

The Cost of Task Switching: Evidence from the Emergency Department

Yige Duan

Vancouver School of Economics, University of British Columbia, Vancouver, BC V6T 1L4, Canada, yige.duan@alumni.ubc.ca

Yiwen Jin

Sauder School of Business, University of British Columbia, Vancouver, BC V6T 1Z2, Canada, yiwen.jin@sauder.ubc.ca

Yichuan Ding

Desautels Faculty of Management, McGill University, Montreal, QC H3A 1G5, Canada, daniel.ding@mcgill.ca

Mahesh Nagarajan

Sauder School of Business, University of British Columbia, Vancouver, BC V6T 1Z2, Canada,
mahesh.nagarajan@sauder.ubc.ca

Garth Hunte

Faculty of Medicine, University of British Columbia, Vancouver, BC V6T 1Z3, Canada, garth.hunte@ubc.ca

Emergency department physicians treat patients with different symptoms that require constant switching of tasks. Using a comprehensive data set with over 650,000 patient visits to four emergency departments, we investigate the impact of task switching on physician productivity. Our instrumental variable estimation, which exploits the exogenous composition of waiting patients, indicates that switching between patients of different types increases the average pick-to-pick time by 4.9 to 14.7 percent. The switch cost likely reflects both a time-consuming cognitive reconfiguration process and interference between tasks. Task switching also affects how physicians route patients, although we find little impact on healthcare quality. Our counterfactual analysis further reveals that the switch cost may have increased the average patient waiting time by 28.2 percent and the congestion level by 27.9 percent. Our research has important managerial implications for how switch costs can be addressed to improve workplace efficiency.

Key words: Task switching; emergency department; empirical healthcare management; discretionary task routing; queuing model.

1. Introduction

It is commonly believed that both technological changes and economic growth in recent decades have led to the rise of skilled and complex jobs in place of routine and monotonous jobs (Katz and Murphy, 1992; Autor et al., 2003). Consequently, workers are required to perform and switch between different tasks more often (Lindbeck and Snower, 2000). This could be challenging for humans, as many psychological experiments have shown that task switching leads to extra cognitive burdens, increased task completion time, and more performance errors (Jersild, 1927; Allport et al., 1994; Rogers and Monsell, 1995; Rubinstein et al., 2001). Thus, the cost of task switching (a.k.a. switch cost) is becoming an increasingly important determinant of worker productivity.

Despite the intensive investigation into task switching by psychological experimentalists, there has been little research on how task switching impairs productivity in real workplaces. An obstacle facing such studies is the measurement of task switching: Detailed diary data on tasks performed in everyday work is usually unavailable or too costly to collect. More importantly, task switching often arises endogenously (Ibanez et al., 2018; KC et al., 2020) and leads to biased estimates (Heckman, 1979). All these complications make it difficult to evaluate the consequences of task switching, especially in a causal sense.

In this paper, we examine the effect of task switching on the productivity of emergency department (ED) physicians. Our study exploits administrative data of patient visits to the four largest EDs in metro Vancouver, Canada. As the jack of all trades, ED physicians master a broad range of medical knowledge, practice a variety of clinical skills, and treat patients with all kinds of medical conditions. Since the conditions of patients and the approaches to treatment vary from case to case, physicians need to switch tasks frequently. In addition, physicians in our EDs of study work individually and keep treating arriving patients at the ED in their shifts. These factors make EDs an ideal environment to study task switching and its impact on physicians’ productivity.

To define task switching, we classify patients into different *types* by their triage acuity codes and chief complaint (CC) categories.¹ Both measures are determined by a triage nurse upon the patient’s arrival at the ED, before the physician’s work starts. Then we create a binary variable to indicate whether a patient has a different type from that of the preceding patient seen by the same physician. The indicator reveals that task switching is ubiquitous at EDs and takes places in over 95 percent of total treatments under the above definition of patient types. Also, it allows us to estimate the physician’s switch cost as the extra time taken to serve a type-switched patient.

Nevertheless, we face two major challenges in identifying the switch cost. First, similar to many other ED data sets, our data does not record all the details about a physician’s activities in the ED. Thus, we have to use available information from the data to estimate the amount of time during which a physician is heavily involved in serving patients. As we shall show later, the pick-to-pick (P2P) time defined between consecutive patient picks proxies this quantity aptly for the sake of our analysis. Second, the observed sequence of patient types, as a result of the physician’s picks, may suffer from selective task switching that undermines the validity of the estimates. To tackle endogenous patient selection, we construct an instrumental variable (IV) reflecting the share of waiting patients of different types in the queue from which the focal patient is selected. As such, we can quantify the causal effect of switching between different patient types.

¹The triage acuity code indicates the severity of a patient’s symptoms and the priority by which the patient should be attended. Meanwhile, the chief complaint is a classification of the patient’s symptoms based on the *Canadian Emergency Department Information System Presenting Complaint List* (Grafstein et al., 2008). In total, we identify 166 CC categories under 14 CC systems.

Following the literature in cognitive psychology, we empirically show that the switch cost likely stems from a time-consuming cognitive process named *task set reconfiguration* (Monsell, 1996; Rubinstein et al., 2001), in which physicians reconfigure clinical tasks when meeting a new patient. Anecdotal evidence suggests that the major cognitive task facing physicians is pattern recognition, which takes a longer time to complete when the patient’s symptoms are new or unfamiliar. In addition, we find evidence that working memories from the previous patient could interfere with the diagnosis of the focal patient. This effect of *task set inertia* (Allport et al., 1994; Mayr and Keele, 2000) also contributes to the observed switch cost.

Moreover, we find that the switch cost on P2P time translates to a much larger increase in patients’ waiting time and the congestion level of EDs. This is because any delay in serving a single patient will simultaneously raise the waiting time of multiple patients and prevent the queue from moving forward. To quantify such multiplier effects, we conduct a counterfactual analysis where we simulate the routing and processing of patients but “shut down” the switch cost on P2P time. The resulting average waiting time and average patient queue length (waiting census) are then compared with the actual data. We find that without switch cost, the patients’ average waiting time is reduced by 28.2 percent relative to the status quo. Analogously, the average waiting census is reduced by 27.9 percent.

Given the above findings, one may ask whether the quality of ED care is also affected by task switching, and whether ED physicians should strategically route their patients to minimize the switch cost for efficiency improvement. To answer these questions, we compare ED revisit-and-readmission (RAD) rates of patients of the same types and different types as the preceding patients but find no statistically significant difference. Meanwhile, we estimate that a patient’s probability of being picked by a physician decreases when he or she has a different type from the physician’s previous patient. Therefore, physicians may have strategically avoided task switching in their routing discretion, possibly in a subconscious way. However, due to the low frequency of RAD incidences, our estimates may not provide a comprehensive picture of the switch cost on care quality. Future works may investigate the potential risks of routing patients to minimize task switching.

Our study contributes to both methodology and practice.

First, we investigate the switch cost on both the efficiency and quality of emergency care delivery. We show that task switching has a detrimental effect on ED physicians’ productivity. Switching across different patient types increases the average P2P time by 4.9 to 14.7 percent in different EDs. This further leads to a 28.2 percent increase in average patient waiting time, and a 27.9 percent increase in the ED congestion level, according to our counterfactual analysis. Nevertheless, We find that task switching has no identifiable impact on RAD incidence.

Second, we propose the P2P time as a novel proxy measure of a physician’s productivity, given that detailed timestamps of physician activities during the P2P period are not available in our data.

This proxy can characterize ED physicians’ service at higher granularity level which is different with the aggregate productivity measurement in the existing literature, e.g., (Zaerpour et al., 2019). The validity of this proxy has been tested through a series of sophisticated designed robustness tests. The construction of this proxy also resonates with the theoretical research on ED service process (Yom-Tov and Mandelbaum, 2014; Huang et al., 2015).

Third, we exploit an IV strategy to overcome endogeneity due to discretionary patient routing in the estimation. Since the above issues likely exist in other service settings, the proposal of this proxy method contributes to empirical research on discretionary task routing in service queues.

Forth, we explore the psychological mechanisms driving the switch cost. Our analysis suggests that the switch cost stems from both task set reconfiguration and task set inertia. This is consistent with findings in the psychological literature and anecdotal evidence collected from our conversation with ED health workers.

Fifth, we provide empirical evidence that ED physicians tend to prioritize patients without task switching during patient routing decisions. Such behavior, referred to as switching aversion, has been identified at statistically significant levels despite the ubiquity of task switching at all four EDs.

Finally, we present practical suggestions that can offset the detrimental effect of task switching. We propose methods including better system design such as the adoption of fast track and the pod system (Valentine and Edmondson, 2015), raising physicians’ awareness of the switch cost in patient routing and assigning breaks to physicians’ shifts. These insights are valuable to ED managers and thus contribute to the ED practice.

The remainder of our paper is organized as follows. Section 2 reviews the literature related to our work. Section 3 describes the ED care delivery process and the data. Section 4 formulates the empirical method. Section 5 presents estimates of the switch cost and discusses the psychological mechanisms. Section 6 estimates the impact of task switching on the quality of ED care and patient routing. Section 7 reports results of robustness tests. Section 8 conducts a counterfactual analysis to quantify the switch cost on patient waiting time and the ED congestion level. Finally, Section 9 discusses the managerial implications of our work, and Section 10 concludes.

2. Literature Review

Our paper is related to five strands of literature that include both empirical and analytical studies. First, our study contributes to the healthcare management literature on physician productivity. Existing research has found various determining factors of physician productivity, 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). More generally, our paper is related to queuing theory with adaptive service behavior that studies the state-dependence of a system’s service

rate (Mandelbaum and Pats, 1998; Bekker et al., 2011; Abouee-Mehrizi and Baron, 2016; Delasay et al., 2016; Wang and Zhou, 2018; Tan and Netessine, 2019). These may lead to improvement or new designs of the care delivery system, especially regarding physician shift scheduling (Liu and Xie, 2018; Zaerpour et al., 2019), new triage policies (Saghafian et al., 2014), and waiting time prediction and announcement (Ang et al., 2016; Ding et al., 2020; Dong et al., 2019). We complement the above literature by providing empirical evidence about a new determinant of physician productivity and a new type of adaptive service behavior—the switch of tasks between consecutive patients or customers. Additionally, we provide managerial insights into novel ED designs.

Second, our paper is related to a growing body of research on discretionary task routing. Ibanez et al. (2018) examine the impact of discretionary task routing on radiologists’ reading time. They find that radiologists tend to prioritize similar tasks or tasks with shorter expected completion time, yet such behaviors have a detrimental effect on productivity. KC et al. (2020) also show that ED physicians selectively complete shorter and cognitively easier tasks when facing increased workload, which hinders throughput rate and learning in the long run. In addition, Ding et al. (2019) show that in general, ED physicians follow a delay-dependent rule and prioritize patients according to their waiting time and triage acuity codes. Instead, Bayati et al. (2017) find that the arrival of low acuity patients may delay the treatment of high acuity patients and increase their waiting time. In addition, Tan and Staats (2020) study a multiserver routing problem where a restaurant host assigns incoming customers to waiters. They find that host prioritizes efficient waiters more often when demand is high. Consistent with existing evidence, we find that physicians batch similar patients to avoid task switching. However, since there is no predetermined patient sequence in our research, we can separate the switch cost from the effect of deviation from a predetermined choice rule. Our estimates reveal that task switching is time-consuming, so avoiding task switching may actually improve the efficiency of ED physicians.

Third, a broad strand of psychological research studies task switching using laboratory experiments. By asking subjects to complete a sequence of tasks, researchers find that task switching increases both task completion time and error rates (e.g., Jersild, 1927; Allport et al., 1994; Rogers and Monsell, 1995; see Monsell and Driver, 2000 for an extensive literature review). Monsell (1996) and Rubinstein et al. (2001) propose that the switch cost reflects a time-consuming preparatory process for the new task. In contrast, Allport et al. (1994) and Mayr and Keele (2000) argue that task switching is an automatic process in which the previous task interferes with the current task. Our paper extends these studies to real workplaces. We show that task switching lowers the efficiency of ED physicians and that the switch cost is a combination of both preparatory and automatic processes. Different from experimental studies, however, we adopt an IV strategy to address the nonrandom switching of tasks.

Fourth, our paper is related to studies of task specialization and task variety, and their effects on productivity (Staats and Gino, 2012; Avgerinos and Gokpinar, 2018; Narayanan et al., 2009). Researchers also study learning from previous tasks, showing that task performance only benefits from experience on related tasks rather than unrelated tasks (Schilling et al. (2003), Boh et al. (2007), and KC and Staats (2012)). Our paper complements theirs by showing that avoiding task switching also improves the performance of subsequent tasks. Conceptually, task specialization and variety are cumulative processes and their effects on productivity dampen or exacerbate over time. However, task switching is an instantaneous event and its effect is more transitory. In addition, our study focuses on ED physicians whose work is cognitively demanding, stressful, and fast-paced, while previous research is based on either repetitive or routine tasks (Staats and Gino, 2012; Narayanan et al., 2009).

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) also review a large number of papers on healthcare delivery quality and summarize the most commonly used metrics of service quality in healthcare settings. They show that an important triptych of quality measures are structure, process and outcome. In addition, Batt and Tong (2020) investigate when and how server-level quality metrics differ from customer-experienced metrics and what effect will such judgment bring. Our paper investigates the potential trade-off between the time and quality of care delivery due to task switching and finds that the impact of task switching mostly materializes in the dimension of time.

3. Clinical Setting and Data

Our study is conducted at four EDs in metro Vancouver, Canada. All the EDs share a similar clinical setting and each provides emergency care for thousands of patients per month on a 24-hour-a-day and seven-day-a-week basis. ED physicians all work independently, delivering healthcare services for patients with a wide range of demographics and medical conditions. Physicians typically follow a shift of seven to eight hours.

3.1. The Care Delivery Process

To understand how patient care is delivered at EDs, we conduct a time-and-motion study by shadowing physician shifts in one of the four EDs. During shadowing, we, as the external observer, record the time a physician spends on each atomic activity using a stop watch. We also interview several healthcare workers in the ED to gather further information.

Figure 1 depicts a representative pathway of an ED patient. Upon arrival, each patient is catered by a triage nurse who collects the patient’s demographic and clinical information, evaluates the medical situation, and assigns a triage acuity code from 1 (resuscitation) to 5 (nonurgent). In particular, the nurse classifies the patient’s symptoms into a CC category. After the triage stage, patients with triage code 1 are usually sent to immediate resuscitation by physicians, while others typically need to wait until being picked by an available physician. A patient will be handled by the physician who picks him throughout his visit to the ED, except near the end of the physician’s shift, when the patient might be handed over to another physician.

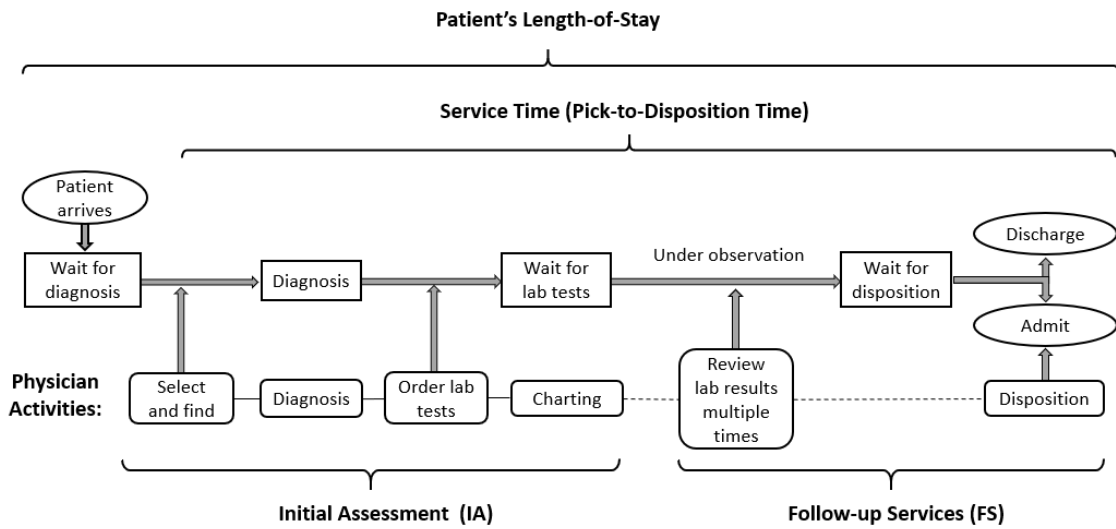


Figure 1 Typical ED Care Delivery Process

Notes: The summation of IA and FS time is referred to as needy time (Yom-Tov and Mandelbaum, 2014), which is much smaller than the service time as the latter includes the time of the patient being under observation without physician attendance.

All waiting patients’ demographic and clinical information is available to every physician for them to select the next patient for treatment. There is no explicit selection rule but it has been shown that physicians take both the patient’s waiting time and triage code into consideration. In particular, their choice patterns exhibit an effort to comply with a fractile response objective, which specifies a triage-specific target waiting time and a target percentage of patients to be picked within the target (Beveridge, 1998; Ding et al., 2019). Nevertheless, the large volume of patients makes it difficult to comply with the objective at most times.

The service process starts when a patient is picked by a physician and ends when the patient is discharged or admitted to an inpatient unit. This pick-to-disposition (P2D) time period is usually

referred to as the service period in the ED literature (Kuntz and Sülz, 2013; Batt and Terwiesch, 2016; Chaou et al., 2018). However, during most times of the service periods, patients are merely waiting for lab results or disposition decisions, requiring minimum time input from the physician. Therefore, the P2D time significantly overestimates the actual time that a physician spends on each patient. To come up with a more precise measurement for the latter quantity, we conduct a time-and-motion study to investigate the care delivery process of each patient. We find that the service process, despite of its high variation from patient to patient, can be modeled as a single-server reentrant queue. Specifically, services delivered can be categorized into two stages along a patient’s pathway, namely the initial assessment and follow-up services. This characterization of the service process is consistent with the existing literature (Yom-Tov and Mandelbaum, 2014; Huang et al., 2015; He et al., 2019). We detail the activities in each service stage as follows.

1. Initial Assessment (IA): A physician needs to select a patient by scanning their information in the patient care information system (PCIS). The physician also needs to physically find the selected patient, talk to the patient, make diagnostic decisions, and order lab tests. After the first meeting, the patient is put under observation, during which the physician could 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 will go over the lab results with the patient. In the latter case, the physician will check up the patient and decide whether to discharge or admit the patient to an inpatient unit.

The order of the above activities may exhibit large within-stage variation. For example, the lab tests can be ordered prior to the diagnosis, as the physician has already read the patient’s information from the patient portal before talking to the patient.

Our time-and-motion study also reveals that physicians may multitask at the patient level but not the activity level. They may finish an activity of the first patient and switch to another activity of the second patient. However, we never see physicians working on two or more activities simultaneously.

In total, we shadow two shifts for four hours each and observed treatments of 22 patients. The average IA time per patient is 13.3 minutes, and the average time of each follow-up service is 3.56 minutes. Despite the small sample size, the shadowing results still provide valuable insights into the ED care delivery process. In particular, the majority of the time is devoted to IA. This is because most of the physician’s activities, including selecting and finding the patient, diagnosis, ordering lab test and charting, are completed in the IA period, and physicians need time to familiarize themselves with the focal patient’s clinical condition. In contrast, the FS time is much shorter because the physician has already ascertained the condition of the patient and only needs to check the updates

or to make a decision. More importantly, should task switching affect physician productivity, the effect would mostly concentrate on the IA period. This is because the most intensive cognitive work takes places in the IA period, including processing information of new patients, pattern recognition (diagnosis), etc.

Unfortunately, the IA time is usually unobserved in administrative data sets that do not record the detailed activities of ED physicians. Instead, we can construct a pick-to-pick (P2P) time as the time difference between the picks of two consecutive patients seen by the same physician. We argue that P2P time is a better proxy for IA time than service time. On the one hand, our data shows that the average P2P time is approximately 20 minutes while the average service time lasts for 3.7 hours. In contrast, the service time severely overestimates the IA time because patients usually spend a long time waiting for lab test results or staying under observation, during which the physician is working on other patients. On the other hand, Figure 2 illustrates that the P2P time corresponds well to activities in the IA period. The major discrepancies between the P2P time and the IA time are interruptions brought by follow-up service requests from preceding reentrant patients. However, such discrepancies are small and mostly unrelated to task switching in our analysis. In Section 7, we use a fluid model to estimate the IA time from the P2P time. The resulting switch cost estimates are robust on both outcomes.

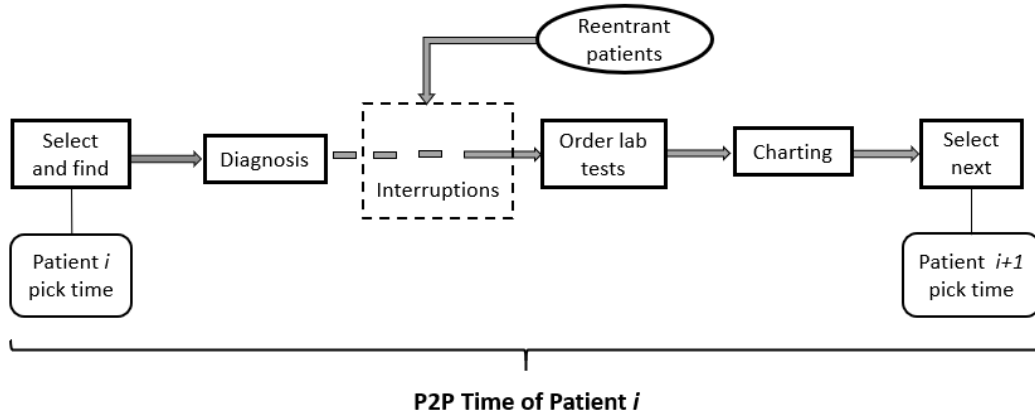


Figure 2 Typical Physician Activities between Two Consecutive Patient Picks

3.2. Data

Our data set, extracted from the National Ambulatory Care Reporting System (NACRS) metadata, comprises all patient visits at the four EDs from April 2013 to March 2019. The data set contains rich information about patient characteristics such as gender, age, homelessness, arriving transportation, triage acuity code, CC system, and CC category. Based on our interview with ED physicians about

how they select patients, we define patients with distinct triage codes or CC categories as having different *types* and define *task switching* as across types. Additionally, we define a patient’s waiting time as the difference between the arrival time and the pick time (Dong et al., 2019; Ding et al., 2019), and use whether a patient returns to any ED within 7 or 30 days and is admitted to an inpatient unit (RAD) as a measure of service quality (Calder et al., 2015; Wang et al., 2019).

To improve the precision of the P2P time in measuring physician productivity, we refine our sample of patient visits in the following steps. First, we focus on visits between 10 am and 10 pm, which is the most heavily loaded period of the EDs and best represents the steady state of the system. Second, we exclude visits when there is no outstanding waiting patient in the queue. For those visits, the P2P time is potentially contaminated by physician idleness. When there are patients waiting in the ED, we hardly see a physician taking any break. Third, we exclude the first and last thirty minutes of each physician’s shift because the number of reentrant requests could differ from the normal level due to hand-over (Batt et al., 2019) or slack-off (Chan, 2018) effects. Fourth, the physician sometimes picks two patients at the same time and treats them in a row, which is referred to as batched picking. We identified only approximately 10 percent picks as batched picking and reconstruct each patient’s P2P time by splitting the total P2P time equally among those picks. Finally, we obtain our working sample by correcting for data errors and outliers. Specifically, we exclude observations with missing values, top one percent outliers and negative values in P2P time, waiting time, and shift time. More details about the sample refinement criteria are provided in Appendix EC.1.

Table 1 summarizes the key variables in the refined sample. At the four EDs, the average P2P time is between 20 and 26 minutes. However, the average waiting time varies from 46 to 72 minutes. To measure the ED congestion level, we calculate the average *waiting census* and average *choice set size*, which are the average number of waiting patients counted every 10 minutes and at each picking epoch, respectively. Both measures vary across the EDs as well. Physicians’ shift time (time since the start of shift) and inpatient admission rates also vary greatly. Last, the 7- and 30-day RAD rates across the EDs range between 1.3–3.1 and 2.4–6.5 percent, respectively.

Regarding patient demographics, the average age of patients is approximately 50 years old at all EDs. The gender distribution is also even. While the EDs differ in the proportion of homeless patients and patients arriving by ambulance, all have a similar share of patients with switched types (approximately 96 percent). This indicates how common task switching is in ED services. Despite the small share of nonswitches, the large sample size still enables us to estimate the effect of task switching with high precision.

3.3. Preliminary Evidence of the Switch Cost

Before diving into the analysis, we graphically illustrate the association between task switching and P2P time. In Figure 3 depicts, in four different cases, the average P2P time of three consecutively

Table 1 Summary Statistics

	ED A	ED B	ED C	ED D
Variables	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)
P2P time	22.37 (14.99)	26.35 (18.89)	20.23 (13.29)	21.16 (17.63)
Waiting time	46.43 (37.59)	71.82 (56.98)	49.61 (35.41)	66.27 (46.84)
Waiting census	4.102 (2.764)	11.44 (5.540)	5.775 (3.725)	7.252 (4.062)
Choice set size	6.368 (3.826)	17.34 (7.478)	6.464 (3.689)	11.56 (5.299)
Shift time	137.8 (74.09)	195.6 (106.3)	180.4 (85.82)	184.3 (96.21)
Hospitalized	0.246 (0.431)	0.223 (0.416)	0.107 (0.309)	0.150 (0.357)
RAD in 7 days	0.031 (0.173)	0.017 (0.131)	0.013 (0.111)	0.019 (0.136)
RAD in 30 days	0.065 (0.247)	0.039 (0.193)	0.024 (0.154)	0.037 (0.189)
Patient age	49.03 (19.10)	52.46 (21.07)	51.06 (21.75)	48.22 (25.79)
Male patient	0.552 (0.497)	0.481 (0.500)	0.450 (0.497)	0.478 (0.499)
Homeless patient	0.072 (0.258)	0.013 (0.114)	0.010 (0.101)	0.005 (0.068)
Arrival by ambulance	0.386 (0.487)	0.235 (0.424)	0.098 (0.297)	0.177 (0.382)
Task switching	0.963 (0.189)	0.973 (0.163)	0.967 (0.178)	0.964 (0.186)
Observations	127,008	256,515	120,633	147,132
Physicians	87	94	85	58

Notes: Samples comprise all patient visits to four study EDs from April 1, 2013 to March 31, 2019, excluding visits before 10 am or after 10 pm, visits without outstanding waiting patients, visits in the first and last 30 minutes of a physician’s shift, and outliers in waiting time, P2P time, and shift time (time since shift starts). Waiting census is calculated by 10-minute time intervals.

treated patients. For example, in the upper-left panel, the first patient’s type is the same type as his precursors’ and the others’ are different. In the upper-right panel, the first and last patients have different types from their precursors while the second has the same type from the first. The lower panels are interpreted in similar ways. Consistent with our argument that task switching increases P2P time, we find that the P2P time is significantly longer whenever the patient’s type differs from the previous patient’s.

4. Empirical Methods

We adopt several empirical strategies to quantify the switch cost at EDs. First, we identify the effect on individual-level P2P time using an IV approach, which isolates bias due to endogenous patient selection. Next, we use alternative response variables and model specifications to tackle the discrepancy between the P2P and IA times. Details of our strategies are described as follows.

4.1. The Regression Model

Consider ED physicians indexed by $i = 1, 2, \dots, I$. Patients seen by physician i are indexed by $t = 1, 2, \dots, N_i$ in the order of arrival time. To simplify notations, we suppress the indices for hospitals and shifts so that I is the number of physicians in a representative hospital and N_i is the number of patients seen by physician i during a representative shift. Following the two-way fixed effect literature (Abowd et al., 1999), we denote $\mathbf{j}(i, t)$ as the t -th patient seen by physician i .

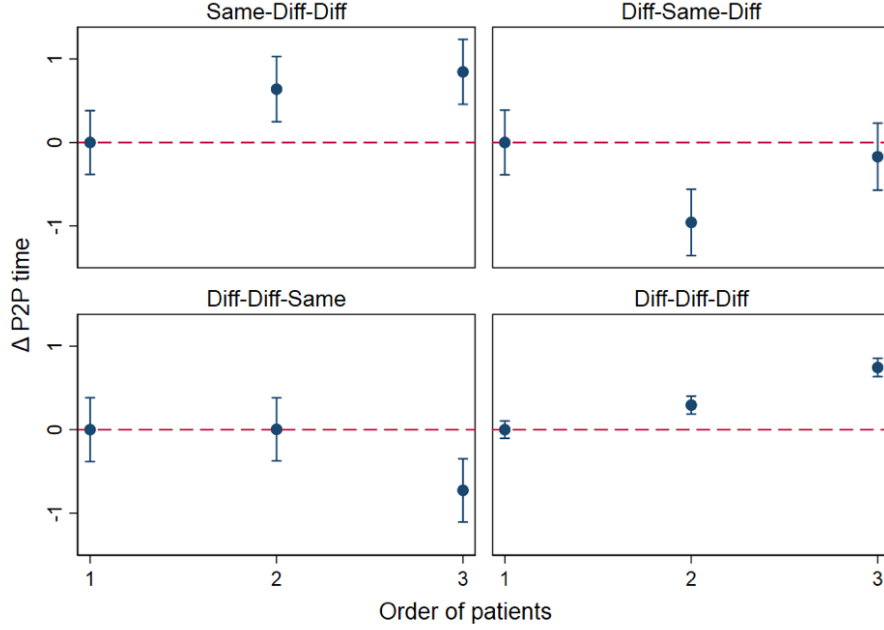


Figure 3 Task Switching and Treatment Time: An “Event Study” Design

Notes: The figure plots the average P2P time of three consecutively treated patients, adjusted for patient characteristics and linear time trends. The first P2P time in each panel is rescaled at zero.

Let $P2P_j$ be the P2P time of patient j . We assume the following linear model:

$$\log(P2P_{j(i,t)}) = D_{j(i,t)}\delta + X'_{it}\beta + Z'_{j(i,t)}\gamma + \epsilon_{it}. \quad (1)$$

The binary variable

$$D_{j(i,t)} = \mathbf{1}\{\text{type}_{j(i,t)} \neq \text{type}_{j(i,t-1)}\}, \quad (2)$$

indicates whether patient $j(i,t)$ has a different type from that of patient $j(i,t-1)$, i.e., the preceding patient seen by physician i . The parameter of interest, δ , reflects the effect of task switching on the P2P time. Following the literature, we use the natural logarithm of P2P time as the dependent variable and thereby assume that the switch cost is a fixed proportion of each patient’s P2P time (KC and Terwiesch, 2009; Staats and Gino, 2012; Chan, 2018). Additionally, we control for X_{it} , a vector of physician-treatment characteristics including the choice set size, the physician’s shift time, clock hour, and indicators of the focal physician-shift when patient $j(i,t)$ is picked; and $Z_{j(i,t)}$, a vector of patient characteristics including waiting time, gender, age, indicators of homelessness, arriving by ambulance, and most importantly, type.

4.2. Instrumental Variable

A concern with the linear regression model (4) is that patients are endogenously selected by physicians. If the physician is aware of the switch cost, she may intentionally avoid selecting patients of a

different type to mitigate the cost. As such, the explanatory variable $D_{\mathbf{j}(i,t)}$ will be endogenous and the ordinary least square (OLS) estimate of δ will be biased (Heckman, 1979).

To rectify the selection bias, we adopt an IV strategy that exploits the exogenous variation in the composition of waiting patients. Let $\mathcal{J}(i,t)$ be the choice set of waiting patients when patient $\mathbf{j}(i,t)$ is selected and $|\mathcal{J}(i,t)|$ be the size of the set. Then our IV is the proportion of waiting patients with a different type from that of the preceding patient $\mathbf{j}(i,t-1)$,

$$D_{\mathcal{J}(i,t)}^{\text{IV}} = \frac{1}{|\mathcal{J}(i,t)|} \sum_{j \in \mathcal{J}(i,t)} \mathbf{1}\{\text{type}_j \neq \text{type}_{\mathbf{j}(i,t-1)}\}. \quad (3)$$

Recall that a valid IV must satisfy two conditions: (1) It must be strongly correlated with the endogenous explanatory variable (the relevance condition); (2) It must not affect the outcome variable in any other ways via the error term (the exclusion condition). We argue that the variable defined in Equation (3) satisfies both conditions. On the one hand, $D_{\mathcal{J}(i,t)}^{\text{IV}}$ is correlated with $D_{\mathbf{j}(i,t)}$ by construction: The greater the proportion of waiting patients with a different type from the preceding patient's, the more likely the selected patient is to have a different type as well. On the other hand, $D_{\mathcal{J}(i,t)}^{\text{IV}}$ is not endogenously selected since the types of patient $\mathbf{j}(i,t-1)$ and the waiting patients are predetermined. Conditional on X_{it} and $Z_{\mathbf{j}(i,t)}$, $D_{\mathcal{J}(i,t)}^{\text{IV}}$ does not affect $\text{P2P}_{\mathbf{j}(i,t)}$ in any other way than $D_{\mathbf{j}(i,t)}$. Once the patient's characteristics are controlled for, $D_{\mathcal{J}(i,t)}^{\text{IV}}$ only depends on other patients' types that are uncorrelated with the P2P time of the focal patient. We also consider other IV forms and test the robustness of our results in Section 7.

We elaborate on the difference between our IV and Ibanez et al.'s (2018) in their study of discretionary task routing. They instrument for voluntary deviation from first-come-first-serve (FCFS) using an indicator of whether there exists a chance to deviate. If we were to directly apply their IV to our model, the resulting IV would be an indicator of the chance of repeating the previous patient's type which is equivalent to the minimum of the switch indicators on the right-hand-side of Equation (3). As an order statistic, however, this minimum depends on the choice set size $|\mathcal{J}(i,t)|$ in a nonlinear way. Consequently, the resulting estimates could be inconsistent if one only controls for a linear term of $|\mathcal{J}(i,t)|$. In contrast, the IV proposed here has a proportion form and therefore is uncorrelated with the size of $\mathcal{J}(i,t)$. Section 7 compares the estimates resulting from both IVs and shows that the performance of our IV is more robust.

4.3. Measurement Errors in the Response Variable

In the above regression, we use P2P time as the dependent variable as it better approximates the unobserved physician's IA time. As discussed above, the P2P time tends to overestimate the IA time by including the follow-up service time of other reentrant patients. However, as long as the measurement errors are uncorrelated with the task switching indicator, the estimated switch cost

will still be unbiased. We believe this is the case because the occurrence and duration of follow-up services does not depend on whether the focal patient has a different type from the previous patient or not. It may depend on the clock hour or physician shift time, both of which have already been controlled in the regression.

Another issue with the P2P time as the dependent variable is a phenomenon called task batching. Task batching happens when a physician picks two or more patients at almost the same time, and then treat them in a row (Jensen, 2011). Such a behavior create additional mismatch between the observed P2P time and the actual IA time. However, our data reveals that the probability of task batching is comparatively low and even during task batching, physicians do not handle more than three patients at a time.

To address the above concerns, we propose an alternative identification strategy where we regress the hourly throughput rate (the number of patients picked in each hour) against the hourly intensity of task switches (the share of patients with task switches). This new dependent variable measures a physician’s productivity at an aggregate level and therefore is more robust to measurement errors at the individual patient visit level. However, in the main regression model we adopt P2P time as the response variable because the estimates of switch cost at the individual level are more relevant for managerial interpretation and performance evaluation.

5. Results and Interpretations

Table 2 presents the estimation results of Equation (4) using two-stage least squares (2SLS). The first row shows the estimated δ , the switch cost on P2P time. Among the four EDs, the estimates range from 0.049 to 0.147 and are all statistically significant: *Ceteris paribus*, a switch in patient types increases the average P2P time by 4.9 to 14.7 percent—that is, task switching negatively influences physicians’ productivity. Evaluated at the sample average level, these values translate to 1.1 to 3.1 minutes of additional P2P time per patient. That says, in an eight-hour shift comprising 24 patients, the total switch costs could result in one to three fewer patients being treated. Thus, task switching also generates sizable financial losses to EDs.

To test the relevance condition of our IV, the bottom of Table 2 reports the Kleibergen-Paap rk Wald F statistics from weak IV tests. In all EDs, we reject the null hypothesis that the IV is barely weakly correlated with the endogenous task switch indicator $D_{j(i,t)}$. Thus our IV satisfies the relevance condition.

The rest of Table 2 reports estimates for other variables of interest. First, we find that a longer waiting time of the focal patient and a larger number of waiting patients at the picking time are both associated with a short P2P time. This is consistent with existing evidence that physicians accelerate the service as workload is high (KC and Terwiesch, 2009). In contrast, the positive coefficients of

Table 2 Effect of Task Switching on P2P Time

Variables	ED A	ED B	ED C	ED D
Task switching	0.049*** (0.017)	0.097*** (0.028)	0.066*** (0.022)	0.147*** (0.022)
Waiting time (log)	-0.068*** (0.004)	-0.036*** (0.003)	-0.064*** (0.005)	-0.060*** (0.005)
Choice set size	-0.062*** (0.002)	-0.021*** (0.001)	-0.052*** (0.002)	-0.034*** (0.001)
Shift time (log)	0.229*** (0.010)	0.232*** (0.008)	0.092*** (0.020)	-0.059*** (0.009)
Patient age	-0.013** (0.006)	0.002 (0.004)	-0.004 (0.003)	0.004 (0.003)
Patient age squared	0.002*** (0.001)	0.000 (0.000)	0.001** (0.000)	-0.000 (0.000)
Male patient	-0.007* (0.004)	-0.010*** (0.003)	-0.003 (0.004)	-0.003 (0.003)
Homeless patient	-0.028*** (0.008)	-0.041*** (0.013)	0.001 (0.025)	0.005 (0.031)
Arrival by ambulance	-0.002 (0.005)	-0.001 (0.004)	-0.001 (0.005)	0.020*** (0.006)
Observations	126,964	255,787	120,630	147,129
rk F	9,005.38	3,400.86	1,472.18	4,037.51

Notes: Estimated by 2SLS. Dependent variable: P2P time in logarithm. Coefficients for age and age squared multiplied by 10 and 100, respectively. All estimates control for indicators of patient's type, physician-shift, and clock hour. Robust standard errors clustered by physicians in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

shift time in the first three EDs imply that the average P2P time is longer for later visits in a shift, due to a fatigue effect eroding productivity (Delasay et al., 2016). We find no systematic effects of other patient characteristics on P2P time.

5.1. Effect Heterogeneity

Next, we investigate whether the switch cost varies across patients and physicians. For the former, we test whether the effect of task switching is moderated by physician shift time, choice set size, patient triage acuity code, gender, age, and day of week. We estimate Equation (4) with an additional interaction term between the switch indicator and the above moderators one at a time. As such, the new regression model is given by

$$\log(\text{P2P}_{j(i,t)}) = D_{j(i,t)}\delta + (D_{j(i,t)} \cdot z_{j(i,t)})\delta_z + X'_{it}\beta + Z'_{j(i,t)}\gamma + \epsilon_{it}, \quad (4)$$

where z is the moderator of interest.

Table 3 reports the estimates of δ and δ_z . In Panels A to D, we standardize z to zero mean and unit standard deviation, so that δ measures the switch cost evaluated at the sample average of the respective moderators, and δ_z measures the change in the switch cost when the moderate increases

by one standard deviation. In Panels E and F, δ_z indicates the difference in the switch cost between male and female patients and between visits on weekdays and weekends, respectively. We find that the patient’s triage level has mixed effects on the switch cost (Panel A), while other moderating effects are overall small and statistically insignificant. Therefore, the switch cost exhibits at most modest heterogeneity with respect to patient characteristics.

Moreover, we find little evidence of heterogeneous switch costs across physicians. Since our data does not contain physician characteristics, we use analysis of variance (ANOVA) to determine whether the switch cost differs across physicians. Specifically, we split the entire sample into small subsamples consisting of 100–500 patient visits, separately estimate the switch cost for each subsample, and apply ANOVA to the resulting estimates. As shown in Appendix EC.2, in none of the four EDs can we reject the null hypothesis that task switching has the same effect on all physicians at the 10 percent significance level. Thus, the analysis yields little evidence of heterogeneous switch costs among physicians in the same ED.

5.2. Mechanisms

The results above provide strong evidence that task switching prolongs P2P time and hence erodes the productivity of ED physicians. In what follows, we explore two potential mechanisms driving this effect.

Task set reconfiguration. Jersild (1927), Monsell (1996), and Meiran (1996) view task switching as a voluntary and controlled preparatory process of “task set reconfiguration”. To perform a task, people have to configure their mental states in accordance with the external task content. This task-specific mental configuration—called a “task set”—has to be reset every time the external task is switched. Hence, they argue that the switch cost reflects the time and cognitive resources consumed during task set reconfiguration.² In our discussion with ED physicians, we learn that the major cognitive task in ED care delivery is pattern recognition, i.e., to match the patient’s clinical characteristics to a certain therapy. According to task set reconfiguration theory, the observed switch cost may well reflect the time to establish new links between symptoms and therapies when there is task switching.

According to Rubinstein et al. (2001) and Yeung and Monsell (2003), task set reconfiguration is more difficult when one switches to a less familiar or more complex task. This allows us to test the task set reconfiguration theory without observing the preparatory time but exploiting the switch cost heterogeneity by type familiarity. In Table 4, Panel A, we add controls for the standardized frequency of the focal type in each ED and its interaction with the switch indicator. At all four EDs, we find the interactive effects to be negative and statistically significant; therefore, the switch cost decreases

²Typical cognitive activities involved in task set reconfiguration include, for example, attention shifting, goal retrieval, and rule setting (Monsell and Driver, 2000).

Table 3 Effect Heterogeneity

Variables	ED A	ED B	ED C	ED D
<i>Panel A</i>				
Task switching	0.042** (0.017)	0.099*** (0.028)	0.065*** (0.023)	0.149*** (0.022)
Task switching $\times$ std. triage level	0.094** (0.046)	-0.017*** (0.003)	0.013 (0.032)	-0.015*** (0.005)
Observations	126,964	255,787	120,630	147,129
rk F	295.03	1,706.74	598.69	2,022.26
<i>Panel B</i>				
Task switching	0.049*** (0.017)	0.097*** (0.028)	0.066*** (0.021)	0.147*** (0.022)
Task switching $\times$ std. shift time (log)	0.009 (0.017)	0.021 (0.021)	-0.004 (0.019)	0.016 (0.020)
Observations	126,964	255,787	120,630	147,129
rk F	3,052.11	1,526.596	401.49	1,333.02
<i>Panel C</i>				
Task switching	0.085*** (0.020)	0.095*** (0.029)	0.048* (0.026)	0.157*** (0.027)
Task switching $\times$ std. choice set size	0.068*** (0.021)	0.035 (0.024)	-0.033* (0.019)	0.026 (0.029)
Observations	126,964	255,787	120,630	147,129
rk F	842.09	679.41	429.68	1,346.82
<i>Panel D</i>				
Task switching	0.047*** (0.018)	0.081*** (0.028)	0.067*** (0.022)	0.150*** (0.024)
Task switching $\times$ std. patient age	-0.033** (0.016)	-0.004 (0.023)	0.022 (0.020)	0.031 (0.040)
Observations	126,964	255,787	120,630	147,129
rk F	1,644.07	1,301.27	533.90	774.61
<i>Panel E</i>				
Task switching	0.077*** (0.027)	0.075** (0.032)	0.104*** (0.029)	0.135*** (0.027)
Task switching $\times$ male patient	-0.050 (0.038)	0.016 (0.050)	-0.090* (0.047)	0.022 (0.039)
Observations	126,964	255,787	120,630	147,129
rk F	1,852.95	1,463.41	723.14	1,257.70
<i>Panel F</i>				
Task switching	0.061*** (0.019)	0.083*** (0.028)	0.067*** (0.022)	0.145*** (0.022)
Task switching $\times$ visiting on weekend	-0.031** (0.015)	-0.003 (0.010)	-0.009 (0.010)	-0.000 (0.017)
Observations	126,964	255,787	120,630	147,129
rk F	4,556.99	1,697.09	736.58	2,018.87

Notes: Estimated by 2SLS. Dependent variable: P2P time in logarithm. All estimates control for patient's waiting time, age, gender, homelessness, arrival by ambulance, choice set size, shift time, indicators of patient's type, physician-shift, and clock hour. Robust standard errors clustered by physicians in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

in the frequency of the focal patient’s type. This is consistent with the reconfiguration theory in that the physician needs extra time to recognize the symptoms and search for a new therapy when she switches to a patient of a less common type.

Table 4 Test of Psychological Mechanisms

Variables	ED A	ED B	ED C	ED D
<i>Panel A. Task set reconfiguration</i>				
Task switching	0.091** (0.036)	0.123*** (0.038)	0.125*** (0.032)	0.206*** (0.033)
Task switching $\times$ focal type frequency	-0.034* (0.020)	-0.030** (0.015)	-0.065*** (0.020)	-0.056*** (0.015)
Observations	126,964	255,787	120,630	147,129
rk F	888.61	936.97	472.90	930.55
<i>Panel B. Task set inertia</i>				
Task switching	0.055** (0.023)	0.108*** (0.028)	0.105*** (0.022)	0.167*** (0.024)
Task switching $\times$ previous type frequency	-0.043** (0.018)	-0.013 (0.025)	-0.035** (0.017)	-0.011 (0.018)
Observations	126,964	255,787	120,630	147,129
rk F	2,548.46	318.18	1,039.73	616.23

Notes: Estimated by 2SLS. Dependent variable: P2P time in logarithm. All estimates control for patient’s waiting time, age, gender, homelessness, arrival by ambulance, choice set size, shift time, indicators of patient’s type, physician-shift, and clock hour. Patient type frequencies are standardized to zero mean and unit standard deviation. Robust standard errors clustered by physicians in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

Task set inertia. Another strand of research claims that task switching is costly because the previously established task set interferes with the focal task set, competes for cognitive resources (e.g. working memory), and thus undermines the performance of the focal task. This “task set inertia” theory (Allport et al., 1994; Meuter and Allport, 1999) argues that complex or unfamiliar tasks often require the activation of stronger task sets that are particularly hard to eliminate. Thus, switching out of such tasks would cause a larger switch cost. We test this hypothesis by interacting the switch indicator with the frequency of the previous patient’s type. In Table 4, Panel B, the interactive effect is also negative, but only significant in two of the four EDs.

To summarize, we find strong evidence of task set reconfiguration at all EDs and evidence for task set inertia in two. It is likely that both mechanisms take place in generating the observed switch cost. As discussed in Meiran (1996) and Pashler (2000), the two mechanisms emphasize different aspects of complex cognitive activities and their relative importance could vary from case to case.

6. Other Effects of Task Switching

In this section, we investigate two important questions related to task switching: (1) How will task switching affect ED service quality? (2) Have ED physicians attempted to avoid task switches when picking the patients?

6.1. Quality of Care

To quantify the effect of task switching on the quality of ED care, we estimate a linear probability model as follows:

$$\text{RAD}_{\mathbf{j}(i,t)}^d = D_{\mathbf{j}(i,t)}\delta + X'_{it}\beta + Z'_{\mathbf{j}(i,t)}\gamma + \epsilon_{it} \quad (5)$$

The right-hand-side of Equation (5) is the same as in Equation (4); the switch indicator $D_{\mathbf{j}(i,t)}$ is analogously instrumented by the IV in Equation (3). However, the response variable $\text{RAD}_{\mathbf{j}(i,t)}^d$ is now an indicator that a patient returns to any ED within d days after the focal visit and is admitted to an inpatient unit during the revisit. Following the literature (Calder et al., 2015; Wang et al., 2019), we focus on 7- and 30-day RAD and set $d = 7$ and 30 in Equation (5). We estimate the coefficients and cluster the standard errors by physicians as before.

Table 5 shows that the effects of task switching on both 7- and 30-day RAD are small and statistically insignificant. The effects of task switching on inpatient admission are also negligible. The estimates reflect the status quo trade-off made by physicians and have two implications. First, task switching has not adversely affected a patient's treatment outcome, which might be due to the extra time that a physician spends in treating a switched patient to counterbalance the associated challenge. Second, we find no evidence that faster service for nonswitch patients is at the expense of compromised quality.

6.2. Patient Routing

To answer the second question of interest, we investigate how task switching affects physicians' choice of patients using the conditional logistic model. Suppose that physicians are homogeneous and their utility from selecting patient j is given by

$$U_j = V_j + e_j. \quad (6)$$

Following (McFadden, 1973), we let the intrinsic valuation $V_j = Z'_j\xi + D_j\zeta$ depend on the patient's characteristics Z_j and whether there is a task switching D_j . The unobserved e_j is assumed to follow the type-I extreme value distribution. Then the probability of choosing j from $\mathcal{J}(i, t)$ is given by

$$\Pr\{j = \mathbf{j}(i, t)\} = \frac{\exp(D_j\pi + Z'_j\xi)}{\sum_{k \in \mathcal{J}(i, t)} \exp(D_k\pi + Z'_k\xi)}. \quad (7)$$

Table 5 Effect of Task Switching on Treatment Quality				
Variables	ED A	ED B	ED C	ED D
<i>Panel A. RAD in 7 days</i>				
Task switching	0.011 (0.007)	0.005 (0.006)	-0.003 (0.005)	0.002 (0.006)
Observations	126,964	255,787	120,630	147,129
rk F	9,005.38	3,400.86	1,472.18	4,037.51
<i>Panel B. RAD in 30 days</i>				
Task switching	0.011 (0.007)	0.005 (0.006)	-0.003 (0.005)	0.002 (0.006)
Observations	126,964	255,787	120,630	147,129
rk F	9,005.38	3,400.86	1,472.18	4,037.51
<i>Panel C. Inpatient admission</i>				
Task switching	0.001 (0.012)	-0.004 (0.012)	-0.007 (0.008)	0.010 (0.011)
Observations	126,964	255,787	120,630	147,129
rk F	9,005.38	3,400.86	1,472.18	4,037.51

Notes: Estimated by 2SLS. Dependent variables: indicators of RAD in 7 and 30 days (Panels A-B) and the indicator of inpatient admission (Panels C). All estimates control for patient's waiting time, age, gender, homelessness, arrival by ambulance, choice set size, shift time, indicators of patient's type, physician-shift, and clock hour. Robust standard errors clustered by physicians in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

The coefficient ζ captures the impact of task switching on the choice probability. If $\zeta < 0$, patients with a switched type are selected with lower priority, and vice versa.

Ding et al. (2019) show that V_j is a triage-specific piecewise linear function of patient j 's waiting time. That is, before the waiting time exceeds some critical value, V_j increases in the waiting time at a constant rate that varies across triage levels. After the critical value, V_j increases in the waiting time at a slower rate. Thus, with a slight abuse of notation, we assume that $Z_j'\xi$ contains the following additive terms,

$$b_0^{Tri(j)} + b_1^{Tri(j)} \cdot wait_j + b_2^{Tri(j)} \cdot (wait_j - \tau^{Tri(j)})^+, \quad Tri(j) = 2, 3, 4, 5. \quad (8)$$

In Equation (8), $Tri(j)$ and $wait_j$ denote, respectively, the triage level and waiting time of patient j in minutes. Meanwhile, $\tau^{Tri(j)}$ denotes the kink of the triage-specific, piecewise linear function. For triage levels 2 to 5, the kinks are 15, 30, 60, and 120 minutes, respectively. Finally, $b_0^{Tri(j)}$ represents the intercept coefficient and $b_1^{Tri(j)}$, $b_2^{Tri(j)}$ represent the slope coefficients. We refer readers to Ding et al. (2019) for a more detailed discussion about this formulation.

Table 6 reports the estimation results of Equation (7) with the coefficients transformed to odds ratios (i.e., b transformed to $\exp(b)$). As before, standard errors are clustered by physicians. We observe that a patient is more likely selected when her triage acuity score is lower (more acute) and

her waiting time is longer, which is consistent with our expectation. In particular, the coefficients for task switching are negative and statistically significant at all four EDs, corresponding to odds ratios of 0.468 to 0.935. Therefore, the odds of selecting a different-type patient is 46.8 to 93.5 percent of that of selecting a same type patient at the four EDs, conditional on all other attributes being the same. These estimates reflect the switch-averse behavior of ED physicians in patient routing decisions.

Table 6 Effect of Task Switching on Physician's Choice of Patients

Variables	ED A (Odds Ratio)	ED B (Odds Ratio)	ED C (Odds Ratio)	ED D (Odds Ratio)
Task switching	0.935*** (0.015)	0.492*** (0.012)	0.706*** (0.045)	0.468*** (0.009)
Triage 2: Intercept (b_0^2)	0.010*** (0.001)	0.143*** (0.008)	0.011*** (0.003)	0.011*** (0.001)
Triage 2: Waiting time (b_1^2)	1.400*** (0.011)	1.067*** (0.004)	1.359*** (0.015)	1.282*** (0.009)
Triage 2: Waiting time over 15 min (b_2^2)	0.707*** (0.005)	0.939*** (0.003)	0.742*** (0.008)	0.783*** (0.006)
Triage 3: Intercept (b_0^3)	0.008*** (0.001)	0.030*** (0.002)	0.007*** (0.002)	0.009*** (0.001)
Triage 3: Waiting time (b_1^3)	1.135*** (0.003)	1.051*** (0.002)	1.141*** (0.003)	1.091*** (0.003)
Triage 3: Waiting time over 30 min (b_2^3)	0.887*** (0.002)	0.959*** (0.002)	0.894*** (0.002)	0.925*** (0.002)
Triage 4: Intercept (b_0^4)	0.020*** (0.002)	0.033*** (0.002)	0.016*** (0.004)	0.020*** (0.003)
Triage 4: Waiting time (b_1^4)	1.058*** (0.001)	1.031*** (0.001)	1.068*** (0.002)	1.039*** (0.002)
Triage 4: Waiting time over 60 min (b_2^4)	0.948*** (0.002)	0.977*** (0.001)	0.949*** (0.001)	0.971*** (0.002)
Triage 5: Intercept (b_0^5)	0.107*** (0.014)	0.100*** (0.007)	0.045*** (0.011)	0.048*** (0.005)
Triage 5: Waiting time (b_1^5)	1.016*** (0.002)	1.009*** (0.000)	1.034*** (0.001)	1.015*** (0.001)
Observations	1,060,626	5,149,582	884,211	2,003,427

Notes: Estimated by conditional logistic regressions. Dependent variable: indicator of selected patients. All estimates control for patient's age, gender, homelessness, arrival by ambulance, and CC systems (controlling for patient's types requires too many dummy variables to execute the estimation). The coefficient b_2^5 is dropped because the corresponding group (patients in triage 5 with waiting time over 120 min) has too few observations. Estimated coefficients converted to odds ratios (i.e., e^b instead of b) and corresponding robust standard errors clustered by physicians in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

7. Robustness Tests

7.1. Aggregate Response Variable

As mentioned in Section 4, our use of the P2P time as a proxy for the IA time raises concerns about measurement errors and task batching, which could potentially confound our estimates. Therefore, we alternatively estimate the effect of task switching on the physician’s throughput rate, which is less affected by confounders at the patient visit level.

The regression model is as follows,

$$\text{PPH}_{ih} = \bar{D}_{ih}\zeta + W'_{ih}\eta + u_{ih}, \quad (9)$$

where we aggregate the patient visit data by physician (i) and shift-hour (h), and define new variables at the physician-shift-hour level. The new dependent variable PPH_{ih} is the number of patients a physician treats per hour, and the hourly *task switching intensity* $\bar{D}_{ih}$ is measured as the share of patients with a switched type in that hour. The corresponding coefficient ζ indicates the impact of task switching on the throughput rate. In the covariate vector W_{ih} , we control for the hourly average choice set size, and the share of patients in different waiting time groups, age groups, triage levels, and CC system groups, the share of male patients, homeless patients, and patients arriving by ambulance. Additionally, W_{ih} includes indicators of the focal physician-shift, shift hour, and clock hour. We estimate Equation (9) by OLS and cluster the standard errors by physicians.

As Table 7 shows, the estimated effect of task switching intensity on the throughput rate is negative and statistically significant across all EDs. That says, if we compare an hour with no switched patient versus another hour with only switched patients but hold everything else constant, on average, the latter will have a lower throughput by 0.164 to 0.24 patients per hour. Evaluated at the sample mean level, this is equivalent to 3.2 to 5.7 foregone patients per eight-hour shift.

7.2. Estimating the IA Time using a Fluid Model

In this subsection, we propose a method to estimate each patient’s IA time IA_j from his P2P time based on fluid approximation. Figure 4 provides a graphical illustration of the fluid model. New patient visits and reentrant visits are modeled as two different types of fluid, and will join two separate queues, namely, the regular queue and the reentrant queue. Both queues share a single server (physician), while the reentrant queue has head-of-line priority over the regular queue, i.e., the server will process new patient fluid only if when all reentrant patient fluid has been processed. Let $M(t)$ denote the number of patients under observation at time t , and γ denote the average rate for a reentrant patient to generate a follow-up request (i.e., either lab results review or disposition). According to the fluid approximation, the instantaneous arrival rate for the reentrant queue is given by $M(t)\gamma$. We assume that the reentrant patient fluid has a constant service speed μ .

Table 7 Effect of Task Switching on Throughput Rate

Variables	ED A	ED B	ED C	ED D
Task switch intensity	-0.185*** (0.043)	-0.164*** (0.027)	-0.240*** (0.055)	-0.172*** (0.051)
Average choice set size	0.111*** (0.003)	0.022*** (0.001)	0.073*** (0.003)	0.034*** (0.002)
Share of waiting time <15 min	-0.158*** (0.042)	-0.011 (0.017)	-0.202*** (0.057)	-0.102** (0.041)
Share of waiting time 15-30 min	0.011 (0.039)	-0.005 (0.015)	0.173*** (0.052)	0.006 (0.035)
Share of waiting time 30-60 min	0.052 (0.037)	0.026* (0.014)	0.257*** (0.048)	0.072** (0.031)
Share of waiting time 60-20 min	0.020 (0.036)	0.006 (0.012)	0.091** (0.046)	0.090*** (0.029)
Share of patients aged <25	0.032 (0.029)	0.125*** (0.017)	0.101*** (0.037)	0.141*** (0.027)
Share of patients aged 25-44	0.041* (0.021)	0.080*** (0.012)	0.091*** (0.028)	0.121*** (0.026)
Share of patients aged 45-64	0.034* (0.020)	0.050*** (0.010)	0.063** (0.027)	0.069*** (0.024)
Share of male patients	0.014 (0.015)	0.023*** (0.008)	0.004 (0.021)	-0.020 (0.018)
Share of homeless patients	0.056* (0.030)	0.053 (0.033)	-0.140 (0.100)	0.080 (0.110)
Share of arrivals by ambulance	0.001 (0.016)	-0.013 (0.010)	-0.120*** (0.035)	-0.116*** (0.023)
Sample average throughput rate	2.677	2.435	2.978	2.650
Observations	72,602	157,888	42,841	74,730
R^2	0.48	0.34	0.38	0.30

Notes: Estimated by OLS. Dependent variable: Throughput rate per hour. Task switching intensity is the share of patients treated in an hour with switched types. All estimates control for shares of patients in each triage level and each CC system, and indicators of physician-shift, shift hour, and clock hour. Robust standard errors clustered by physicians in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

Given the above assumptions, the fraction of capacity to be allocated to the new patient queue is $(1 - M(t)\gamma/\mu)^+$. This fraction is usually positive because $M(t)$ is less than 10 and $M(t)\gamma < \mu$. If patient j is picked by a physician at time t_j , his P2P time can be expressed as

$$\text{P2P}_j = \inf \left\{ s \left| \int_{t_j}^{t_j+s} \left(1 - \frac{M(t)\gamma}{\mu} \right)^+ dt = \text{IA}_j \right. \right\}. \quad (10)$$

Following Hawkins (2007) and MMH (2017), we assume $\gamma = 1/60$ per minute. Using the results from the time-and-motion study (Section 3), we set $\mu = 0.281$ per minute. Plugging these parameters into Equation (10) allows us to estimate the IA time from the P2P time. Then we can estimate the switch cost on the IA time using Equation (4) with the response variable $\text{P2P}_{j(i,t)}$ replaced by $\text{IA}_{j(i,t)}$. All explanatory variables and the IV strategy remain unchanged.

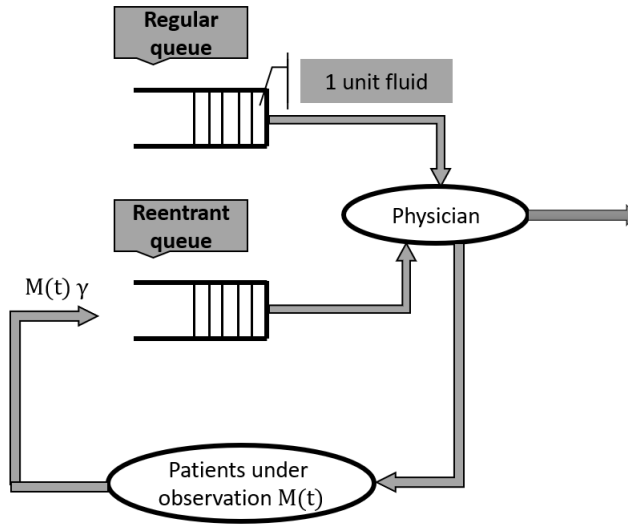


Figure 4 Typical Sequence of Patient Activities in the ED

Table 8 presents the resulting estimates. First, we note that the estimated switch cost on the IA time is close to our baseline estimates in Table 2. Furthermore, the average estimated IA time from the fluid model is of a similar magnitude to the average IA time observed from our time-and-motion study. The results validate our argument that the P2P time is a plausible proxy of the IA time.

Table 8 Effect of Task Switching on IA Time

Variables	ED A	ED B	ED C	ED D
Task switching	0.057*** (0.018)	0.098*** (0.028)	0.072*** (0.023)	0.159*** (0.025)
Sample average IA time	14.43	19.29	14.62	15.36
Observations	126,905	255,779	120,630	146,903
rk F	9,003.63	3,400.75	1,472.18	3,991.03

Notes: Estimated by 2SLS. Dependent variable: IA time in logarithm. All estimates control for patient's waiting time, age, gender, homelessness, arrival by ambulance, choice set size, shift time, indicators of patient's type, physician-shift, and clock hour. Robust standard errors clustered by physicians in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

7.3. Alternative Definition of Task Switching

Next, we test the robustness of our results to the levels of switch. Table 9, Panel A provides the estimation results of model (4), where we define the patient *type* by CC categories only. As such, serving two patients with the same CC category but different triage levels is no longer considered a switch. We find positive and statistically significant estimates at all EDs, suggesting that switching across CC categories is still costly in terms of P2P time. In Table 9, Panels B, we define task switching as switching across CC systems. Again, we find significant switch costs in three EDs.

7.4. Alternative IV

Finally, our estimates are robust to an alternative IV for task switching. We follow Ibanez et al. (2018) and define the IV for task switching as an indicator that at least one waiting patient has a switched type. As mentioned in Section 4, the new IV is an order statistic and depends on the choice set size in a nonlinear way. As such, we additionally control for a cubic polynomial of the choice set size. As shown in Table 9, Panel C, the estimated switch cost is still positive and significant, and the effect magnitude is close to our baseline estimates.

Table 9 Additional Robustness Tests

Variables	ED A	ED B	ED C	ED D
<i>Panel A. Switch by CC category</i>				
Task switching	0.039** (0.016)	0.080*** (0.021)	0.053*** (0.016)	0.143*** (0.017)
Observations	126,964	255,787	120,630	147,129
rk F	11,958.05	5,425.63	2,067.36	6,779.32
<i>Panel B. Switch by CC system</i>				
Task switching	0.002 (0.009)	0.059*** (0.012)	0.045*** (0.010)	0.157*** (0.012)
Observations	126,964	255,787	120,630	147,129
rk F	34,353.88	19,569.01	6,523.29	34,816.06
<i>Panel C. Alternative IV</i>				
Task switching	0.076*** (0.020)	0.119*** (0.030)	0.072*** (0.024)	0.183*** (0.024)
Observations	126,964	255,787	120,630	147,129
rk F	3,123.75	3,512.80	1,336.96	2,487.72

Notes: Estimated by 2SLS. Dependent variable: P2P time in logarithm. All estimates control for patient's waiting time, age, gender, homelessness, arrival by ambulance, choice set size, shift time, indicators of patient's type, physician-shift, and clock hour. Panel C additionally controls for a cubic polynomial of the choice set size. Robust standard errors clustered by physicians in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

8. Counterfactual Analysis

The reduced-form regression in Section 4 allows us to estimate the negative effect of task switching on physicians' productivity at the individual patient level. In this section, we present a counterfactual analysis to evaluate the effect of task switching at the ED system level. Using historical data, we construct a counterfactual scenario in which task switching would not prolong a physician's IA time or P2P time. Although this may not happen in reality, comparing the counterfactual and real scenarios helps researchers and ED managers comprehend the system-wide impact of task switching. In fact, we expect the switch cost to be more prominent at the system level than at the individual level because a delay in the focal patient's service will introduce further delay for all patients in the queue.

The counterfactual analysis is performed in one of the four study EDs and covers the entire study period, i.e., from 10 am to 10 pm every day from April 1, 2013 to March 31, 2019, excluding physicians' idle periods, the first and last thirty minutes of a physician's shift, and missing values and outliers (see Section 3 and Appendix EC.1). To construct the counterfactual scenario, we use exactly the same patient arrival time, patient attributes, and physician shift schedules as those in the historical data. However, the P2P time is reconstructed by eliminating the switch cost identified from the regression model. The new P2P time is calculated using the following formula:

$$\text{P2P}_{j(i,t)}^{\text{new}} = \text{P2P}_{j(i,t)} \cdot \left(1 - D_{j(i,t)} \hat{\delta}\right), \quad (11)$$

where the variables follow the same definition as in Equation (4). In particular, the term $D_{j(i,t)} \hat{\delta}$ captures the estimated switch cost from Table 2.

Next, we define the patient selection rule in the counterfactual scenario. In many cases, since the arrival time and patient attributes follow from the data, physicians still face the original choice set. As such, we will let them pick the same patient as they did in the history. However, due to the modification of the P2P time, the patient pick time in the counterfactual scenario could be different, and some physicians might face a different choice set. We therefore have to devise a new patient selection rule that aligns the patient pick sequence as close as that in the data. The new rule is determined by the following procedures.

First, if we observe that a physician selects a patient, say A , from a choice set, say $\{A, B, C\}$ in the historical data, we infer that the selected patient A has a higher priority over the rest, i.e., $A \succ B$ and $A \succ C$. Then we find the closure of the above partial order by transitivity. For example, if later a physician selects patient B from a new choice set $\{B, C, D\}$, we can then infer $B \succ D$, and thus $A \succ D$ by transitivity.

Second, for patient pairs whose relative order cannot be determined in the first step, we assume they are ordered by their arrival time. In the above example, if later a patient E joined the choice set and a physician picked E , then we cannot infer the partial order between A (or B) and E from the first step. We then let $A \succ B \succ E$ following the first-come-first-serve (FCFS) rule. We enforce the FCFS rule when partial order information is absent because in reality, 60.5 percent of the patient picks follow the FCFS rule and it is easier to simulate.

An implicit assumption in defining the above picking rule is that all ED physicians are homogeneous in their patient picking behavior. The validity of this assumption has been empirically tested in Ding et al. (2019).

Table 10 compares key performance metrics in the counterfactual scenario and the real scenario. In row (1), we see that the actual average waiting time is 49.6 minutes per patient and the actual average

waiting census (number of waiting patients) is 5.8 patients. Next, row (2) shows that the average waiting time and waiting census in the counterfactual scenario are 36.6 minutes and 4.2 patients, respectively. Comparing row (2) with row (1), we find significant efficiency gains from eliminating the switch cost: For each patient visit, the average waiting time is reduced by 13.99 minutes, which equals 28.2 percent of the actual waiting time and 8.7 times the switch cost on the average P2P time. The average waiting census also decreases by 1.6 patients, which alleviates congestion in the waiting room by 27.9 percent. The large effect of task switching on the system level compared to the individual level arises because any delay in a single treatment is borne by multiple patients waiting at the same time. Essentially, the high-congestion environment in the ED amplifies the switch cost on ED physicians, which in turn further exacerbates the congestion level in the ED.

Table 10 Results of Counterfactual Analysis

	Waiting Time Mean (S.D.)	Waiting Census Mean (S.D.)
(1) Real Data	49.61 (35.41)	5.775 (3.730)
(2) Counterfactual	35.62 (29.38)	4.162 (3.888)
(3) Difference	13.99	1.613

Notes: Samples comprise all patient visits to four study EDs from April 1, 2013 to March 31, 2019, excluding visits before 10 am or after 10 pm, visits in idle periods (10-minute intervals with no waiting patient), visits in the first and last 30 minutes of a physician’s shift, and outliers in waiting time, P2P time, and shift time (time since shift starts). Waiting census is calculated by 10-minute time intervals. Row (1): key performance metrics computed based on real data. Row (2): the counterfactual results. Row (3): difference between row (1) and row (2).

9. Managerial Implications

We summarize the managerial and policy insights from our analysis that could help ED managers and healthcare practitioners improve operational efficiency.

First, our study highlights the existence of the task switching cost in emergency healthcare and its large impact on the efficiency of the ED system. We find that task switching extends the average P2P time by 4.9 to 14.7 percent. However, we do not find evidence that serving same-type patients with less time affects the quality of care. As such, efforts to reduce the switch cost can render great value while researchers and ED managers should still be careful of the potential risk of compromised care quality. More work is required to explore this question.

In reality, it is impossible to eliminate task switching due to the complexity of ED treatments and diversity of ED patients, but several approaches can be implemented to reduce task switching. The first angle is ED layout design improvement. For example, as some EDs have done in practice, patients with modest symptoms and requiring relatively simple treatment can be treated exclusively

on a fast track. Physicians who work in the fast track will then treat a more homogeneous group of patients, resulting in a lower frequency of task switching. Another effective method is the pod system for ED workers (Valentine and Edmondson (2015)). This can enhance inter-professional communication efficiency that may mitigate the delay caused by task switching. For example, in case a physician encounters a new patient type and needs some information, he may have easy access to information resources nearby. In addition, the computer system can highlight waiting patients of the same type as the focal patient to help physicians avoid task switching. Ultimately, it is the physician's responsibility to carefully weight the switch cost against other factors (e.g., acuteness, waiting time, etc.) in making patient prioritization decisions.

Second, our empirical evidence shows that ED physicians are switch averse. *Ceteris paribus*, the odds of selecting a same-type patient can be twice as high as that of selecting a different-type patient in some EDs. According to our conversation with ED physicians, such switch-averse behavior is usually subconscious. If ED managers were to increase physicians' awareness of the task switch effect, physicians would possibly make better trade-offs in patient sequencing and increase their service efficiency.

Finally, our study points to task set reconfiguration and task set inertia as two potential mechanisms driving the switch cost in EDs. According to our interview with ED physicians, when a patient has a different or unfamiliar type, a physician needs a longer time matching the patient's patterns to a therapy. Additionally, it is possible for the preceding patient to interfere with the diagnosis of the current patient. To cope with this, ED managers can consider assigning breaks to ED physicians after a certain period of time. During the shadowing study at the ED, we find that the physicians hardly get any rest in the entire seven-to-eight-hour shift. Some physicians may have the incentive to keep working along the shift to finish earlier or to treat more patients. A new shift can add short breaks after two or three hours of work while keeping the same working hours to help physicians refresh and relieve the cognitive burden due to task switching.

10. Conclusion and Future Research

Although research on task scheduling and adaptive service rates is abundant, little is known about whether and how task switching affects productivity. In this paper, we investigate the switch cost of ED physicians to bridge this gap. EDs are well known as extremely busy workplaces heavily loaded at most times. Patients in EDs are characterized by heterogeneous symptoms and acuity levels so that physicians need to provide a variety of medical care at a fast pace. As a result, physicians experience a sizable and robust productivity loss due to task switching. Our analysis reveals that the switch cost leads to an average of 4.9 to 14.7 percent longer P2P time and physicians try to avoid task switching by batching patients with the same CC category and triage level. Using a counterfactual

analysis, we show that the switch cost on the patient’s waiting time and the ED congestion level is even more pronounced. These findings are particularly meaningful to ED managers and policymakers. Academics should also pay close attention to switch costs in heavily loaded service contexts.

It is worth noting that our study has wide implications for occupations and industries beyond healthcare. Task switching is prevalent in both the service and production sectors of the economy and is ubiquitous in the lives of humans. 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 ready for intensive task switching, multitasking, or densely interrupting working environments. Therefore, the cost of task switching could become an increasingly important factor that harms productivity at workplaces. The negative impact of task switching also highlights 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; Bloom et al., 2013, 2015).

We conclude our paper by suggesting some avenues for future research. Our paper focuses on ED physicians’ switches across patient types and the consequences on productivity. Future studies can further 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, sustainability of work, as well as patient satisfaction and treatment quality. As task switching renders few patients to be treated and less fee-for-service earned in a time unit, it is important to quantify the monetary value of the switch cost. Future research may also compare task switching, multitasking, and work interruptions in a unified setting, and to 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; Roback, 1982; Mas and Pallais, 2017) will help researchers recover the monetary equivalent cost of task switching.

References

- Abouee-Mehrizi, H. and Baron, O. State-Dependent M/G/1 Queueing Systems. *Queueing Systems*, 82(1):121–148, 2016.
- Abowd, J. M., Kramarz, F., and Margolis, D. N. High Wage Workers and High Wage Firms. *Econometrica*, 67(2):251–333, 1999.
- Allon, G. and Kremer, M. Behavioral Foundations of Queueing Systems. In Donohue, K., Katok, E., and Leider, S., editors, *The Handbook of Behavioral Operations*, pages 325–366. John Wiley & Sons, Hoboken, NJ, 2019.

- Allport, A., Styles, E. A., and Hsieh, S. Shifting Intentional Set: Exploring the Dynamic Control of Tasks. In Umiltà, C. and Moscovitch, M., editors, *Attention and Performance XV: Conscious and Nonconscious Information Processing*, pages 421–452. MIT Press, Cambridge, MA, 1994.
- Ang, E., Kwasnick, S., Bayati, M., Plambeck, E. L., and Aratow, M. Accurate Emergency Department Wait Time Prediction. *Manufacturing & Service Operations Management*, 18(1):141–156, 2016.
- Autor, D. H., Levy, F., and Murnane, R. J. The Skill Content of Recent Technological Change: An Empirical Exploration. *Quarterly Journal of Economics*, 118(4):1279–1333, 2003.
- Avgerinos, E. and Gokpinar, B. Task Variety in Professional Service Work: When It Helps and When It Hurts. *Production and Operations Management*, 27(7):1368–1389, 2018.
- Batt, R. J. and Terwiesch, C. Early Task Initiation and Other Load-Adaptive Mechanisms in the Emergency Department. *Management Science*, 63(11):3531–3551, 2016.
- Batt, R. J. and Tong, J. D. Mean Service Metrics: Biased Quality Judgment and the Customer–Server Quality Gap. *Manufacturing & Service Operations Management*, 22(5):975–995, 2020.
- Batt, R. J., KC, D. S., Staats, B. R., and Patterson, B. W. The Effects of Discrete Work Shifts on a Nonterminating Service System. *Production and Operations Management*, 28(6):1528–1544, 2019.
- Bayati, M., Kwasnick, S., Luo, D., and Plambeck, E. L. Low-Acuity Patients Delay High-Acuity Patients in an Emergency Department. *SSRN Working Paper No. 3095039*, 2017.
- Bekker, R., Koole, G. M., Nielsen, B. F., and Nielsen, T. B. Queues with Waiting Time Dependent Service. *Queueing Systems*, 68(1):61–78, 2011.
- Beveridge, R. CAEP Issues. The Canadian Triage and Acuity Scale: A New and Critical Element in Health Care Reform. Canadian Association of Emergency Physicians. *Journal of Emergency Medicine*, 16(3):507–511, 1998.
- Bloom, N. and Van Reenen, J. Human Resource Management and Productivity. In Card, D. and Orley, A., editors, *Handbook of Labor Economics*, volume 4, pages 1697–1767. Elsevier, 2011.
- Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. Does Management Matter? Evidence from India. *Quarterly Journal of Economics*, 128(1):1–51, 2013.
- Bloom, N., Liang, J., Roberts, J., and Ying, Z. J. Does Working from Home Work? Evidence from a Chinese Experiment. *Quarterly Journal of Economics*, 130(1):165–218, 2015.
- Boh, W., Slaughter, S. A., and Espinosa, A. J. Learning from Experience in Software Development: A Multilevel Analysis. *Management Science*, 53(8):1315–1331, 2007.
- Calder, L., Pozgay, A., Riff, S., Rothwell, D., Youngson, E., Mojaverian, N., Cwinn, A., and Forster, A. Adverse Events in Patients with Return Emergency Department Visits. *BMJ Quality & Safety in Health Care*, 24(2):142–148, 2015.
- Chan, D. C. Teamwork and Moral Hazard: Evidence from the Emergency Department. *Journal of Political Economy*, 124(3):734–770, 2016.

-
- Chan, D. C. The Efficiency of Slacking Off: Evidence from the Emergency Department. *Econometrica*, 86(3):997–1030, 2018.
- Chaou, C.-H., Chen, H.-H., Tang, P., Yen, A. M.-F., Wu, K.-H., Hsiao, C.-T., and Chiu, T.-F. Traffic Intensity of Patients and Physicians in the Emergency Department: A Queueing Approach for Physician Utilization. *Journal of Emergency Medicine*, 55(5):718–725, 2018.
- Delasay, M., Ingolfsson, A., and Kolfal, B. Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research*, 64(4):867–885, 2016.
- Ding, Y., Park, E., Nagarajan, M., and Grafstein, E. 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, 2019.
- Ding, Y., Nagarajan, M., and Zhang, G. Parallel Queues with Discrete-Choice Arrival Pattern: Empirical Evidence and Asymptotic Characterization. *SSRN Working Paper No. 3584880*, 2020.
- Dong, J., Yom-Tov, E., and Yom-Tov, G. B. The Impact of Delay Announcements on Hospital Network Coordination and Waiting Times. *Management Science*, 65(5):1969–1994, 2019.
- Grafstein, E., Bullard, M. J., Warren, D., Unger, B., and Group, C. N. W. Revision of the Canadian Emergency Department Information System (CEDIS) Presenting Complaint List Version 1.1. *Canadian Journal of Emergency Medicine*, 10(2):151–161, 2008.
- Hawkins, R. C. Laboratory Turnaround Time. *Clinical Biochemist Reviews*, 28(4):179–194, 2007.
- He, S., Sim, M., and Zhang, M. Data-Driven Patient Scheduling in Emergency Departments: A Hybrid Robust-Stochastic Approach. *Management Science*, 65(9):4123–4140, 2019.
- Heckman, J. J. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161, 1979.
- Huang, J., Carmeli, B., and Mandelbaum, A. Control of Patient Flow in Emergency Departments, or Multiclass Queues with Deadlines and Feedback. *Operations Research*, 63(4):892–908, 2015.
- Ibanez, M. R., Clark, J. R., Huckman, R. S., and Staats, B. R. Discretionary Task Ordering: Queue Management in Radiological Services. *Management Science*, 64(9):4389–4407, 2018.
- Jensen, K. B. Fast and Efficient Practice: The Emergency Department Clinician on the Emergency Department Autobahn, 2011. Technical Report.
- Jersild, A. T. Mental Set and Shift. *Archives of Psychology*, 89:5–82, 1927.
- Katz, L. F. and Murphy, K. M. Changes in Relative Wages, 1963–1987: Supply and Demand Factors. *Quarterly Journal of Economics*, 107(1):35–78, 1992.
- KC, D. S. and Staats, B. R. Accumulating A Portfolio of Experience: The Effect of Focal and Related Experience on Surgeon Performance. *Manufacturing & Service Operations Management*, 14(4): 618–633, 2012.
- KC, D. S. and Terwiesch, C. Impact of Workload on Service Time and Patient Safety: An Econometric Analysis of Hospital Operations. *Management Science*, 55(9):1486–1498, 2009.

- KC, D. S., Staats, B. R., Kouchaki, M., and Gino, F. Task Selection and Workload: A Focus on Completing Easy Tasks Hurts Performance. *Management Science*, 66(10):4397–4416, 2020.
- Kuntz, L. and Sülz, S. Treatment Speed and High Load in the Emergency Department—Does Staff Quality Matter? *Health Care Management Science*, 16(4):366–376, 2013.
- Lindbeck, A. and Snower, D. J. Multitask Learning and the Reorganization of Work: From Tayloristic to Holistic Organization. *Journal of Labor Economics*, 18(3):353–376, 2000.
- Liu, R. and Xie, X. Physician Staffing for Emergency Departments with Time-Varying Demand. *INFORMS Journal on Computing*, 30(3):588–607, 2018.
- Mandelbaum, A. and Pats, G. State-Dependent Stochastic Networks. Part I. Approximations and Applications with Continuous Diffusion Limits. *The Annals of Applied Probability*, 8(2):569–646, 1998.
- Mas, A. and Pallais, A. Valuing Alternative Work Arrangements. *American Economic Review*, 107(12):3722–3759, 2017.
- Mayr, U. and Keele, S. W. Changing Internal Constraints on Action: The Role of Backward Inhibition. *Journal of Experimental Psychology: General*, 129(1):4–26, 2000.
- McFadden, D. Conditional Logit Analysis of Qualitative Choice Behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142, New York, 1973. Academic Press.
- Meiran, N. Reconfiguration of Processing Mode Prior to Task Performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1423–1442, 1996.
- Meuter, R. F. I. and Allport, A. Bilingual Language Switching in Naming: Asymmetrical Costs of Language Selection. *Journal of Memory and Language*, 40(1):25–40, 1999.
- MMH. *Wait Times Reduction Task Force: Final Report*. Manitoba Minister of Health (MMH), Seniors and Active Living, Canada, 2017.
- Monsell, S. Control of Mental Processes. In Bruce, V., editor, *Unsolved Mysteries of the Mind: Tutorial Essays in Cognition*, pages 93–148. Psychology Press, Hove, U.K., 1996.
- Monsell, S. and Driver, J. *Control of Cognitive Processes: Attention and Performance XVIII*. MIT Press, Cambridge, Massachusetts, 2000.
- Narayanan, S., Balasubramanian, S., and Swaminathan, J. M. A Matter of Balance: Specialization, Task Variety, and Individual Learning in a Software Maintenance Environment. *Management Science*, 55(11):1861–1876, 2009.
- Pashler, H. Task Switching and Multitask Performance. In Monsell, S. and Driver, J., editors, *Control of Cognitive Processes: Attention and Performance XVIII*, pages 277–307. MIT Press, Cambridge, Massachusetts, 2000.
- Roback, J. Wages, Rents, and the Quality of Life. *Journal of Political Economy*, 90(6):1257–1278, 1982.

-
- Rogers, R. D. and Monsell, S. Costs of a Predictable Switch between Simple Cognitive Tasks. *Journal of Experimental Psychology: General*, 124(2):207–231, 1995.
- Rosen, S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 82(1):34–55, 1974.
- Rubinstein, J. S., Meyer, D. E., and Evans, J. E. Executive Control of Cognitive Processes in Task Switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4):763–797, 2001.
- Saghafian, S., Hopp, W. J., Van Oyen, M. P., Desmond, J. S., and Kronick, S. L. Complexity-Augmented Triage: A Tool for Improving Patient Safety and Operational Efficiency. *Manufacturing & Service Operations Management*, 16(3):329–345, 2014.
- Schilling, M. A., Vidal, P., Ployhart, R. E., and Marangoni, A. Learning by Doing Something Else: Variation, Relatedness, and the Learning Curve. *Management Science*, 49(1):39–56, 2003.
- Silver, D. Haste or Waste? Peer Pressure and the Distribution of Marginal Returns to Health Care. *Review of Economic Studies*, forthcoming, 2020.
- Song, H. and Veeraraghavan, S. Quality of Care. *Handbook of Healthcare Analytics: Theoretical Minimum for Conducting 21st Century Research on Healthcare Operations*, pages 79–108, 2018.
- Staats, B. R. and Gino, F. Specialization and Variety in Repetitive Tasks: Evidence from a Japanese Bank. *Management Science*, 58(6):1141–1159, 2012.
- Tan, T. F. and Staats, B. R. Behavioral Drivers of Routing Decisions: Evidence from Restaurant Table Assignment. *Production and Operations Management*, 29(4):1050–1070, 2020.
- Tan, T. F. and Netessine, S. When you work with a superman, will you also fly? an empirical study of the impact of coworkers on performance. *Management Science*, 65(8):3495–3517, 2019.
- Valentine, M. A. and Edmondson, A. C. Team Scaffolds: How Mesolevel Structures Enable Role-Based Coordination in Temporary Groups. *Organization Science*, 26(2):405–422, 2015.
- Wang, J. and Zhou, Y.-P. Impact of queue configuration on service time: Evidence from a supermarket. *Management Science*, 64(7):3055–3075, 2018.
- Wang, Y., Ding, Y., Park, E., and Hunte, G. 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, 2019.
- Yeung, N. and Monsell, S. Switching between Tasks of Unequal Familiarity: The Role of Stimulus-Attribute and Response-Set Selection. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2):455–469, 2003.
- Yom-Tov, G. B. and Mandelbaum, A. Erlang-R: A Time-Varying Queue with Reentrant Customers, in Support of Healthcare Staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014.

Zaerpour, F., Bijvank, M., and Sun, Z. Scheduling of Physicians with Time-Varying Productivity Levels in Emergency Departments. *SSRN Working Paper No. 3519625*, 2019.

Online Appendix

EC.1. Sample Refinement

As mentioned in Section 3, we refine our sample to exclude patients whose P2P time poorly proxies the IA time. This appendix will discuss the refinement process in greater detail.

We begin with all patient visits to the four EDs between April 2013 and March 2019. As shown in row (1) of Table EC.1, we obtain a total of over 126 million visits, with the numbers varying across the EDs considerably. First, we want to exclude those idle periods of the physicians since idleness may contaminate our P2P time measure. Thus, we remove visits before 10 am or after 10 pm each day when physicians are less busy but more likely idle. We also drop visits during whose P2P time there is no waiting patient. This is because when there is no outstanding workload, we cannot determine whether the physician is working or taking a break. The resulting sample sizes are reported in Table EC.1, row (2) and row (3).

Next, Table EC.1, row (4) excludes patients in the first and last thirty minutes per shift. Previous research has shown that physicians often need to attend patients handed over from other physicians at the beginning of their own shifts (Batt et al., 2019), and they may slack off close to the end of their shifts by shortening treatments or cherry-picking easier tasks (Chan, 2018a). More importantly, the number of patients requiring follow-up services could be nonstationary at the two ends of a shift.

In the last step, we delete patients with missing characteristics such as age and gender, and whose P2P, waiting, or shift time exceed reasonable levels (i.e., the 99th percentile). Additionally, we drop patients whose type is unique and who are the only patient in a physician-shift. These patients form singletons in the fixed effect regression and they do not contribute to the identification of parameters (Abowd et al. 1999). The final sample size is in Table EC.1, row (5).

Table EC.1 Sample Refinement				
Sample	ED A	ED B	ED C	ED D
(1) Initial	271,833	518,653	179,710	296,280
(2) Limit to 10 am – 10 pm	169,881	330,247	154,224	200,242
(3) Drop treatments without waiting patients	164,128	306,844	140,142	179,114
(4) Drop first and last 30 minutes per shift	130,411	263,067	123,137	150,711
(5) Drop missing values, outliers, and singletons	127,008	256,515	120,633	147,132

Notes: The initial sample (row (1)) comprises all patient visits to each ED between April 1, 2013 and March 31, 2019. Rows (2)-(4) exclude, respectively, visits before 10 am or after 10 pm, without waiting patients, and in the first and last 30 minutes per shift. Row (5) excludes observations with missing values, outliers in waiting time, P2P time, and shift time, and singletons with respect to types and physician shifts.

EC.2. Switch Cost Heterogeneity across Physicians

To investigate whether the estimated switch cost varies across physicians, we perform an ANOVA test whose details are described as follows.

In the first step, we split the entire data set into subsamples where we separately estimate the switch cost. For each ED, we index physicians using $i = 1, \dots, N$ and patients seen by physician i using $t = 1, \dots, N_i$ in the order of arrival date and time (see Section 3). For each physician, we take her first to the 500th patients, the 501st to the 1000th patients and so on as distinct subsamples. For subsamples comprising fewer than 500 patients, we only keep those with more than 100 patients. In this way, we obtain a total of 1,479 subsamples for 290 physicians at four EDs.

Second, we estimate Equation (4) using the baseline IV strategy on every subsample and obtain 1,479 switch cost estimates. We exclude estimates for which the rk F statistic is below 10 (the relevance condition is violated). This yields 1,124 switch cost estimates for 261 physicians.

Now we perform the ANOVA test separately for each ED. The unit of analysis is the estimated switch cost weighted by the subsample size (70.1 percent of subsamples have exactly 500 patients). As shown in the first part of Table EC.2 below, the total variance in the switch cost estimates (TSS) is much larger than the variance across physicians (ESS) in every ED. Measured by R^2 , physician heterogeneity only explains up to 25 percent of the total variation. Accordingly, the ANOVA tests yield small F statistics that are statistically insignificant at all EDs. Thus, we cannot reject the null hypothesis that there is no difference in the average switch cost across physicians.

Table EC.2 Switch Cost Heterogeneity across Physicians: ANOVA Results

	ED A	ED B	ED C	ED D
ESS	9.685 (71)	24.95 (77)	6.300 (57)	10.26 (52)
RSS	28.63 (185)	86.22 (268)	23.45 (180)	51.80 (230)
TSS	38.32 (256)	111.2 (345)	29.75 (237)	62.06 (282)
R^2	0.2527	0.2244	0.2118	0.1653
F	0.88	1.01	0.85	0.88
p -value	0.7628	0.1311	0.7625	0.7107
Observations	257	346	238	283
Physicians	72	78	58	53

Notes: Sample comprises 1,124 switch cost estimates for 261 physicians. All observations weighted by subsample size. Degrees of freedom in parenthesis.