and switching frequency associated with each CCS pair in Table EC.4. We then use these values for all patients picked in the interval.

Now we take into account the switch cost impact using the estimates $\widehat{\delta}_l$ from Table EC.4. In the simulation, we may obtain different patient selection results, which result in modified task switching frequency in this period of time. Compared with the switching frequency in the real data, we obtain the difference between the switching frequency denoted by $\mathbf{d}_i$ and we obtain the modified PPH denoted by $\widetilde{\text{PPH}}_i$ as:

$$\widetilde{\text{PPH}}_i = \text{PPH}_i + \widehat{\delta}^\top \mathbf{d}_i$$

where PPH_i represents the original PPH level in the data. We then obtain the proportional relation as

$$\frac{\widetilde{\text{PPH}}_i}{\text{PPH}_i} = 1 + \frac{\widehat{\delta}^\top \mathbf{d}_i}{\text{PPH}_i}$$

To translate the impact into P2P duration, we need the reciprocal of that. Therefore, the modified average P2P duration $\widetilde{\text{P2P}}_i$ for each patient i in the simulated paths is:

$$\widetilde{\text{P2P}}_i = \frac{\text{P2P}_i}{1 + \frac{\widehat{\delta}^\top \mathbf{d}_i}{\text{PPH}_i}}$$

where both P2P_i and PPH_i are known based on original data. The above equations show that, as the switching frequency decreases, PPH should increase and the average P2P duration will decrease and vice versa.