Emergency Care Access vs. Quality: Uncovering Hidden Consequences of Fast-Track Routing Decisions

Shuai Hao
Gies College of Business, University of Illinois at Urbana-Champaign, Champaign, IL 61820, shuaih2@illinois.edu

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

Yuqian Xu
Kenan-Flagler Business School, University of North Carolina at Chapel-Hill, NC 27599, yuqian_xu@kenan-flagler.unc.edu

Problem definition: This work aims to examine the impact of the emergency department (ED) fast-track (FT) routing decisions on patient outcomes and propose evidence-based routing policies to guide the FT routing decisions. **Methodology/results:** In this paper, we utilize a two-year dataset from two hospital EDs in Alberta, Canada, and adopt an instrumental variable (IV) approach to quantify the impact of the FT routing decisions on patient outcomes. Based on the empirical findings, we propose a multi-class queueing model to derive the optimal routing policy and then utilize a data-calibrated simulation to compare the performance of different routing policies. First, our study reveals that FT routing decisions are not purely clinical-driven, and ED operational status related to congestion is also associated with FT routing decisions. Second, we find that being routed to FT can improve patient access to emergency care by reducing the average ED length of stay (LOS). However, this efficiency improvement comes at the cost of potential quality decline. In particular, we find that being routed to the FT leads to a 6.8% (6.6%) increase in the 48-hour (72-hour) revisit rate for high-complexity patients, and a 5.8% (5.8%) increase in the 48-hour (72-hour) revisit rate for medium-complexity patients. Third, we show that our proposed optimal state-dependent routing policy can lead to a 5.44% reduction in the 48-hour patient revisits and a 21.9% reduction in the average patient waiting time compared to the current routing policy used by our study hospitals. **Managerial implications:** Our empirical findings call for immediate attention from healthcare practitioners to carefully balance the trade-off between the access to emergency care and the quality of care. Using the estimated effects of FT routing decisions on patient outcomes for different patient groups, we also propose potentially implementable routing policies for hospital EDs.

Key words: emergency department, empirical healthcare, behavioral operations, fast-track routing, queueing.

1. Introduction

Emergency department (ED) congestion has been observed in many hospitals across the world and poses critical challenges to both healthcare practitioners and policy makers. According to the National Center for Health Statistics, 40%–50% of US hospitals have experienced ED congestion (Burt and McCaig 2006). As a result, patients have to spend hours in the waiting area, leading to an increased risk of cross-infection, mortality, and patient readmission (Guttmann et al. 2011). Hence, a crowded ED is more than a nuisance; it is a threat to both individual patients and overall public health (Maa 2011). Many strategies have been proposed to regulate patient flow and reduce ED congestion. Among these, fast-track (FT) has been highlighted by the American College of Emergency Physicians (ACEP) as a high-impact initiative (Liu et al. 2013). In particular, FT is a separate ED area that provides dedicated pathways aimed toward fast care delivery and rapid discharge for patients with less urgent conditions, which becomes more prevalent in recent years and has been implemented by nearly 80% of academic EDs in the US (Liu et al. 2013).

It has been documented in earlier medical studies that the implementation of FT is a great success in serving low acuity patients and improving ED operational efficiency in terms of reduced patient waiting time, LOS, and left without being seen (LWBS); see, for example, Sanchez et al. (2006), Ieraci et al. (2008), Devkaran et al. (2009), Chrusciel et al. (2019), and Grant et al. (2020). It is, therefore, natural to expect that the adoption of FT improves healthcare quality, for example, through reduced patient revisits, if the FT routing decision can always select the "right" patients to be treated in the FT area. However, since the FT is a rigidly separated area, the mismatch between demand and supply in both the FT and main areas can occur if the routing decisions are purely based on clinical conditions, which leads to operational inefficiency. Particularly in a congested system, the workload in the two treatment areas can be heavily unbalanced as a result of high demand variation. Therefore, the triage nurse, who serves as a "dispatcher" and determines whether a patient should be routed to the main or FT area, may consider the ED operational conditions in the FT routing decisions to better match healthcare resources with demands. As a result, patients with similar clinical conditions might receive treatment in different ED areas (i.e., main vs. FT) under different ED congestion conditions. Consequently, it is unclear whether being routed to FT will lead to adverse effects on patients. Therefore, this study first aims to answer the following two questions: (i) whether non-clinical factors such as ED congestion status are also associated with the FT routing decision; and (ii) whether potential adverse effects exist if being routed to the FT.

Moreover, so far, hospitals have not yet established consistent guidelines for determining which patients should be routed to the FT, which might be due to the lack of a more comprehensive understanding of how FT routing decisions impact patient outcomes as we discussed earlier. Upon arriving at an ED, a patient who does not have life-threatening conditions is first triaged by the nursing staff, who (i) assigns the patient a triage score that indicates the urgency level of the patient's care needs and (ii) routes the patient to either the main ED area, where most patients are treated, or to the FT area. Standard protocols have been established for assigning triage scores, such as the Canadian Triage and Acuity Scale (CTAS), the most commonly used triage protocol in Canada, and the Emergency Severity Index (ESI), the algorithm commonly adopted in the US. Both protocols are five-point scoring systems (1 to 5) with smaller numbers indicating higher levels of urgency. However, neither of these protocols specifies the type of patients that should be routed to the FT. Hence, EDs currently make FT routing decisions at their own discretion. Some EDs adopt a flexible routing policy, under which triage nurses make routing decisions based on both triage scores and other patient and ED factors (which is the practice in our study hospitals). On the other hand, many EDs simply implement triage-score-based routing policies. Specifically, it has been observed in both American (Peck and Kim 2010, Arya et al. 2013, Song et al. 2015) and Canadian EDs (Ding et al. 2019) that all (and only) patients of triage levels 4 and 5 are routed to the FT. Such a policy is simple and easy to implement, but it is rigid and may lead to mismatch between demand and supply in different treatment areas. Meanwhile, flexible policies could route patients with similar clinical conditions to different ED areas under different congestion conditions, of which the consequence is not clear. Therefore, it is inherently important to establish evidence-based policies to guide FT routing decisions (Peck and Kim 2010), which is the third goal of this study.

To achieve our research goals, we obtain unique access to a two-year patient health record dataset from two hospitals in Alberta, Canada, which have established dedicated FT areas. Our dataset is unique in that the study hospitals adopt a flexible routing policy such that the FT routing decisions do not entirely rely on triage scores; hence, patient and ED characteristics may also affect FT routing decisions, enabling the investigation of our research questions. Our findings and contributions can be summarized as follows.

First, our work uncovers an important correlation between ED congestion and FT routing decisions (i.e., the likelihood of being routed to FT) made by triage nurses. This finding suggests that FT routing decisions are not purely clinical-driven, and operational factors related to ED congestion are also crucial to the decisions made. As a result, it is possible that patients with similar clinical conditions might receive treatment in different ED areas (i.e., main vs. FT) under different ED congestion conditions. This finding calls for a more comprehensive examination of how FT routing decisions might impact patient outcomes.

Next, we examine the impact of FT routing decisions on patient outcomes. Consistent with earlier medical literature, we find that being routed to FT can improve patient access to emergency care by reducing the average ED LOS. This finding supports the merit of establishing the FT area, that is, to provide fast care delivery and improve emergency care access. However, we also find that being routed to the FT can lead to a 6.8% (6.6%) increase in the 48-hour (72-hour) revisit rate for high-complexity patients and a 5.8% (5.8%) increase in the 48-hour (72-hour) revisit rate for medium-complexity patients. These findings uncover an important trade-off between fast care access and quality assurance.

Finally, we propose evidence-based policies to guide FT routing decisions. In particular, to balance the trade-off between emergency care access and quality assurance, we develop a multi-class queueing model with two stations to study the optimal routing policy. Through a data-calibrated simulation study, we show that our proposed optimal state-dependent routing policy can achieve a 5.44% reduction in the 48-hour patient revisits and 21.9% reduction in the average patient waiting time compared to the current policy used by our study EDs. Moreover, we compare several easy-to-implement heuristic policies and find that a simple static policy can also reduce the 48-hour patient revisits over the current policy by 2.47%. It is worth noting that the triage-score-based policy that simply routes all patients of triage levels 4 and 5 to the FT area has the worst performance (despite being popularly adopted) among all the tested policies, potentially due to its rigid separation of patient streams.

The rest of this paper is organized as follows. Section 2 discusses the relevant literature. Section 3 presents the study setting and our data. Section 4 describes the econometric models. Section 5 shows the main empirical results. Section 6 develops a simulation model to compare the current policy in practice with alternatives and provides policy recommendations. Finally, Section 7 summarizes the main findings and discusses the practical implications.

2. Literature Review

In recent years, studies on healthcare worker behaviors, especially in congested systems, have attracted growing attention from the operations management (OM) community (KC et al. 2020). Existing studies have shown that healthcare workers respond to system congestion and heavy workload by varying their behavior and rationing decisions, which leads to, among others, accelerated service (KC and Terwiesch 2009, Long and Mathews 2018), compromised patient safety (Kuntz et al. 2015), early task initiation (Batt and Terwiesch 2016), higher referral rates (Freeman et al. 2017), increased post-ED care utilization (Soltani et al. 2022), biased admission decisions (Kim et al. 2020), and patient undercoding (Powell et al. 2012). Our study contributes to this stream of literature by uncovering a positive relationship between ED congestion and FT routing decisions. Therefore, despite that the purpose of FT is to provide fast care delivery to patients with less urgent conditions, FT routing decisions are not purely clinical-driven, and operational factors related to ED congestion are also critical in making the decision. We then subsequently examine the impact of the FT routing decision on patient outcomes and propose new routing policies.

Consequently, our work is closely related to studies that empirically examine the impact of routing decisions in healthcare settings (e.g., Kim et al. 2015, Chan et al. 2018, and Song et al. 2020). In particular, Kim et al. (2015) investigate the impact of the routing decisions (i.e., admission or denied admission) to a hospital's intensive care unit (ICU) on patient outcomes. By quantifying the cost of denied ICU admission, they provide a simulation framework to compare various admission strategies. Chan et al. (2018) empirically estimate the costs and benefits associated with routing patients to the general wards, ICUs, and step-down units. To address the uncertain patient needs, the authors propose a data-driven approach to classify patients based on their severity. Song et al. (2020) study the off-service placement in hospitals, i.e., routing a patient to hospital beds designated for a different service due to capacity constraints on the unit designed for this patient's service needs. Routing decisions in other healthcare settings have also been investigated. For example, Lu and Lu (2018) probe the inter-hospital routing of heart attack patients, and Webb and Mills (2019) discuss how to increase the pre-hospital triage adoption so as to route patients to appropriate care providers before transport to the ED. Interested readers can refer to Section 3.3 in KC et al. (2020) for a review of studies on patient routing decisions in healthcare systems. Built upon earlier works, this paper contributes to the literature by studying triage nurses' FT routing decisions and examining their impact on emergency care access and quality assurance.

Next, motivated by our empirical results, we devise new routing policies through the analysis of a queueing model with multiple classes of customers and multiple pools of servers, where customer classes refer to patient complexity groups and server pools represent the main and FT treatment areas. Hence, our work is also related to the literature on skill-based routing in service systems (Gans et al. 2003); see also Chen et al. (2020a) for an overview that highlights the complications brought by healthcare applications. The models reviewed in Chen et al. (2020a) assume that the routing decision for a customer is only made when at least one server becomes available to serve the customer; therefore, there is no forced idling in their models. In our model, however, patients are routed to one of the two queues with dedicated servers upon their arrival to the ED, which creates the "anti-pooling" effect (i.e., servers at one queue may be idle while servers at the other queue are overwhelmed). In the setting of parallel symmetric queues, the policy that routes customers to the shortest queue has been shown to be optimal (Winston 1977, Weber 1978). When it comes to routing between asymmetric queues, the optimal routing policy can be described by a monotone switching curve (Hajek 1984, Xu and Zhao 1996). Our study differs in that existing works focus on systems with homogeneous customers, whereas our model considers patient heterogeneity based on their complexity levels. We note that the routing mechanism in Xu et al. (1992) is similar to ours. They study the routing decision in a system with two stations (each with parallel servers) serving two classes of customers. However, in their setting, customers of class 1 can only be served by a designated station, whereas an incoming customer can be served by any station in our model. Although our work does not focus on theoretically analyzing our proposed setting, our data-driven simulation study with the Markov decision process formulation enables us to evaluate and compare different FT routing policies, propose the optimal state-dependent policy, and provide managerial implications for practitioners. Moreover, our study adds to this stream of literature by introducing a new practice-driven application setting, which might be of interest to future theoretical research.

Finally, as an initiative to improve ED front-end operations, the effectiveness of introducing FT has been investigated in the emergency medicine literature; see, e.g., Sanchez et al. (2006), Ieraci et al. (2008), Devkaran et al. (2009), Chrusciel et al. (2019) and Grant et al. (2020). Most existing papers in this stream of literature conclude that the implementation of FT improves ED efficiency by reducing the average patient waiting time, LOS, and the rate of LWBS; see a recent review of Grant et al. (2020) on this stream of studies. So far, only two papers (see Ieraci et al. 2008 and Chrusciel et al. 2019) have documented potential adverse effects of the FT area. In particular, Ieraci et al. (2008) use t-tests along with linear and logistic regressions to compare patient outcomes before and after the implementation of FT area and find a slight increase in the 48-hour revisit rate for patients discharged from the ED. One limitation of this observational pre-post analysis is the potential existence of the temporal trend for the 48-hour revisit rate during the study period. Besides, as noted by the authors, the net effect of introducing FT could be confounded by the addition of new staff and physicians to the FT area. More recently, Chrusciel et al. (2019) find a rise in the 7- and 30-day readmission rates after the implementation of the FT (although the readmission rates are not the key focus of the paper) by comparing the sample average before and after the implementation with t-tests. However, this approach ignores potential individual-level confounders. Therefore, to examine how FT routing decisions impact patient outcomes (especially whether potential adverse effects might exist), it is crucial to have a more comprehensive empirical examination with patient-level analysis and control for potential confounders. Moreover, to devise evidence-based routing policies and improve FT routing performance, it is also critical to have a more accurate estimation of the effects of FT routing decisions on patient outcomes. Therefore, our work contributes to this stream of literature by (i) documenting an important correlation between ED congestion and FT routing decisions, (ii) providing a comprehensive examination with patient-level analyses to uncover potential adverse effects of being routed to FT, and (iii) using the estimated results to propose new evidence-based routing policies that are potentially implementable by hospital EDs.

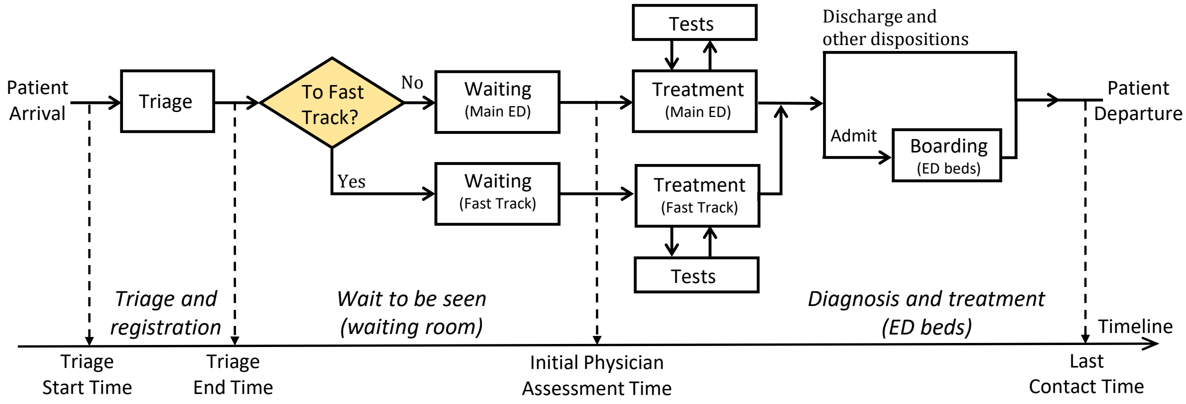
3. Study Setting and Data

This section describes our study setting with details on the ED patient flow process and the data used to conduct our analyses. Section 3.1 describes the setting, Section 3.2 presents the details of our data, and Section 3.3 discusses the choice of key variables used in the analyses.

3.1. Patient Flow Process

We first describe our setting with details on the patient flow process in our collaborator hospitals. The two EDs adopt a similar patient flow process, depicted in Figure 1. Note that our description is based on EDs in Alberta, Canada, and EDs of different regions may operate differently. However, we believe that the key features are shared in most EDs.

Figure 1 A depiction of the ED patient flow process with the timeline of a patient's care pathway.



Upon arrival at the ED, patients are first assigned a triage score, following the CTAS protocol. The timestamps at the start and end of the triage process are referred to as triage start and end times, respectively. The time duration between triage start and end is referred to as the *triage time*, during which triage nurses assign triage scores and route patients to either the main ED or the FT area. These two are separate treatment areas with separate medical facilities and dedicated care teams while sharing the same pool of attending physicians and having similar configurations except that the FT area contains fewer beds and physicians. In other words, a physician may serve in the main area in one shift and in the FT unit at a different time; however, each physician is dedicated to one area during each specific shift. During the study period, the FT area operates 10 hours every day in the two EDs from 10 am to midnight. The average daily traffic (i.e., number of patient arrivals) to the two EDs are 178.9 and 183.7, respectively; the average daily traffic to the FT areas of the two EDs are 33.4 and 41.7, respectively.

After triage, patients wait in the waiting room until being signed up by physicians, and this timestamp is the start of the *initial physician assessment*. The period between triage end time and initial physician assessment time is referred to as the *patient waiting time*. When the ED treatment is completed, physicians make disposition decisions. After that, patients are either discharged home or admitted to the hospital, and the corresponding time is the last contact time. The period between the initial assessment time and the last contact time is the *diagnosis and treatment time*. Finally, the period from the triage end time to the last contact time is referred to as the ED *length of stay* (LOS).

3.2. Data Description

Our data contain patient visit records from the EDs of two urban hospitals in Alberta, Canada, from August 2013 to July 2015, involving a total of 264,551 visits from 169,752 patients. Note that a patient may have visited the EDs more than once during the study period. Each observation in our data includes patient demographics (e.g., age and gender) and the details of their ED visits (e.g., chief complaint, triage score, and attending physician ID).

We now discuss how we clean our data for empirical analyses. To start with, we exclude patient visits that occurred outside the FT operation time (i.e., from midnight to 10 am), which leaves us with a total of 203,974 observations (22.9% removed) from 139,830 patients. We then remove observations with LOS greater than 48 hours, as those extreme cases could bias our results (Song et al. 2015). Note that only 299 observations are identified with LOS greater than 48 hours in our data. Next, we remove observations of the first and last weeks of our study period to avoid censored estimates (Kim et al. 2015, Song et al. 2020, Chan et al. 2018), which deletes 3,595 observations (1.4% of the original data). We further exclude patients coming to the ED through ambulance, because those patients are normally associated with urgent healthcare conditions that require immediate care from physicians. In addition, we remove patients with dispositions of "left without being seen," "left against medical advice," and "transferred," because those patients did not receive care from their visits. These two steps leave us with 143,566 observations (21.4% of the original data removed) from 106,510 patients, which are used for the patient classification in Section 4.4 and the simulation study in Section 6. For our main empirical analyses in Section 4.3, we further remove patients without hospital discharge information and patients of triage level 1 (7.5% of the original data). Note that we need discharge information to compute outcome measures of 48- and 72-hour revisits. Besides, we exclude patients of triage level 1, as their conditions are usually very urgent, requiring immediate attention (Ding et al. 2019). Later in Section 5.3, we also conduct a robustness check including patients with triage score 1 and show that our main results still hold. These two steps leave us with 123,655 observations from 94,448 patients used for our empirical analyses.

3.3. Choice of Variables

This section presents the choice of key variables used in our empirical analyses; see Table 1 for the summary statistics.

Table 1 Summary statistics of key variables.

	Main Area				Fast-Track Area			
Variables	Mean	SD	Min	Max	Mean	SD	Min	Max
Patient Outcomes								
$Revisit_{48h}$ (%)	6.63	24.88	0	100	3.89	19.35	0	100
$Revisit_{72h}$ (%)	7.89	26.96	0	100	4.71	21.19	0	100
LOS (in hours)	4.29	2.90	0.00	39.84	2.60	1.69	0.04	25.30
Operational Characteristics								
EDCongestion	0.78	0.14	0.13	1.31	0.79	0.14	0.14	1.30
MainCongestion	0.78	0.14	0.13	1.32	0.78	0.15	0.15	1.32
FTCongestion	0.60	0.27	0	2.00	0.59	0.27	0	1.91
AvgOccTreated	0.58	0.21	0	1.30	0.48	0.23	0	1.25
WaitTime (in hours)	1.56	1.25	0	17.64	1.34	0.95	0	9.97
TriageTime (in minutes)	4.50	1.88	0.62	43.43	3.82	1.62	0.70	49.22
Physician Characteristics								
Workload	3.59	2.36	0	18	2.69	1.80	0	14
Patient Characteristics								
Gender (Male %)	40.83	49.15	0	100	55.43	49.70	0	100
Age group (%)								
0 to 25 years	18.02	38.44	0	100	21.51	41.09	0	100
25 to 40 years	30.80	46.17	0	100	29.64	45.67	0	100
40 to 55 years	22.18	41.54	0	100	22.14	41.52	0	100
55 to 70 years	17.03	37.59	0	100	17.12	37.67	0	100
Over 70 years	11.97	32.46	0	100	9.58	29.43	0	100
Triage score (%)								
CTAS 2	35.54	47.86	0	100	15.13	35.84	0	100
CTAS 3	44.68	49.72	0	100	37.26	48.35	0	100
CTAS 4	15.06	35.76	0	100	33.12	47.06	0	100
CTAS 5	4.73	21.22	0	100	14.49	35.20	0	100
N	85,091				38,564			

Notes. SD = standard deviation; CTAS = Canadian Triage and Acuity Scale.

3.3.1. Dependent Variables We consider three outcome measures: the 48-hour revisit rate, the 72-hour revisit rate, and patient LOS, denoted by $Revisit_{48h}$, $Revisit_{72h}$, and *LOS*, respectively. The variable $Revisit_{48h}$

($\mathit{Revisit}_{72h}$) equals 1 if the patient visited one of the two EDs within 48 (72) hours after being discharged and 0 otherwise. The 48- and 72-hour revisit rates are widely used in the healthcare literature to measure the quality of emergency care (e.g., Ieraci et al. 2008, Trivedy and Cooke 2015, Song et al. 2015, and Batt et al. 2019). Later, we also consider a robustness check with the 7-day revisit rate in Section 5.4 and show consistent results. Finally, as mentioned in Section 3.1, the variable *LOS* is the time duration from the triage end time to the last contact time.

3.3.2. Independent Variables Next, we describe the independent variables in our estimation. The key variable of interest in our study is the FT routing decision for patient i, denoted by FT_i, which equals 1 if patient i is routed to the FT area and 0 otherwise. In what follows, we discuss a set of control variables on the system-, physician-, and patient-level operational metrics and patient characteristics.

We start with the system-level operational metric: the area occupancy level during patient i's treatment period, denoted by $\mathit{AvgOccTreated}_i$. Following similar ideas in Kim et al. (2015), Chan et al. (2018), and Song et al. (2020), we define the area occupancy level as the total number of hours other patients spend in a particular area (i.e., main or FT) during patient i's stay in this area divided by the length of treatment for patient i. Next, we introduce the physician-level operational metric: physician workload $\mathit{Workload}_i$. Following the earlier literature (Song et al. 2015, Soltani et al. 2022), physician workload is defined as the number of patients other than patient i that have been assigned to patient i's attending physician and have yet been discharged at the time when patient i is assigned. Besides, we include two patient-level operational characteristics: waiting time ($\mathit{WaitTime}_i$) and triage time ($\mathit{TriageTime}_i$). The patient waiting time is the period from triage end till patient i was picked up by a physician in a particular area. The triage time is the period from the triage start to end.

Finally, we include the following patient characteristics: age, gender, triage score, and chief complaints. To account for the possible nonlinear effect of age, we use categorized age groups instead of numerical values. We then use triage score to control the patient's urgency level. Besides, we control the heterogeneity in patient health conditions within the same triage score using chief complaint codes, which are categorical variables with 170 levels in our data, such as "abdominal pain," "upper extremity injury," and "shortness of breath." To reduce the dimension, especially for complaints with very few observations, we follow the chief complaint classification protocol in Grafstein et al. (2003) and group the 170 complaints into 18 major categories. Later in our robustness check, we also incorporate patients' comorbidity information to show the consistency of our main results.

4. Econometric Model

This section describes the econometric model and identification strategy used in our paper. Section 4.1 presents the baseline econometric model and describes potential endogeneity issues involved. Section 4.2 then explains our identification strategy with instrumental variables (IVs). Section 4.3 introduces the details of the two econometric models (with IVs) used for our main analyses. Finally, Section 4.4 discusses the patient classification method for our subgroup analyses.

4.1. Baseline Econometric Model

Our paper aims to understand the impact of FT routing decisions on patient outcomes (i.e., 48- and 72-hour revisits and ED LOS). The best way to quantify the impact on *revisit* and *LOS* is through field experiments by randomly assigning patients to either the main or FT area. However, this method is impracticable for various reasons, including ethical concerns. Therefore, we use retrospective observational data to answer this question instead. We start with the following baseline econometric model for patient i:

$$Outcome_i = \tilde{\beta}\mathbf{X_i} + \tilde{\gamma}FT_i + \tilde{\omega}_h + \tilde{\tau}_m + \tilde{\theta}_t + \tilde{\xi}_i, \tag{1}$$

where the dependent variable $Outcome_i$ represents either the binary outcome measures on 48- and 72-hour revisits or continuous outcome measure on ED LOS, and the vector $\mathbf{X_i}$ includes the age group, gender, chief complaint, triage score, and triage time of patient i. The variables $\tilde{\omega}_h$, $\tilde{\tau}_m$, and $\tilde{\theta}_t$ represent the hospital, month-year, and weekday fixed effects. The error term $\tilde{\xi}_i$ follows a standard normal distribution.

One may estimate the above Equation (1) and then interpret the estimated parameter $\tilde{\gamma}$ as the impact of being routed to FT on patient outcomes. However, such an approach ignores that the FT routing decisions may be endogenous due to factors that were observed by triage nurses when making the decisions but are unobservable in our data, such as patient mental state and the level of pain. These unobserved factors could simultaneously affect both the FT routing decisions and patient outcomes, which raises endogeneity issues and can lead to omitted variable bias in the estimation (Wooldridge 2012). Next, we discuss how we address this issue in our estimation.

4.2. Instrumental Variables

To address the endogeneity issue raised in Section 4.1, we adopt an IV approach. A valid IV should satisfy two requirements: (i) inclusion condition—IVs should be correlated with the endogenous variable; and (ii) exclusion condition—IVs cannot directly affect the dependent variable except through the endogenous variable. Following the empirical healthcare literature, we consider IVs related to operational factors of the ED; see, e.g., Kim et al. (2015), Chan et al. (2018), and Song et al. (2020). Specifically, following closely Song et al. (2020), we propose an ED congestion-related IV: the relative congestion level between the main area and the entire ED at patient i's triage start time, denoted by $MEBusyRatio_i$. To compute this variable, we first measure congestion levels in the main area, the FT area, and the entire ED, denoted by *MainCongestion*, *FTCongestion*, and *EDCongestion*, respectively. This area congestion measure is calculated as the area workload divided by the area capacity. In particular, the area workload is computed as the total number of patients waiting and being treated in this area divided by the number of physicians on duty at that time. Next, the area capacity is defined as the 95th percentile of the distribution of the area workload, where we use the 95th percentile instead of the maximum to avoid observations under extreme situations (Kim et al. 2015). Note that we compute this capacity measure for each hospital separately. In general, the congestion measure here captures the extent to which the area workload takes up to its service capacity. Based on these congestion measures, we can then compute our proposed IV on the relative congestion level between the main area and the entire ED at patient *i*'s triage time. Note that by adjusting the number of physicians on duty, our proposed relative congestion measure is independent of the scale of the supply side. However, from the triage nurses' perspective, the congestion level might be purely determined by the number of patients in a particular area, which is directly observable. Therefore, later in the robustness check in Section 5.3, we also consider an alternative congestion measure without adjusting the number of physicians on duty.

Table 2 Summary statistics of the instrumental variable.

	Main Area				Fast-Track			
Variables	Mean	SD	Min	Max	Mean	SD	Min	Max
MEBusyRatio	1.13	0.07	0.72	1.38	1.13	0.07	0.71	1.38
N	85,091				38,564			

Notes. SD = standard deviation.

Next, we discuss the validity of this IV. We start with the inclusion condition. It has been shown in the earlier work that healthcare admission controllers take into account hospital congestion or utilization when making admission decisions; see, for example, Kim et al. (2015). Similarly, in our setting, when a patient arrives at an ED, without explicit guidelines for FT routing decisions, a triage nurse may consider both clinical and ED operational factors to decide where to route the patient during the triage process. Being aware that a prolonged waiting time may increase the risk of adverse patient outcomes (Guttmann et al. 2011, Maa 2011, Affleck et al. 2013), triage nurses may intentionally route patients to the FT area to reduce their waiting time when the main area is busy, indicating a potential correlation between our relative congestion measure and the FT routing decision. We further validate this inclusion condition statistically through the first-stage regression results (see the full estimation results in Tables 9–12 in the Appendix). The coefficients of our proposed IV $MEBusyRatio_i$ in all these first-stage regressions are statistically significant. Finally, we conduct the weak identification test. The Cragg-Donald Wald F statistics reported for all the estimation equations later described in Section 4.3 are greater than 16.38, which is the critical value of the Stock-Yogo weak IV test (Stock and Yogo 2005). This result indicates that our identification is not weak.

Finally, we discuss the exclusion condition, i.e., the busyness ratio $MEBusyRatio_i$ affects patient outcomes only through the FT routing decision. To start with, we note that our IV (the busyness ratio $MEBusyRatio_i$) measures the relative congestion condition at the starting time of triage. Therefore, ideally, our proposed IV only affects the FT routing decision but not patient outcomes that occurred after the treatment. However, one may argue that the relative congestion condition at the time of triage might be correlated with the area congestion during the patient treatment, thus affecting patient outcomes. Although our proposed IV is a comparison measure (i.e., the relative congestion level between the main area and the entire ED), we still cannot fully rule out the possibility that this relative congestion measure might be correlated with the area congestion during the patient's treatment process. Therefore, we introduce the following two important control variables in our estimation to block the indirect impact of our proposed IV on patient outcomes through channels other than the FT routing decision.

First, following Kim et al. (2015), we control for the area occupancy level ($AvgOccTreated_i$) during the focal patient's diagnosis and treatment period, which allows us to separate the impact of congestion on the FT routing decision from its direct impact on patient outcome. This step is important as earlier work (e.g., Kuntz et al. 2015, Long and Mathews 2018) has shown that area occupancy level might adversely affect patient outcomes. Second, similar to the control on the area occupancy level, we also control for the workload of patient i's attending physician ($Workload_i$) at the time when patient i was assigned to this physician. Following a similar logic, this control variable allows us to separate the impact of area congestion on the FT routing decision from its impact on physician behaviors. This is another important step because earlier work (e.g., KC and Terwiesch 2009) has shown that area congestion increases physician workload, hence leading to physician behavioral changes that might negatively affect patient outcomes. As a result, conditional on the occupancy level and the individual physician workload, the busyness ratio $MEBusyRatio_i$ can only affect patient outcomes through the FT routing decision.

4.3. Estimation

In this section, we describe our IV estimation approaches for patient outcomes. We consider two types of patient outcomes: LOS_i (a continuous variable) and $Revisit_i$ (a binary variable). The variable of interest here is the FT routing decision FT_i. As mentioned earlier, this binary variable of the FT routing decision could be endogenous; hence, we adopt an IV approach to estimate its impact on patient outcomes. We remark that all the continuous variables used in our estimation are standardized (i.e., subtract the mean and then divide by the standard deviation).

4.3.1. Outcome Variable: *LOS* We start with the outcome on *LOS*. Since LOS_i is continuous and the endogenous variable FT_i is binary, directly applying a standard two-stage least square (2SLS) approach to nonlinear models by incorporating the nonlinear first-stage fitted value into the second stage will lead to estimation bias (i.e., forbidden regression, see Angrist and Pischke 2009). Following closely the earlier empirical healthcare literature (see Chan et al. 2018), we consider a similar nonlinear parametric model approach to jointly estimate the FT routing decision model and the patient outcome model. We first model the FT routing decision using a latent variable approach as follows:

$$FT_i^* = \beta \mathbf{X_i} + \alpha MEBusyRatio_i + \omega_h + \tau_m + \theta_t + \varepsilon_i, \tag{2}$$

$$FT_i = \mathbb{1}\{FT_i^* > 0\}, \tag{3}$$

where FT_i^* is the latent variable associated with the binary outcome FT_i. The vector $\mathbf{X_i}$ includes the age group, gender, chief complaint, triage score, and triage time of patient i. The variables ω_h, τ_m, and θ_t represent the hospital, month-year, and weekday fixed effects, respectively, and ε_i is the error term for the FT routing model. We also include our IV ($MEBusyRatio_i$) in the first stage.

Next, we estimate the impact of the FT routing decision on the patient outcome *LOS* using the following second-stage equation:

$$\log(LOS_i) = \beta'\mathbf{X_i} + \gamma FT_i + \delta AvgOccTreated_i + \kappa Workload_i + \omega'_h + \tau'_m + \theta'_t + \xi_i, \tag{4}$$

where FT_i is the binary FT routing decision variable and vector $\mathbf{X_i}$ includes same variables as in Equation (2). As mentioned earlier, we also control for the average treatment area occupancy level ($AvgOccTreated_i$) and the workload of patient i's attending physician ($Workload_i$). Similarly, variables ω'_h, τ'_m, and θ'_t represent the hospital, month-year, and weekday fixed effects, and ξ_i is the error term for the outcome model. Standard errors are clustered at the physician level (i.e., by physician ID). To account for the endogeneity of the FT routing variable in Equation (4), we allow for the error terms ε_i and ξ_i to be jointly distributed as a bivariate normal distribution $\Phi_2(\varepsilon_i, \xi_i; \rho)$ with correlation coefficient ρ. Finally, we jointly estimate the FT routing decision and outcome equations through the full maximum likelihood estimation (FMLE). The dependent variable LOS_i here is log-transformed due to the skewness concern of its distribution.

4.3.2. Outcome Variable: *Revisit* We next consider the binary outcome on patient revisit $Revisit_i$. Specifically, we study both the 48-hour ($Revisit_{48h}$) and 72-hour revisit ($Revisit_{72h}$) for patient i. Because both the endogenous variable (i.e., the FT routing decision) and the outcome variable are binary, directly incorporating the nonlinear first-stage fitted value into the second-stage regression will lead to estimation bias. Therefore, we again follow Kim et al. (2015) and Chan et al. (2018) and use a nonlinear parametric model approach to jointly estimate $Revisit_i$ and FT_i. More specifically, we employ the recursive bivariate probit model (see Maddala 1986, Greene 2018, Kim et al. 2015, Liu et al. 2019, and Chan et al. 2018), which contains two probit models with correlated error terms as follows:

$$FT_i^* = \beta\mathbf{X_i} + \alpha MEBusyRatio_i + \omega_h + \tau_m + \theta_t + \varepsilon_i, \tag{5}$$

$$FT_i = \mathbb{1}\{FT_i^* > 0\}, \tag{6}$$

$$Revisit_i^* = \beta'\mathbf{X_i} + \gamma FT_i + \delta AvgOccTreated_i + \kappa Workload_i + \eta WaitTime_i + \omega'_h + \tau'_m + \theta'_t + \xi_i, \tag{7}$$

$$Revisit_i = \mathbb{1}\{Revisit_i^* > 0\}, \tag{8}$$

where FT_i^* and $Revisit_i^*$ are the latent variables associated with FT_i and $Revisit_i$, respectively; ε_i and ξ_i are the error terms of the FT routing decision and patient outcome models, respectively, and are jointly distributed following a bivariate normal distribution $\Phi_2(\varepsilon_i, \xi_i; \rho)$ with correlation coefficient ρ. We further control for a patient's waiting time $WaitTime_i$. Note that $WaitTime_i$ is not included in the LOS regression model in Equation (4) because LOS_i is the sum of $WaitTime_i$ and patient i's diagnosis and treatment time (see Figure 1). If we control $WaitTime_i$ (equivalent to conditional on patient waiting time), Equation (4) examines the variation in the diagnosis and treatment time only. All other control variables are the same as those described in Section 4.3.1. Finally, we cluster standard errors at the physician level and estimate the model through FMLE.

4.4. Patient Classification

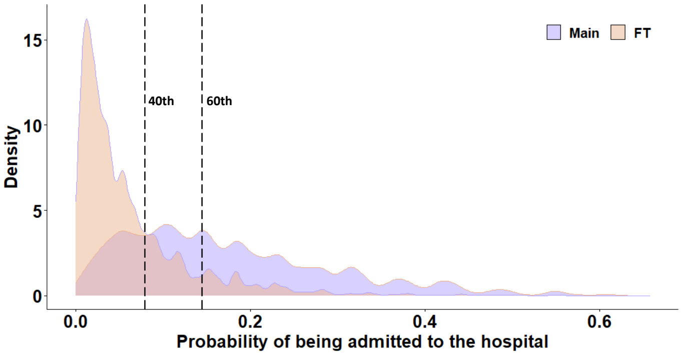
As mentioned earlier, the FT area is designated to treat patients with less urgent and less complex health issues so as to deliver care more quickly. However, triage nurses may consider both clinical and operational factors when making routing decisions, given the lack of consistent guidelines for the FT routing process. As a result, patients with similar clinical conditions might receive treatment in different ED areas (i.e., main vs. FT) under different congestion conditions. It is thus unclear whether any hidden unintended consequences may occur. Moreover, the impact of FT routing decisions might vary across patients of different complexity levels. For instance, patients with high-complex conditions (who should be routed to the main area under less congested situations) may have been routed to the FT area when the main area is highly crowded and may experience adverse outcomes. On the other hand, patients with low-complex conditions might not experience adverse effects or even benefit from being routed to the FT area. However, since there is no consistent guideline for who should be treated in the FT area, such patient complexity categorization could be highly varied across hospitals or even across triage nurses (especially when hospitals adopt a flexible routing policy such as our studied hospitals). As such, similar to Chan et al. (2018), we consider a data-driven approach to classify patients into different complexity categories.

In this regard, a patient streaming strategy based on predicted disposition (i.e., admitted to the hospital vs. discharged from the ED) has been found to be successful by ED practitioners (O'Brien et al. 2006, Kelly et al. 2007). Moreover, in the OM literature, Saghafian et al. (2012, 2014) demonstrate that streaming patients by the predicted disposition during the triage process can improve ED performance. Following this line of work, we classify patients into different complexity levels based on their likelihood of admission. Specifically, we consider disposition decision as the outcome variable and estimate the following probit model:

$$M_i = \begin{cases} 1 \text{ (Admitted to the hospital)} & \text{if } \beta_p \mathbf{X_i^p} + \delta_p EDCongestion_i^d + \phi_i \geq 0, \\ 0 \text{ (Discharged home)} & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathbf{X_i^p}$ is a vector of patient characteristics, including triage score, age group, gender, and chief complaint, and ϕ_i is the unobserved component following a standard normal distribution. Besides, since previous work has shown that ED congestion may affect hospital admission decision (Gorski et al. 2017, Chen et al. 2020b), we also control for the ED congestion at the time when the attending physician makes the disposition decision for patient i, denoted by $EDCongestion_i^d$, to separate the impact of ED congestion and patient characteristics on the admission decision. We then create patient complexity classes by partitioning the fitted probability of admission (denoted as $\hat{M}_i$) based on patient clinical characteristics collected during triage. The fitted probability $\hat{M}_i$ here is computed as $\hat{M}_i = \Phi(\hat{\beta}_p \mathbf{X_i^p})$, where $\hat{\beta}_p$ is the estimated β_p and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Intuitively, the higher the fitted probability, the more likely the patient would be admitted to the hospital, and hence, this patient is more likely to be classified as of a higher complexity level. Figure 2 depicts the fitted probability distribution for patients routed to the main and FT area, respectively. We observe that most patients with a high value of $\hat{M}_i$ were routed to the main area, whereas most patients with a low value of $\hat{M}_i$ were routed to the FT area. This observation is consistent with our intuition that patients with a higher probability of being admitted to the hospital are likely to be higher-complexity patients who should be treated in the main area. Nevertheless, we still observe a few patients with a high value of $\hat{M}_i$ who were routed to the FT area and vice versa. Therefore, we are interested in understanding whether any hidden consequence exists for patients treated in the FT area but would have been routed to the main area in a less congested ED.

Figure 2 **Patient complexity classification based on fitted probability of admission**



Next, we consider the following complexity classification approach: a patient belongs to (i) the high-complexity class if $\hat{M}_i > t_2$, (ii) the low-complexity class if $\hat{M}_i < t_1$, and (iii) the medium-complexity class if $t_1 \leq \hat{M}_i \leq t_2$, where the two thresholds t_1 and t_2 are determined based on the density function of the fitted probability $\hat{M}_i$. Following a similar logic as the thresholds choice in Chan et al. (2018), a larger t_2 increases the percentage of patients with $\hat{M}_i > t_2$ being routed to the main area; similarly, a smaller t_1 increases the percentage of patients with $\hat{M}_i < t_1$ being routed to the FT area. The goal of our selected thresholds (t_1, t_2) is then to balance the increasing percentage of patients in the high- (low-) complexity group being routed to the main (FT) area while maintaining a large enough patient sample in each group for meaningful statistical analyses. As a result, we find the 40th (t_1) and 60th (t_2) percentiles of $\hat{M}_i$ achieve a proper balance. Tables 6 and 7 in the Appendix present the summary statistics of patient characteristics and outcomes (i.e., revisits and *LOS*) for the three complexity classes. We can then estimate the impact of FT routing decisions on patient outcomes based on patient complexity subgroups. Later, we also conduct robustness checks with alternative choices of t_1 and t_2 and show our empirical results remain consistent.

5. Estimation Results

In this section, we present our estimation results. Section 5.1 discusses the correlation between operational status and FT routing decisions. Section 5.2 examines the impact of FT routing decisions on patient outcomes and its heterogeneous effects across patient complexity classes. Section 5.3 presents our robustness checks.

5.1. Correlation Between Operational Status and FT Routing Decisions

We start our discussion with the relationship between operational status and FT routing decisions. In particular, we employ the following probit model:

$$FT_i = \begin{cases} 1 & \text{if } \beta \mathbf{X_i} + \alpha MEBusyRatio_i + \omega_h + \tau_m + \theta_t + \varepsilon_i > 0, \\ 0 & \text{otherwise}, \end{cases} \tag{10}$$

where vector $\mathbf{X_i}$ again includes the age group, gender, chief complaint, triage score, and triage time of patient i. The variables ω_h, τ_m, and θ_t represent the hospital, month-year, and weekday fixed effects. The error term ε_i follows a standard normal distribution. The variable of interest $MEBusyRatio_i$ here measures the relative congestion level between the main area and the entire ED. Table 3 below presents the estimation results for all patients as well as patients in each complexity group; see Table 8 in the Appendix for the full estimation results. In addition, we also include the average marginal effect (AME) of $MEBusyRatio_i$ on the FT routing decision FT_i computed based on the estimated coefficients.

Based on results in Table 3, we find that the coefficient of $MEBusyRatio_i$ is positive and significant (p-value < 0.01) for all the analyses (i.e., all patients sample, high-, medium-, and low-complexity groups), indicating a positive correlation between the relative congestion level of the main area to the entire ED and the likelihood of being routed to the FT area. Specifically, based on the AME in Table 3, we find that a 10% increase in $MEBusyRatio_i$ is associated with a 1.0%, 1.8%, and 2.3% increase in the likelihood of being routed to the FT area for the high-, medium-, and low-complexity groups, respectively. Note that the AME values in Table 3 are based on the standardized value of $MEBusyRatio_i$; hence, we cannot use the AME values directly and the calculation of the percentage changes involves transforming $MEBusyRatio_i$ back to its original scale. These results suggest that FT routing decisions are not purely clinical-driven, ED operational status related to congestion is also a critical factor in the FT routing decision-making process. Besides, based on results in Table 8 in the Appendix, we find that clinical factors, such as age group, triage score, gender, and triage time, are also associated with FT routing decisions.

5.2. Impact of FT Routing Decisions on Patient Outcomes

This section discusses our main results on the impact of FT routing decisions on patient outcomes. Table 4 presents results both with and without IV to illustrate the potential estimation bias without IV. Note that instead of the estimated coefficient of the FT routing variable FT_i (γ in Equations (4) and (7)), we present the AME of FT_i on each patient outcome variable for the interpretation purpose. The estimated coefficient γ can be found in Tables 9–12 in the Appendix, which present the full estimation results.

Table 3 **Results on the correlation between operational status and the FT routing decisions**

	All patients	High-complexity	Medium-complexity	Low-complexity
MEBusyRatio	0.073^{***}	0.073^{***}	0.080^{***}	0.074^{***}
	(0.005)	(0.011)	(0.012)	(0.007)
AME	0.012^{***}	0.007^{***}	0.013^{***}	0.016^{***}
	(0.001)	(0.001)	(0.002)	(0.002)
N	123,655	50,514	23,377	49,434

Standard errors in parentheses. Some observations are dropped due to the perfect separation. See Table 8 in the Appendix for the full estimation results. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

To start with, we consider analyses with all patients. From panel A of Table 4, we find that being routed to the FT area reduces the average *LOS* (i.e., a negative AME of -0.247 with p-value < 0.01). To understand the *LOS* reduction in hours, we compute the predicted values of *LOS* when patients were routed to the main versus FT area using our estimation results, which gives us an average reduction of 0.60 hospital hours in *LOS* (i.e., $3.43 - 2.83 = 0.60$). We remark that here we cannot directly interpret the *LOS* reduction in hours using AME values in Table 4 because the dependent variable *LOS* is log-transformed and the AME measures the marginal effect of log(*LOS*). As a result, following a similar approach in Chan et al. (2018), we interpret our results using predicted values. Although earlier medical literature has shown the effectiveness of FT on reducing patient *LOS* (see Sanchez et al. 2006, Devkaran et al. 2009, Chrusciel et al. 2019, and Grant et al. 2020), our work further validates this result with a more comprehensive approach and a new hospital setting (data) in Canada. Besides, by simply estimating the impact of FT routing on patient revisits using all patients data without considering the variation in care needs across patient complexity groups, we do not find statistically significant effects on $Revisit_{48h}$ or $Revisit_{72h}$. However, as we have discussed earlier, the impact of FT routing decisions might vary across different patient complexity conditions. Therefore, we proceed to investigate the effects based on patient complexity groups.

5.2.1. High-Complexity Patients Panel B of Table 4 presents the impact of being routed to FT on high-complexity patients. First, we find that being routed to FT reduces the *LOS* for high-complexity patients (a negative coefficient -0.303, p-value < 0.01). Specifically, by computing the predicted values of *LOS* when patients were routed to the main versus FT area using our estimation results, we get an average reduction of 0.86 hospital hours in *LOS* (i.e., $4.12 - 3.26 = 0.86$) for high-complexity patients. Next, by restricting our analyses to high-complexity patients, we find that being routed to the FT area hurts the quality of care by increasing the likelihood of revisits (positive coefficients for the 48-hour and 72-hour revisit: 0.068 and 0.066, respectively, with p-value < 0.05). By interpreting the AME values directly, these results indicate that being routed to the FT increases the 48-hour and 72-hour revisits by 6.8% and 6.6%, respectively. Besides, the full estimation results in Table 10 in the Appendix further show that waiting time is also positively correlated with the 48-hour and 72-hour revisits; however, waiting for an additional hour is only associated with a 0.6% increase in both the 48-hour and 72-hour revisits. These findings call for attention from hospital and ED managers to carefully balance the tradeoff between care access and quality of care.

Table 4 The AME of being routed to FT on patient outcomes.

	With IV			Without IV
Outcome variables	AME (SE)	ρ (SE)	Test $\rho = 0$	AME (SE)
Panel A: All patients ($N = 123,655$)				
$Revisit_{48h}$	-0.003 (0.007)	0.004 (0.033)	0.898	-0.002 (0.004)
$Revisit_{72h}$	-0.006 (0.008)	0.003 (0.032)	0.922	-0.005 (0.004)
$\log(LOS)$	-0.247***(0.055)	-0.016 (0.037)	0.671	-0.266***(0.017)
Panel B: High-complexity patients ($N = 50,768$)				
$Revisit_{48h}$	0.068** (0.027)	-0.157** (0.062)	0.013	0.015** (0.006)
$Revisit_{72h}$	0.066** (0.028)	-0.144** (0.062)	0.022	0.014** (0.006)
$\log(LOS)$	-0.303***(0.055)	-0.045 (0.035)	0.204	-0.361***(0.019)
Panel C: Medium-complexity patients ($N = 23,453$)				
$Revisit_{48h}$	0.058***(0.022)	-0.194***(0.064)	0.003	0.007 (0.006)
$Revisit_{72h}$	0.058** (0.025)	-0.194***(0.067)	0.005	0.002 (0.007)
$\log(LOS)$	-0.457***(0.061)	0.158***(0.044)	0.000	-0.263***(0.021)
Panel D: Low-complexity patients ($N = 49,434$)				
$Revisit_{48h}$	-0.002 (0.009)	-0.048 (0.059)	0.419	-0.009***(0.003)
$Revisit_{72h}$	-0.004 (0.009)	-0.046 (0.054)	0.397	-0.012***(0.003)
$\log(LOS)$	-0.421***(0.072)	0.152***(0.055)	0.006	-0.233***(0.018)

Notes. Standard errors (SEs) clustered by the physician who conducted the initial assessment are shown in parentheses. Controls not shown include patient characteristics, operational factors, and the fixed effects (hospital, month-year, and weekday). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.2.2. Medium-Complexity Patients Next, we discuss the impact on medium-complexity patients; see panel C of Table 4. We again observe that being routed to FT reduces the *LOS* for medium-complexity patients (a negative coefficient -0.457 with p-value < 0.01). Specifically, by computing the predicted values of *LOS* when patients were routed to the main versus FT area using our estimation results, we get an average reduction of 1.04 hospital hours in *LOS* (i.e., $3.49 - 2.45 = 1.04$) for medium-complexity patients. Next, we find positive and statistically significant effects of being routed to FT on both 48- and 72-hour revisits for medium-complexity patients. Specifically, being routed to FT increases the likelihood of both 48-hour and 72-hour revisits by 5.8%. Besides, similar to the analyses for high-complexity patients, the full estimation results in Table 11 in the Appendix again confirm that waiting time is positively correlated with the 48-hour and 72-hour revisits.

5.2.3. Low-Complexity Patients Finally, we discuss the impact on low-complexity patients; see panel D of Table 4. Similar to the previous results, we again find that being routed to FT reduces the *LOS* for low-complexity patients (a negative coefficient -0.421 with p-value < 0.01). Comparing the predicted values of *LOS* when patients were routed to the main versus FT area using our estimation results, we obtain an average reduction of 0.82 hospital hours in *LOS* (i.e., $2.98 - 2.16 = 0.82$) for low-complexity patients. However, for low-complexity patients, we do not find statistically significant effects of being routed to FT on the 48- or 72-hour revisits. This finding supports the purpose of introducing the FT area: to treat low-complexity patients faster and improve operational efficiency without compromising the quality of care.

5.2.4. Discussion on the Likelihood Ratio Tests The third column of Table 4 shows the estimated correlation $\rho(SE)$ between the error terms of the FT routing decision equation and the outcome equation. The fourth column of Table 4 presents the p-values of the likelihood ratio test results "Test $\rho = 0$" that compares the log-likelihood of our full model with the sum of log-likelihood of two separate models. Similar to the Hausman test, the likelihood ratio test checks the exogeneity of a dummy independent variable with a dummy dependent variable (Knapp and Seaks 1998). We can see from panel B of Table 4 that for high-complexity patients, the p-values of the likelihood ratio test for the 48-hour and 72-hour revisits are both less than 0.05, indicating a strong endogeneity issue, which also explains the difference between the results with IV and without IV. Similarly, for medium-complexity patients, the likelihood ratio test indicates the existence of a strong endogeneity issue across all the outcome variables; see panel C of Table 4. Finally, for low-complexity patients, the likelihood ratio test suggests the existence of endogeneity issues in the estimation of the outcome variable *LOS*. These results indicate the importance of adopting an IV approach to get consistent estimates of the impact of FT routing decisions on patient outcomes.

5.3. Robustness Checks

This section presents our robustness checks; see Table 13 in the Appendix for a summary of the estimation results. In particular, we consider alternative IVs, alternative patient classification cutoffs, alternative samples, alternative outcome measures, and alternative model specifications.

5.3.1. Alternative IVs The IV in our main model measures the relative congestion level between the main area and the entire ED. As mentioned in Section 4.2, the congestion level in a particular area is computed as the area workload divided by the area capacity, both of which have been adjusted by the number of physicians on duty. Hence, our IV is independent of the scale of the supply side. However, from triage nurses' perspective, the congestion level might be purely determined by the number of patients in a particular area, which is directly observable. Therefore, we consider an alternative IV that does not adjust the number of physicians. More specifically, the new congestion level is calculated as the total number of patients waiting and being treated in a particular area divided by the area capacity. Panel (1) of Table 13 in the Appendix shows the estimation results using this alternative IV, which are consistent with our main results.

Next, the IV used in our main analyses is computed at the focal patient's triage start time. However, triage nurses may use past congestion information to inform current routing decisions. Therefore, we also consider alternative IVs computed using information that is 0.5, 1, and 2 hours before patient i's triage start time, respectively. Panels (2), (3), and (4) in Table 13 in the Appendix show the estimation results with these alternative IVs, respectively, and we find all the results are consistent with our main findings.

5.3.2. Alternative Patient Classification Cutoffs We next consider alternative cutoffs to partition the patient complexity classes. In our main analyses, the cutoffs t_1 and t_2 are the 40th and 60th percentiles of the distribution of the likelihood of being admitted to the hospital, respectively. To show the robustness of our findings, we consider alternative cutoffs of t_1 and t_2. In particular, we test four pairs of the thresholds (t_1, t_2), i.e., (35, 60)th, (45, 60)th, (40, 55)th, and (40, 65)th percentiles, respectively; see the estimation results in panels (5)–(8) in Table 13 in the Appendix, which are consistent with our main findings.

5.3.3. Alternative Samples We now consider two alternative samples. First, as shown in Table 1, the triage time could be as long as 49.22 mins, which is abnormal. Therefore, we consider an alternative sample by removing extreme observations with triage time longer than 17 mins (i.e., outside the 99.9th percentile); see panel (9) in Table 13 in the Appendix. The estimation results are consistent with our main findings. Second, in our main analyses, we exclude patients of triage level 1, as their conditions are usually very urgent, requiring immediate attention (Ding et al. 2019). We repeat our analysis by including these patients, and the results are shown in panel (10) in Table 13 in the Appendix. Again, we find the estimation results consistent with our main findings.

5.3.4. An Alternative Outcome Measure In addition to the 48- and 72-hour revisits used in our main analyses, the 7-day revisit has also been used in prior studies (e.g., Song et al. 2015 and Michelson et al. 2018) to measure a longer-term impact on patient outcome. Hence, we use the 7-day revisit as the outcome variable and run our analyses again; see Table 14 in the Appendix. We find that being routed to the FT also has a potential longer-term adverse effect in terms of the increased 7-day revisit for high-complexity patients. This result further supports our main findings on the potential hidden consequences on the quality of care.

5.3.5. Alternative Model Specification with Patient Comorbidity It is natural to expect that patient comorbidities may affect patient outcomes and associate with complexity classes. Hence, we would like to conduct a robustness check by controlling patient comorbidity information. However, our dataset does not contain a numerical measure of patient comorbidity that can be directly incorporated into our econometric model. Hence, we use the textual medical history data—collected by triage nurses—to construct the Charlson comorbidity index (Charlson et al. 1987) as the control for patient comorbidity. Since the medical history data only covers 11 months of our study period (from September 2014 to July 2015), we decided to use this information in the robustness check instead of including in our main analysis. Specifically, we first include the Charlson comorbidity index ($Charlson_i$) in the patient classification model discussed in Section 4.4 and then incorporate it into Equations (2), (4), (5), and (7) of our main empirical analyses. The results shown in panel (11) of Table 13 in the Appendix are consistent with our main findings.

6. Evaluation of Alternative Fast-Track Routing Policies

In previous sections, we have empirically investigated the impact of being routed to FT on patient outcomes. In this section, we propose a multi-class routing model with two parallel queues to study the optimal routing policy. Specifically, we model the problem using the Markov decision process (MDP) and leverage our empirical results from Section 5 to estimate the model parameters. We solve for the optimal policy numerically and then draw insights from its structure to propose several heuristic routing policies. Finally, we evaluate the performance of different policies via simulation.

6.1. Model of Fast-Track Routing

We model the ED patient flow process as a multi-class queueing system with two parallel stations. Station 1 represents the main treatment area, and station 2 represents the FT area. Patients of class i arrive to the ED according to a time-homogeneous Poisson process with arrival rate λ_i, where $i = 1, 2, 3$, representing patients of high-, medium-, and low-complexity classes as defined in Section 4.4, respectively. We are aware that a nonstationary Poisson process with time-dependent arrival rates is a better model for the patient arrival process (Kim and Whitt 2014). We make the stationary assumption to simplify the MDP formulation and will relax it in our simulation model. Each station has a single server (which is relaxed to multiple shift-based servers in our simulation model) and a queue with infinite capacity. At station j, the service time (*diagnosis and treatment time*) is independent and identically distributed, following exponential distribution with mean $1/\mu_j,\ j = 1, 2$, for all patients (again, this assumption will be relaxed in the simulation). We further assume that patients are served on a first-come-first-served (FCFS) basis at each station for the MDP formulation. We are aware that ED decision makers do not always adhere to the FCFS rule in real settings (Ding et al. 2019). Hence, in our simulation, the process of selecting the next available patient to treat is formulated by a discrete choice model whose parameters are estimated from our data.

Upon arrival, patients will be routed to one of the two queues by the decision maker (i.e., triage nurses), waiting to be seen. If a patient of class i is routed to queue j, a cost $r_{ij}(t)$ is incurred upon the completion of service at station $j,\ j = 1, 2$, given that the patient waited t units of time in the queue before being seen by a physician. The cost is associated with the inconvenience of waiting and fees encountered if a patient needs to revisit the ED within a short period of time (e.g., 48 hours) after being discharged from the ED, which also reflects the quality of care. The dependence of $r_{ij}(t)$ on the station and patient class reflects the discrepancy in the quality of care between the main area and the FT area for patients of different classes (see Table 4). The cost also depends on the patient's waiting time, as shown by our empirical results (see Tables 10–12 in the Appendix), which aligns with the literature (Guttmann et al. 2011). Note that the dependence of the cost term on a patient's characteristics (e.g., age, gender) is reflected by the patient's class. In our simulation study, we explicitly account for patient characteristic information when estimating the cost term $r_{ij}(t)$. The decision maker's objective is to find a routing policy to minimize the expected long-run average cost over an infinite time horizon. Note that we assume any class of patients can be routed to any queue to keep our model general. However, as we show later in Figure 4, the optimal policy rarely routes any patient of high-complexity level to the FT area based on the model parameters estimated from our data.

6.1.1. The MDP Formulation Next, we formulate the decision problem for FT routing using an MDP formulation. The decision epochs correspond to patient arrival times to the ED. Denote the system state at time t by $\boldsymbol{x} = (x_1, x_2)$, where x_1 and x_2 represent the number of patients in the main and the FT area, respectively. Hence, the state space is $\mathcal{S} \equiv \{\boldsymbol{x} = (x_1, x_2) : x_i \in \mathbb{N}, i = 1, 2\}$. Upon the arrival of a new patient, the triage nurse needs to decide which area to route this patient to after triage. Hence, the action space is $\mathcal{A} \equiv \{1, 2\}$, where 1 and 2 represent routing the patient to the main and the FT area, respectively.

Let $V_t(\pi, \boldsymbol{x})$ be the total expected t-period cost starting from state $\boldsymbol{x}$ under policy π, which is a sequence of decision rules that map from $\mathcal{S}$ to $\mathcal{A}$ to specify the actions taken at any state and time. Then, the expected long-run average cost starting from state $\boldsymbol{x}$ under policy π is defined as $g(\pi, \boldsymbol{x}) = \limsup_{t\to\infty} V_t(\pi, \boldsymbol{x})/t$, $\forall \boldsymbol{x} \in \mathcal{S}$, and the optimal expected long-run average cost is defined as $g^*(\boldsymbol{x}) = \inf_\pi g(\pi, \boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{S}$. Following Lippman (1975), we apply *uniformization* with the uniformization constant $\Gamma = \sum_{i=1}^{3} \lambda_i + \sum_{j=1}^{2} \mu_j$. Without loss of generality, we can redefine the time unit so that $\Gamma = 1$, and then λ_i and μ_j become, respectively, the probability that the next uniformized transition is a new arrival from class i and a service completion at station j, where $i = 1, 2, 3$ and $j = 1, 2$. Let $v(\boldsymbol{x})$ be the relative value function, $\boldsymbol{e_1} \equiv (1, 0)$, and $\boldsymbol{e_2} \equiv (0, 1)$. Then, the Bellman equation can be written as $g + v(\boldsymbol{x}) = Tv(\boldsymbol{x})$, where g is the optimal long-run average cost, and the operator T is defined as

$$Tv(\boldsymbol{x}) = \sum_{i=1}^{3} \lambda_i \min_{j\in\mathcal{A}} \left\{ r_{ij}(x_j/\mu_j) + v(\boldsymbol{x} + \boldsymbol{e_j}) \right\} + \sum_{j=1}^{2} \mu_j v(\boldsymbol{x} - \mathbb{1}_{\{x_j \geq 1\}} \boldsymbol{e_j}), \forall \boldsymbol{x} \in \mathcal{S}, \tag{11}$$

where $\mathbb{1}_{\{x_j \geq 1\}} = 1$ indicates $x_j \geq 1$, and $\mathbb{1}_{\{x_j \geq 1\}} = 0$ indicates otherwise. Note that we estimate the waiting time of patient i who joins queue j by x_j/μ_j in our MDP formulation since the service times are station-specific and the service discipline at both queues is assumed to be FCFS. Hence, the expected waiting time of a patient is uniquely determined by the number of patients in the queue upon this patient's arrival. In our simulation, we use the actual waiting time so that our model can better reflect reality.

6.1.2. Solve for the Optimal Policy A theoretical study of the optimal policy of our MDP would be of interest. However, it deviates from the main focus of this paper, so we leave it for future research. The relatively low dimension of the MDP allows us to focus on numerical solutions instead. Hence, we solve the MDP by the value iteration algorithm with the value iteration operator defined in (11). The arrival rates and service times are estimated from data under the stationary assumption. It is however challenging to estimate the cost terms $r_{ij}(t)$, $i = 1, 2, 3,\ j = 1, 2$. Next, we leverage the results of our econometric model for the binary patient outcome variable to estimate the 48-hour revisit cost for a class i patient with characteristics $\mathbf{X}$ who joins queue j and waits t units of time before being seen by physicians as follows:

$$r_{ij}(t) = \mathrm{E}\left(Revisit_i | FT_i = j, \mathbf{X}\right) = \mathrm{P}\left(\xi_i \geq -\beta_i \mathbf{X} - \mathbb{1}_{\{j=2\}} \gamma_i - h_i t\right), \tag{12}$$

where γ_i is the coefficient of FT_i estimated from Equation (7), h_i is the cost per unit time a class i patient waits in the system, and ξ_i is the error term that follows a standard normal distribution based on the observed information from data. Note that for each class i patient, we compute costs associated with both $FT_i = 1$ and $FT_i = 2$ for the optimization problem.

6.2. Simulation Design, Input Modeling, and Validation

We build a discrete event simulation model to simulate the ED patient flow process. The objective of the simulation is to compare different routing policies, which will be described in Section 6.3. Next, we describe the simulation design, input modeling, and validation in detail.

Patient Arrival. The patient arrival process is modeled as a nonstationary Poisson process with a time-dependent rate based on hourly resolution. Upon each arrival, we randomly draw a patient from the corresponding set of patients that arrive at this time of the day in our dataset and apply this patient's information (e.g., age, gender, and triage score) to the newly arrived patient in our simulation. We then follow the approach in Section 4.4 to determine the patient's complexity class.

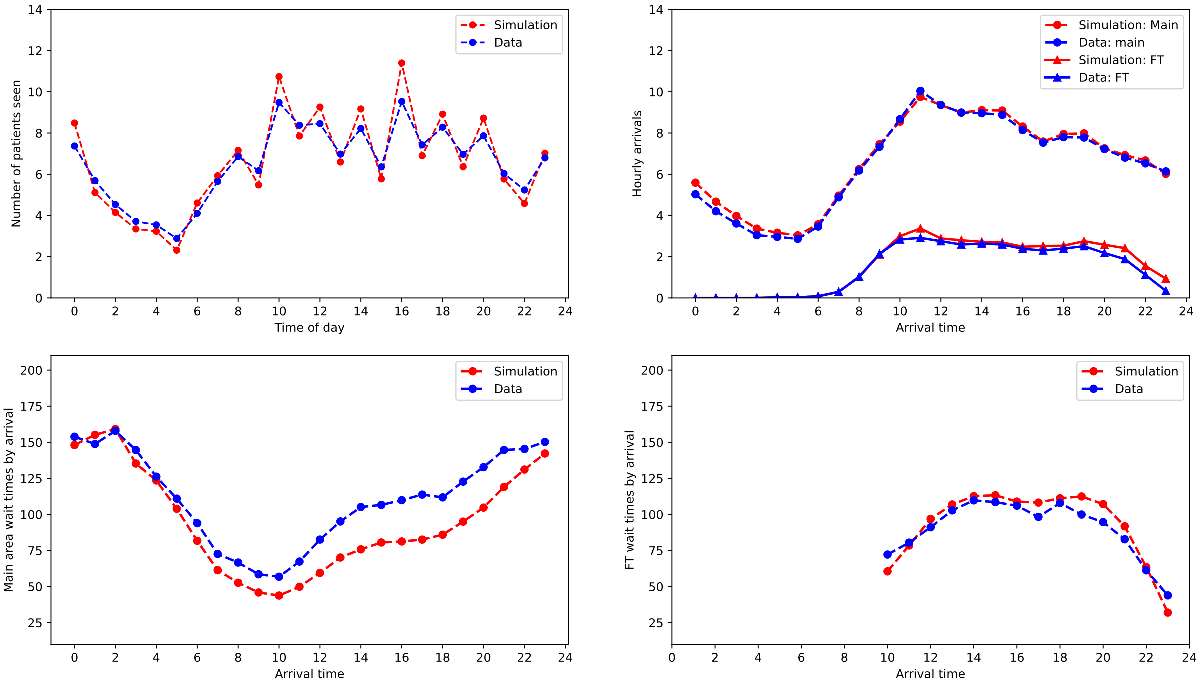
Patient Routing and Abandonment. Based on predefined routing policies (see more details in Section 6.3), patients will be routed to the main treatment area or the FT area and join the corresponding queue. An exponentially distributed patience time is generated for each patient upon joining the queue, and the patient will leave the ED if her waiting time exceeds this patience time and the disposition of this patient will be considered as LWBS. The FT area in our study hospitals operates from 10 am to midnight; hence, no patients will be routed to FT outside this period. When the FT area closes at midnight, we assume that an exhaustive service discipline is applied (Ingolfsson et al. 2007), i.e., the FT physician completes the treatment of the patient whose diagnosis is in process before they leave work. Other patients waiting in the FT area are moved to the main area instantaneously.

Service Process. Physicians go to work according to a shift-based schedule. In the simulation, we use the actual schedule from our study hospital. As a result, the number of physicians on duty is time-varying, determined by the shift schedule (the FT area always has one working physician). We assume that physicians do not idle if there are patients waiting to be seen. Physicians select the next patient to treat based on a discrete choice model, in which a patient's priority of being seen mainly depends on the triage score and the current waiting time (each triage level is associated with a quadratic marginal waiting cost function, see, e.g., Ding et al. 2019). We generate the exponentially distributed service times with the rates given by the number of new patients seen by a physician at the corresponding shift hour observed from the data. This level of abstraction has been shown to be sufficient to generate dynamics that match the actual ED process (Ouyang et al. 2021).

Implementation and Validation. The simulation model is written in Python using SimPy 4.0. For the purpose of model validation, we start the simulation with an empty ED and run 30 replications with a replication length of 500 weeks (the first 100 weeks are identified as the warm-up period and thus removed). The routing policy used in the simulation for validation is based on the estimated current routing policy (Policy CP in Section 6.3). All parameters related to the inter-arrival time generation, service time generation, and the current routing policy are estimated using data from one of our study hospitals from January 2015 to July 2015, as the shift schedule was fixed during this period.

The average patient waiting times from the simulation and the data are shown in the bottom two panels of Figure 3 for the main and the FT areas, respectively, which provide evidence that our simulation model captures the trend of the average waiting time from the data reasonably well. We also compare the simulated number of patients seen by all physicians on duty per hour and the hourly arrivals with that from the data (shown in the top two panels of Figure 3), which further shows the validity of our simulation model.

Figure 3 Comparison of the number of patients seen, the number of patients arrived, and the average patient waiting times between the simulated and the real data.



6.3. Fast-Track Routing Policies

In this section, we compare five FT routing policies through simulations. We first describe the policies of interest explicitly.

Current Routing Policy (CP): We first estimate the current routing policy implemented in our study hospitals. Particularly, we estimate the following probit model based on the patient's characteristics and ED system state to predict the patient's disposition: $FT(\mathbf{X_i}, x_1, x_2) = \mathbb{1}\left(\beta\mathbf{X_i} + \nu_1 x_1 + \nu_2 x_2 + \nu_3 x_1^2 + \nu_4 x_2^2 > \epsilon_i\right)$, where $\mathbf{X_i}$ represents patient characteristics, such as age group and gender; and x_1 and x_2 are the numbers of patients waiting in the main and the FT area, respectively. We include both the linear and quadratic terms of x_i, $i = 1, 2$ to account for potential non-linear effects.

Optimal Routing Policy (OP): We follow the procedure described in Section 6.1.2 to solve for the optimal routing policy based on our MDP. Note that the MDP formulation assumes time-independent patient arrivals and transitions. Hence, the optimal policy for the MDP model is *not* necessarily the optimal policy for our simulation setup.

Figure 4 **An illustration of the optimal routing policy (policy OP) used in our simulation study.**

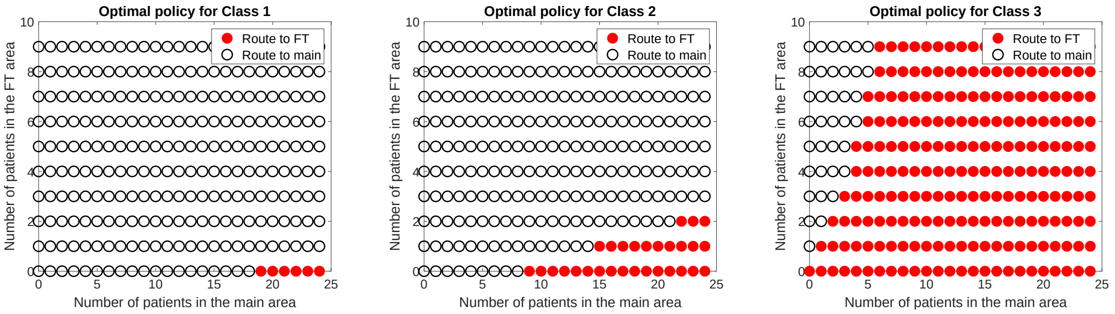


Figure 4 illustrates the optimal policy used in the simulation study. From Figure 4, we observe that Class 1 (i.e., high-complexity) patients should almost always be routed to the main area, whereas it is optimal to route most Class 3 (i.e., low-complexity) patients to the FT area under most circumstances. The dynamic routing mainly applies to Class 2 (i.e., medium-complexity) patients. Specifically, when the main area is crowded while the FT area is almost empty, it is optimal to route more patients of Class 2 to the FT area to reduce their waiting time, which also eases the congestion level in the main area.

Static Routing Policy (SP): Motivated by the structure of the optimal routing policy and the insights noted, we propose the following static routing policies, which are easier to implement because they are state-independent and do not require solving an MDP. Specifically, patient i is routed to the FT area if the predicted admission probability $\hat{M}_i$ is lower than the ηth percentile; otherwise, the patient is routed to the main area. Note that the admission prediction is based on the same approach as described in Section 5. Based on the numerical solution of the optimal policy, we choose the thresholds at the 25th and 30th percentiles and denote the corresponding static routing policies as SP-25 and SP-30, respectively.

Triage-Score-Based Routing Policy (TP): In the simulation study, we also consider the routing policy that routes (i) patients with triage scores 4 and 5 to the FT area, and (ii) patients with triage scores 1, 2, and 3 to the main ED area. Potentially due to its simplicity, such a purely triage-score-based routing policy has been implemented in many EDs under various triage protocols—for example, CTAS in Canada (Ding et al. 2019) and ESI in the US (Peck and Kim 2010)—despite the lack of understanding of its effectiveness.

6.4. Results and Discussion

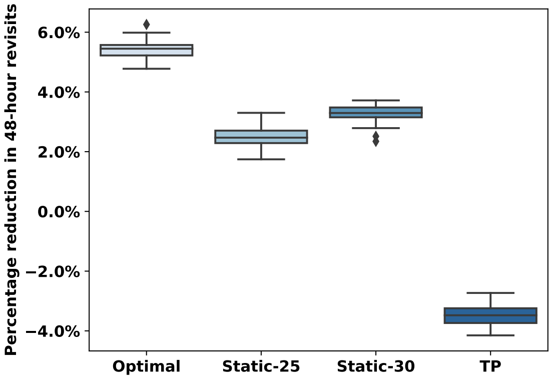
In the simulation, we use common random numbers for variance reduction when creating the patient arrival process under different routing policies. We run the simulation under each routing policy for 30 replications, where each replication has a length of 500 weeks. For each replication, we identify the first 100 weeks as the warm-up period by Welch's method (Law and Kelton 2000), and thus, the patient visit records during this period are removed from the output. We use the remaining data to calculate the 48-hour patient revisits and the average patient waiting time for each of the five policies described in Section 6.3. Table 5 shows the average waiting time and the 48-hour patient revisits and their corresponding 95% confidence intervals for the five routing policies. The 95% confidence intervals for the percentage reduction in the 48-hour patient revisits for policies OP, TP, SP-25, and SP-30 over policy CP are also included in Table 5 (see a graphical comparison in Figure 5), from which we make the following observations.

Table 5 The 95% confidence interval for the 48-hour revisits and the average waiting time under each routing policy, and the percentage reduction in the 48-hour revisits by using policies OP, TP, SP-25 and SP-30 over policy CP.

Routing policy	CP	OP	TP	SP-25	SP-30
The 48-hour patient revisits	5212 ± 6	4928 ± 5	5392 ± 5	5083 ± 4	5042 ± 5
Reduction in 48-hour patient revisits (%)		5.44 ± 0.12	−3.46 ± 0.14	2.47 ± 0.13	3.27 ± 0.12
Average waiting time (hours)					
All patients	1.51 ± 0.01	1.18 ± 0.01	1.51 ± 0.01	1.55 ± 0.01	1.41 ± 0.01
Patients in main area	1.49 ± 0.01	1.00 ± 0.01	1.42 ± 0.01	1.56 ± 0.01	1.10 ± 0.01
Patients in FT area	1.63 ± 0.01	1.89 ± 0.01	1.88 ± 0.01	1.53 ± 0.01	2.58 ± 0.01

Notes: The calculation of the 48-hour patient revisits is based on the total number of discharged patients during FT open hours for the two EDs in 24 months (i.e., a total of 123,655 observations).

Figure 5 Percentage reductions in the 48-hour patient revisits for the proposed routing policies over CP.



Observation 1. *The state-dependent policy OP performs the best among all the routing policies in terms of reducing both the 48-hour patient revisits and the average patient waiting time.*

Our simulation results show that 26.37% patients are routed to the FT area under policy OP, whereas the FT area treats 23.66% patients under policy CP. The percentage reduction in the 48-hour patient revisits by policy OP over the current routing policy used in our study EDs (Policy CP) is 5.44%. At the same time, policy OP reduces the average waiting times of all patients by 21.9%, compared to CP. A closer look finds that the waiting time reduction comes from the reduced waiting time of patients in the main area, but at the cost of longer waiting for patients treated in FT, as more patients are routed to FT by OP.

Observation 2. *Both policies SP-25 and SP-30 reduce the 48-hour patient revisits.*

The static routing policy SP-25 is interesting because, under this policy, almost the same percentage of patients are routed to FT as under policy CP. However, SP-25 can reduce the 48-hour patient revisits over CP by 2.47%, which implies that our patient classification can pick out the "right" patients to be routed to FT to reduce revisits and improve patient outcomes. Similarly, SP-30 also reduces the 48-hour patient revisits over CP with an even higher percentage reduction, i.e., 3.27%. Although SP-30 results in a shorter average waiting time for all patients compared to policy CP, the average waiting time for FT patients becomes significantly longer, mainly due to the higher workload in the FT area.

Observation 3. *The triage-score-based routing policy TP performs the worst among all policies under consideration, despite being the most popular policy implemented in many hospitals.*

The percentage reductions in the 48-hour patient revisits by OP and CP over TP are 8.6% and 3.3%, respectively. The performance of TP is not surprising, as it is the only policy that does not consider ED congestion levels among CP, OP, and TP. Policy TP is also outperformed by SP-25 and SP-30 since these two static routing policies can pick out the relatively "right" patients who are safer to be treated in the FT area.

To summarize, the state-dependent routing policy OP achieves the best performance in terms of reducing the 48-hour patient revisits and the average waiting time of all patients. The intuition is that the dynamic routing policy benefits from the server pooling effect, which, to a certain extent, makes up the "anti-pooling" deficit from setting up the FT area by placing physicians (also nurses and beds) into separate areas with dedicated queues. The current routing policy implemented in our study hospitals (CP) performs significantly better than the triage-score-based policy (TP); however, it is outperformed by our proposed policies OP, SP-25, and SP-30 because our patient classification helps identify the "right" patients to be routed to the FT area when the ED is congested. Despite being a popular policy in practice, TP is not recommended based on our simulation results. If management sees value in the simplicity of TP, then SP-25 can be a better alternative.

7. Conclusion and Future Research

This paper studies the impact of being routed to FT on patient outcomes using data from two Canadian EDs. The purpose of introducing an FT area is to reduce the waiting time for less urgent and less complex patients. However, the FT area forms a separate queue with a fixed allocation of medical resources, which may create the “anti-pooling” effect, as Saghafian et al. (2012) cautioned in their study. Triage nurses, the decision makers of FT routing, are aware of the congestion levels at both the main and the FT areas. Hence, it seems to be an intuitive and sensible decision to route patients who would be sent to the main area when the ED is less congested into the FT area when the main area is significantly more crowded, so as to reduce their waiting times. In fact, we find a positive correlation between the ED congestion level and the likelihood of being routed to the FT area. To a certain extent, routing decisions based on congestion levels achieve resource pooling between the main area and FT. Indeed, our results show that the congestion-dependent routing practice in our study EDs improves patient access to emergency care by reducing patient *LOS*, which aligns with triage nurses’ intuition.

However, through a subgroup analysis based on patient complexity classification, we uncover a hidden consequence of the congestion-influenced FT routing decisions: the 48- and 72-hour revisits increase respectively by 6.8% and 6.6% for high-complexity patients, and by 5.8% for medium-complexity patients. Therefore, we advise caution since it has unintended consequences on the quality of care, especially for patients with more complex care conditions. Being aware of this important trade-off between the care access and quality of care, we propose a multi-class queueing model to devise new routing policies and evaluate their performances through simulation studies. Our results show that a better-informed routing policy can improve both care access and quality of care compared to the current routing policy in our study hospitals. Interestingly, the triage-score-based policy, which routes all (and only) patients with triage scores 4 and 5 to the FT area, performs the worst among all the policies under consideration, despite its prevalent use as a guideline for making FT routing decisions in many hospitals. Our work, therefore, calls for attention from healthcare decision makers to carefully balance the trade-off between access to emergency care and the quality of care when making FT routing decisions.

As more hospitals have implemented FT areas in their EDs, it becomes increasingly important to establish consistent and evidence-based guidelines for FT routing decisions. Our study serves as an important step towards this goal. In what follows, we discuss some limitations of our study and point out opportunities for future research. First, our study focuses on two Canadian EDs where the physician scheduled to work in the FT area has similar training as physicians working in the main ED area. While we believe many EDs have similar settings to ours, we note that the staffing of FT areas in some hospitals can be different. For example, the ED studied by Sanchez et al. (2006) staffed physician assistants and nurse practitioners to provide care for patients routed to FT. Therefore, our results may not be directly applied to those hospitals, and it would be valuable to conduct analyses using data from more hospitals based on our framework. Second, we stratify patients into three complexity classes based on their predicted dispositions. It would be of interest for future studies to examine other classification methods that reflect patients’ heterogeneous care needs from alternative perspectives. For example, Ieraci et al. (2008) classify a patient as of low complexity if the patient's clinical requirements are evident and do not need intensive nursing care based on triage nurses' assessment. Finally, from a stochastic modeling perspective, it would be interesting to study the optimal routing policy theoretically based on our proposed multi-class queueing model, which adds to the growing body of work on patient admission and routing decisions in healthcare systems; see, e.g., Helm and Van Oyen (2014), Samiedaluie et al. (2017), Dai and Shi (2019), and Dong et al. (2019). We believe further investigations on these issues would be beneficial for the implementation of evidence-based guidelines for FT routing decisions in ED practice.

References

Affleck A, Parks P, Drummond A, Rowe BH, Ovens HJ (2013) Emergency department overcrowding and access block. *Canadian Journal of Emergency Medicine* 15(6):359–370.

Angrist JD, Pischke JS (2009) *Mostly harmless econometrics: An empiricist's companion* (Princeton university press).

Arya R, Wei G, McCoy JV, Crane J, Ohman-Strickland P, Eisenstein RM (2013) Decreasing length of stay in the emergency department with a split emergency severity index 3 patient flow model. *Academic Emergency Medicine* 20(11):1171–1179.

Batt RJ, Kc DS, Staats BR, Patterson BW (2019) The effects of discrete work shifts on a nonterminating service system. *Production and operations management* 28(6):1528–1544.

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

Burt CW, McCaig LF (2006) Staffing, capacity, and ambulance diversion in emergency departments, United States, 2003-04. *Adv Data* 376:1–23.

Chan CW, Green LV, Lekwijit S, Lu L, Escobar G (2018) Assessing the impact of service level when customer needs are uncertain: An empirical investigation of hospital step-down units. *Management Science* 65(2):751–775.

Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987) A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of chronic diseases* 40(5):373–383.

Chen J, Dong J, Shi P (2020a) A survey on skill-based routing with applications to service operations management. *Queueing Systems* 1–30.

Chen W, Linthicum B, Argon NT, Bohrmann T, Lopiano K, Mehrotra A, Travers D, Ziya S (2020b) The effects of emergency department crowding on triage and hospital admission decisions. *The American Journal of Emergency Medicine* 38(4):774–779.

Chrusciel J, Fontaine X, Devillard A, Cordonnier A, Kanagaratnam L, Laplanche D, Sanchez S (2019) Impact of the implementation of a fast-track on emergency department length of stay and quality of care indicators in the Champagne-Ardenne region: a before–after study. *BMJ Open* 9(6), ISSN 2044-6055.

Dai JG, Shi P (2019) Inpatient overflow: An approximate dynamic programming approach. *Manufacturing & Service Operations Management* 21(4):894–911.

Devkaran S, Parsons H, Van Dyke M, Drennan J, Rajah J (2009) The impact of a fast track area on quality and effectiveness outcomes: A Middle Eastern emergency department perspective. *BMC Emergency Medicine* 9(1):11.

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

Dong J, Shi P, Zheng F, Jin X (2019) Off-service placement in inpatient ward network: Resource pooling versus service slowdown. *Working paper* .

Freeman M, Savva N, Scholtes S (2017) Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science* 63(10):3147–3167.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5(2):79–141.

Gorski JK, Batt RJ, Otles E, Shah MN, Hamedani AG, Patterson BW (2017) The impact of emergency department census on the decision to admit. *Academic Emergency Medicine* 24(1):13–21.

Grafstein E, Unger B, Bullard M, Innes G, et al. (2003) Canadian emergency department information system (CEDIS) presenting complaint list (version 1.0). *Canadian Journal of Emergency Medicine* 5(1):27–34.

Grant KL, Bayley CJ, Premji Z, Lang E, Innes G (2020) Throughput interventions to reduce emergency department crowding: A systematic review. *Canadian Journal of Emergency Medicine* 22(6):864–874.

Greene WH (2018) *Econometric analysis* (Prentice Hall, Englewood Cliffs, NJ).

Guttmann A, Schull MJ, Vermeulen MJ, Stukel TA (2011) Association between waiting times and short term mortality and hospital admission after departure from emergency department: Population based cohort study from Ontario, Canada. *BMJ* 342.

Hajek B (1984) Optimal control of two interacting service stations. *IEEE transactions on automatic control* 29(6):491–499.

Helm JE, Van Oyen MP (2014) Design and optimization methods for elective hospital admissions. *Operations Research* 62(6):1265–1282.

Ieraci S, Digiusto E, Sonntag P, Dann L, Fox D (2008) Streaming by case complexity: Evaluation of a model for emergency department fast track. *Emergency Medicine Australasia* 20(3):241–249.

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

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

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

Kelly AM, Bryant M, Cox L, Jolley D (2007) Improving emergency department efficiency by patient streaming to outcomes-based teams. *Australian Health Review* 31(1):16–21.

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

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

Kim SH, Whitt W (2014) Choosing arrival process models for service systems: Tests of a nonhomogeneous poisson process. *Naval Research Logistics* 61(1):66–90.

Knapp LG, Seaks TG (1998) A hausman test for a dummy variable in probit. *Applied Economics Letters* 5(5):321–323.

Kuntz L, Mennicken R, Scholtes S (2015) Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science* 61(4):754–771.

Law AM, Kelton WD (2000) *Simulation modeling and analysis* (McGraw-Hill New York), 3rd edition.

Lippman SA (1975) Applying a new device in the optimization of exponential queuing systems. *Operations Research* 23(4):687–710.

Liu J, Xie J, Yang KK, Zheng Z (2019) Effects of rescheduling on patient no-show behavior in outpatient clinics. *Manufacturing & Service Operations Management* .

Liu SW, Hamedani AG, Brown DF, Asplin B, Camargo Jr CA (2013) Established and novel initiatives to reduce crowding in emergency departments. *Western Journal of Emergency Medicine* 14(2):85.

Long EF, Mathews KS (2018) The boarding patient: Effects of ICU and hospital occupancy surges on patient flow. *Production and Operations Management* 27(12):2122–2143.

Lu LX, Lu SF (2018) Distance, quality, or relationship? Interhospital transfer of heart attack patients. *Production and Operations Management* 27(12):2251–2269.

Maa J (2011) The waits that matter. *New England Journal of Medicine* 364(24):2279–2281.

Maddala GS (1986) *Limited-dependent and qualitative variables in econometrics*. Number 3 (Cambridge university press).

Michelson KA, Lyons TW, Bachur RG, Monuteaux MC, Finkelstein JA (2018) Timing and location of emergency department revisits. *Pediatrics* 141(5).

O'Brien D, Williams A, Blondell K, Jelinek GA (2006) Impact of streaming "fast track" emergency department patients. *Australian Health Review* 30(4):525–532.

Ouyang H, Liu R, Sun Z (2021) Emergency department modeling and staffing: Time-varying physician productivity. *Available at SSRN 3963226* .

Peck JS, Kim SG (2010) Improving patient flow through axiomatic design of hospital emergency departments. *CIRP Journal of Manufacturing Science and Technology* 2(4):255–260.

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

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* 60(5):1080–1097.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2014) Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management* 16(3):329–345.

Samiedaluie S, Kucukyazici B, Verter V, Zhang D (2017) Managing patient admissions in a neurology ward. *Operations Research* 65(3):635–656.

Sanchez M, Smally AJ, Grant RJ, Jacobs LM (2006) Effects of a fast-track area on emergency department performance. *The Journal of Emergency Medicine* 31(1):117–120.

Soltani M, Batt RJ, Bavafa H, Patterson B (2022) Does what happens in the ED stay in the ED? The effects of emergency department physician workload on post-ed care use. *Manufacturing & Service Operations Management* .

Song H, Tucker AL, Graue R, Moravick S, Yang JJ (2020) Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science* 66(9):3825–3842.

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

Stock J, Yogo M (2005) *Testing for Weak Instruments in Linear IV Regression*, 80–108 (New York: Cambridge University Press).

Trivedy CR, Cooke MW (2015) Unscheduled return visits (URV) in adults to the emergency department (ed): A rapid evidence assessment policy review. *Emergency Medicine Journal* 32(4):324–329.

Webb EM, Mills AF (2019) Incentive–compatible prehospital triage in emergency medical services. *Production and Operations Management* 28(9):2221–2241.

Weber RR (1978) On the optimal assignment of customers to parallel servers. *Journal of Applied Probability* 15(2):406–413.

Winston W (1977) Optimality of the shortest line discipline. *Journal of Applied Probability* 14(1):181–189.

Wooldridge JM (2012) *Introductory econometrics: A modern approach* (South-Western Cengage Learning).

Xu SH, Righter R, Shanthikumar JG (1992) Optimal dynamic assignment of customers to heterogeneous servers in parallel. *Operations Research* 40(6):1126–1138.

Xu SH, Zhao YQ (1996) Dynamic routing and jockeying controls in a two-station queueing system. *Advances in Applied Probability* 1201–1226.

Appendix: Tables

Table 6 Summary statistics for patients of different complexity classes

	High-complexity patients				Medium-complexity patients				Low-complexity patients			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Age (years)	54.20	18.76	0	106.5	39.98	17.73	0.0	104.2	33.88	15.54	0.0	100.3
Gender (Male %)	47.26	49.93	0	100	41.01	49.19	0	100	45.53	49.80	0	100
Triage score (%)												
CTAS 2	47.71	49.95	0	100	26.95	44.37	0	100	11.19	31.53	0	100
CTAS 3	45.30	49.78	0	100	53.15	49.90	0	100	34.24	47.45	0	100
CTAS 4	5.48	22.75	0	100	16.19	36.84	0	100	38.44	48.65	0	100
CTAS 5	1.51	12.21	0	100	3.71	18.90	0	100	16.12	36.78	0	100

Notes. SD = standard deviation; CTAS = Canadian Triage and Acuity Scale.

Table 7 Summary statistics for patient outcomes of different complexity classes

	High-complexity patients		Medium-complexity patients		Low-complexity patients	
	Main Area	Fast-Track	Main Area	Fast-Track	Main Area	Fast-Track
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
$Revisit_{48h}$ (%)	6.96 (25.44)	6.41 (24.50)	7.15 (25.77)	5.02 (21.85)	5.42 (22.64)	3.32 (17.92)
$Revisit_{72h}$ (%)	8.38 (27.72)	7.59 (26.49)	8.28 (27.56)	5.93 (23.61)	6.43 (24.52)	4.07 (19.76)
LOS (hours)	4.81 (3.08)	3.13 (2.06)	3.97 (2.68)	2.90 (1.83)	3.42 (2.35)	2.47 (1.58)

Notes. SD = standard deviation; *LOS* = length of stay.

Table 8 Full results on the correlation between operational status and FT routing decisions

	All patients	High-complexity	Medium-complexity	Low-complexity
MEBusyRatio	0.073***	0.073***	0.080***	0.074***
	(0.005)	(0.011)	(0.012)	(0.007)
Age group (Base=0–25 years)				
25–40 years	-0.080***	0.033	-0.020	-0.103***
	(0.015)	(0.076)	(0.055)	(0.018)
40–55 years	-0.102***	-0.064	-0.061	-0.072***
	(0.016)	(0.076)	(0.080)	(0.022)
55–70 years	-0.140***	-0.092	0.005	-0.053*
	(0.017)	(0.077)	(0.113)	(0.030)
$>$ 70 *years*	-0.247***	-0.124	-0.323**	-0.138**
	(0.020)	(0.080)	(0.160)	(0.055)
Triage score (Base=CTAS 2)				
CTAS 3	0.507***	0.466***	0.505***	0.433***
	(0.015)	(0.025)	(0.053)	(0.028)
CTAS 4	0.842***	0.759***	0.962***	0.718***
	(0.016)	(0.045)	(0.103)	(0.029)
CTAS 5	0.899***	1.073***	1.211***	0.726***
	(0.021)	(0.063)	(0.131)	(0.033)
Gender (Base=Female)				
Male	0.219***	0.109***	0.217***	0.298***
	(0.011)	(0.021)	(0.031)	(0.015)
Hospital (Base=ED A)				
ED B	-0.144***	-0.162***	-0.175***	-0.125***
	(0.011)	(0.021)	(0.025)	(0.015)
TriageTime	-0.184***	-0.156***	-0.149***	-0.209***
	(0.006)	(0.011)	(0.013)	(0.008)
N	123,655	50,514	23,377	49,434

Standard errors in parentheses. Some observations are dropped because of the perfect separation.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9 Full estimation results (with IV) based on visit records of all patients.

	48-hour revisit		72-hour revisit		Length of stay	
	FT	$Revisit_{48h}$	*FT*	$Revisit_{72h}$	*FT*	$\log(LOS)$
MEBusyRatio	0.073***		0.073***		0.072***	
	(0.013)		(0.013)		(0.016)	
Age group (Base=0–25 years)						
25–40 years	-0.080***	0.108***	-0.080***	0.120***	-0.080***	0.068***
	(0.015)	(0.020)	(0.015)	(0.020)	(0.015)	(0.006)
40–55 years	-0.102***	0.047*	-0.102***	0.064***	-0.102***	0.142***
	(0.020)	(0.024)	(0.020)	(0.023)	(0.020)	(0.007)
55–70 years	-0.140***	0.049*	-0.140***	0.074***	-0.140***	0.186***
	(0.024)	(0.026)	(0.024)	(0.025)	(0.024)	(0.007)
$>$ 70 *years*	-0.247***	0.118***	-0.247***	0.155***	-0.247***	0.269***
	(0.027)	(0.028)	(0.027)	(0.027)	(0.027)	(0.010)
Triage score (Base=CTAS 2)						
CTAS 3	0.507***	-0.065***	0.507***	-0.067***	0.508***	-0.078***
	(0.015)	(0.016)	(0.015)	(0.016)	(0.016)	(0.007)
CTAS 4	0.842***	-0.191***	0.842***	-0.188***	0.843***	-0.168***
	(0.020)	(0.019)	(0.020)	(0.018)	(0.020)	(0.012)
CTAS 5	0.899***	-0.252***	0.899***	-0.246***	0.900***	-0.212***
	(0.020)	(0.028)	(0.020)	(0.027)	(0.020)	(0.018)
Gender (Base=Female)						
Male	0.219***	-0.021	0.219***	-0.005	0.219***	-0.025***
	(0.012)	(0.013)	(0.012)	(0.013)	(0.012)	(0.005)
Hospital (Base=ED A)						
ED B	-0.144	0.115***	-0.144	0.104***	-0.144	-0.085***
	(0.135)	(0.019)	(0.135)	(0.017)	(0.134)	(0.026)
TriageTime	-0.184***	0.005	-0.184***	0.008	-0.184***	0.038***
	(0.009)	(0.007)	(0.009)	(0.006)	(0.009)	(0.003)
Workload		0.012*		0.012*		0.025***
		(0.007)		(0.006)		(0.008)
AvgOccTreated		0.032***		0.039***		0.341***
		(0.008)		(0.008)		(0.008)
WaitTime		0.056***		0.048***		
		(0.007)		(0.006)		
FT		-0.028		-0.047		-0.247***
		(0.068)		(0.065)		(0.055)
N	123,655		123,655		123,655	

Notes. Standard errors in parentheses. CTAS = Canadian Triage and Acuity Scale.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10 Full estimation results (with IV) on patient outcome variables for high-complexity patients.

	48-hour revisit		72-hour revisit		Length of stay	
	FT	$Revisit_{48h}$	*FT*	$Revisit_{72h}$	*FT*	$\log(LOS)$
MEBusyRatio	0.074***		0.073***		0.072***	
	(0.013)		(0.013)		(0.013)	
Age group (Base=0–25 years)						
25–40 years	0.037	0.026	0.035	0.066*	0.032	0.097***
	(0.077)	(0.039)	(0.077)	(0.035)	(0.077)	(0.018)
40–55 years	-0.061	-0.109***	-0.063	-0.067*	-0.066	0.185***
	(0.074)	(0.042)	(0.074)	(0.039)	(0.075)	(0.019)
55–70 years	-0.087	-0.112**	-0.089	-0.063	-0.094	0.240***
	(0.072)	(0.043)	(0.072)	(0.039)	(0.073)	(0.020)
> 70 *years*	-0.117	-0.036	-0.119	0.018	-0.126	0.340***
	(0.079)	(0.046)	(0.079)	(0.041)	(0.079)	(0.022)
Triage score (Base=CTAS 2)						
CTAS 3	0.467***	-0.097***	0.468***	-0.105***	0.470***	-0.099***
	(0.026)	(0.020)	(0.026)	(0.019)	(0.027)	(0.008)
CTAS 4	0.761***	-0.216***	0.760***	-0.192***	0.763***	-0.259***
	(0.046)	(0.045)	(0.046)	(0.038)	(0.047)	(0.016)
CTAS 5	1.073***	-0.259***	1.073***	-0.263***	1.076***	-0.379***
	(0.064)	(0.078)	(0.064)	(0.075)	(0.064)	(0.032)
Gender (Base=Female)						
Male	0.110***	-0.008	0.110***	0.020	0.109***	-0.020***
	(0.024)	(0.017)	(0.024)	(0.017)	(0.024)	(0.006)
Hospital (Base=ED A)						
ED B	-0.164	0.118***	-0.163	0.108***	-0.163	-0.100***
	(0.127)	(0.025)	(0.128)	(0.022)	(0.128)	(0.028)
TriageTime	-0.155***	0.006	-0.155***	0.009	-0.155***	0.041***
	(0.013)	(0.010)	(0.013)	(0.009)	(0.013)	(0.003)
Workload		0.022**		0.028***		0.010
		(0.010)		(0.009)		(0.008)
AvgOccTreated		-0.016		-0.007		0.338***
		(0.014)		(0.013)		(0.010)
WaitTime		0.044***		0.037***		
		(0.008)		(0.008)		
FT		0.415***		0.363***		-0.303***
		(0.134)		(0.132)		(0.055)
N	50,768		50,768		50,768	

Notes. Standard errors in parentheses. CTAS = Canadian Triage and Acuity Scale.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 11 Full estimation results (with IV) on patient outcome variables for medium-complexity patients.

	48-hour revisit		72-hour revisit		Length of stay	
	FT	$Revisit_{48h}$	*FT*	$Revisit_{72h}$	*FT*	$\log(LOS)$
MEBusyRatio	0.081^{***}		0.082^{***}		0.089^{***}	
	(0.016)		(0.016)		(0.017)	
Age group (Base=0–25 years)						
25–40 years	-0.018	-0.023	-0.016	0.007	-0.019	0.046^{**}
	(0.056)	(0.063)	(0.056)	(0.063)	(0.056)	(0.023)
40–55 years	-0.054	-0.095	-0.051	-0.064	-0.060	0.092^{**}
	(0.082)	(0.113)	(0.082)	(0.115)	(0.082)	(0.037)
55–70 years	0.013	-0.236	0.017	-0.178	0.005	0.166^{***}
	(0.117)	(0.157)	(0.117)	(0.159)	(0.118)	(0.052)
> 70 *years*	-0.310^{*}	-0.297	-0.305^{*}	-0.165	-0.323^{*}	0.237^{***}
	(0.168)	(0.217)	(0.168)	(0.226)	(0.169)	(0.076)
Triage score (Base=CTAS 2)						
CTAS 3	0.504^{***}	0.014	0.504^{***}	-0.009	0.503^{***}	-0.024
	(0.055)	(0.065)	(0.055)	(0.064)	(0.055)	(0.023)
CTAS 4	0.957^{***}	-0.007	0.954^{***}	-0.040	0.959^{***}	-0.096^{**}
	(0.116)	(0.140)	(0.115)	(0.137)	(0.116)	(0.048)
CTAS 5	1.204^{***}	0.006	1.201^{***}	-0.047	1.216^{***}	-0.130^{**}
	(0.142)	(0.179)	(0.141)	(0.181)	(0.141)	(0.062)
Gender (Base=Female)						
Male	0.220^{***}	-0.143^{***}	0.220^{***}	-0.128^{***}	0.211^{***}	-0.044^{***}
	(0.034)	(0.036)	(0.034)	(0.035)	(0.034)	(0.013)
Hospital (Base=ED A)						
ED B	-0.177	0.194^{***}	-0.178	0.174^{***}	-0.175	-0.088^{***}
	(0.137)	(0.032)	(0.137)	(0.031)	(0.137)	(0.026)
TriageTime	-0.148^{***}	-0.013	-0.148^{***}	-0.004	-0.152^{***}	0.029^{***}
	(0.015)	(0.014)	(0.015)	(0.013)	(0.015)	(0.006)
Workload		0.029^{**}		0.020		0.029^{***}
		(0.015)		(0.014)		(0.010)
AvgOccTreated		0.012		0.032^{**}		0.368^{***}
		(0.014)		(0.013)		(0.009)
WaitTime		0.085^{***}		0.074^{***}		
		(0.013)		(0.011)		
FT		0.399^{***}		0.361^{***}		-0.457^{***}
		(0.129)		(0.133)		(0.061)
N	23,453		23,453		23,453	

Notes. Standard errors in parentheses. CTAS = Canadian Triage and Acuity Scale.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 12 Full estimation results (with IV) on patient outcome variables for low-complexity patients.

	48-hour revisit		72-hour revisit		Length of stay	
	FT	$Revisit_{48h}$	*FT*	$Revisit_{72h}$	*FT*	log(*LOS*)
MEBusyRatio	0.074***		0.074***		0.095***	
	(0.016)		(0.016)		(0.022)	
Age group (Base=0–25 years)						
25–40 years	-0.103***	0.169***	-0.103***	0.163***	-0.104***	0.055***
	(0.019)	(0.030)	(0.019)	(0.031)	(0.019)	(0.008)
40–55 years	-0.072***	0.199***	-0.072***	0.218***	-0.072***	0.117***
	(0.025)	(0.035)	(0.025)	(0.035)	(0.025)	(0.009)
55–70 years	-0.052*	0.267***	-0.052*	0.286***	-0.055*	0.139***
	(0.032)	(0.044)	(0.032)	(0.042)	(0.032)	(0.012)
$>$ 70 *years*	-0.138**	0.374***	-0.138**	0.426***	-0.143**	0.134***
	(0.063)	(0.072)	(0.063)	(0.069)	(0.062)	(0.024)
Triage score (Base=CTAS 2)						
CTAS 3	0.433***	-0.046	0.433***	-0.036	0.434***	0.001
	(0.027)	(0.045)	(0.027)	(0.039)	(0.027)	(0.013)
CTAS 4	0.718***	-0.202***	0.718***	-0.200***	0.719***	-0.050***
	(0.032)	(0.047)	(0.032)	(0.042)	(0.032)	(0.017)
CTAS 5	0.726***	-0.256***	0.726***	-0.245***	0.730***	-0.072***
	(0.031)	(0.054)	(0.031)	(0.049)	(0.031)	(0.019)
Gender (Base=Female)						
Male	0.298***	0.021	0.298***	0.020	0.297***	-0.016**
	(0.015)	(0.023)	(0.015)	(0.024)	(0.015)	(0.008)
Hospital (Base=ED A)						
ED B	-0.125	0.078***	-0.125	0.070***	-0.126	-0.072**
	(0.140)	(0.027)	(0.140)	(0.025)	(0.139)	(0.030)
TriageTime	-0.209***	0.035***	-0.209***	0.032***	-0.208***	0.026***
	(0.015)	(0.013)	(0.015)	(0.011)	(0.015)	(0.005)
Workload		-0.022**		-0.021**		0.041***
		(0.009)		(0.009)		(0.011)
AvgOccTreated		0.092***		0.088***		0.332***
		(0.009)		(0.009)		(0.010)
WaitTime		0.047***		0.043***		
		(0.012)		(0.011)		
FT		-0.028		-0.044		-0.421***
		(0.107)		(0.096)		(0.072)
N	49,434		49,434		49,434	

Notes. Standard errors in parentheses. CTAS = Canadian Triage and Acuity Scale.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 13 Average marginal effect of FT routing on patient outcomes for the eleven robustness checks.

	All patients			High-complexity			Medium-complexity			Low-complexity		
Panel	Re_{48h}	Re_{72h}	$\log(LOS)$	Re_{48h}	Re_{72h}	$\log(LOS)$	Re_{48h}	Re_{72h}	$\log(LOS)$	Re_{48h}	Re_{72h}	$\log(LOS)$
(1)	-0.00	-0.01	-0.34***	0.07**	0.06**	-0.39***	0.06**	0.06**	-0.54***	-0.00	-0.01	-0.57***
	(0.01)	(0.01)	(0.08)	(0.03)	(0.03)	(0.07)	(0.02)	(0.03)	(0.07)	(0.01)	(0.01)	(0.07)
(2)	-0.00	-0.01	-0.23***	0.07**	0.06**	-0.31***	0.06**	0.05*	-0.45***	-0.00	-0.01	-0.38***
	(0.01)	(0.01)	(0.05)	(0.03)	(0.03)	(0.06)	(0.02)	(0.03)	(0.06)	(0.01)	(0.01)	(0.07)
(3)	-0.01	-0.01	-0.20***	0.07**	0.07**	-0.30***	0.05**	0.05**	-0.43***	-0.00	-0.01	-0.33***
	(0.01)	(0.01)	(0.05)	(0.03)	(0.03)	(0.06)	(0.02)	(0.03)	(0.06)	(0.01)	(0.01)	(0.06)
(4)	-0.01	-0.01	-0.17***	0.07**	0.06**	-0.29***	0.05**	0.05**	-0.40***	-0.00	-0.01	-0.25***
	(0.01)	(0.01)	(0.04)	(0.03)	(0.03)	(0.06)	(0.02)	(0.03)	(0.06)	(0.01)	(0.01)	(0.05)
(5)	-0.00	-0.00	-0.25***	0.07**	0.07**	-0.30***	0.04**	0.04**	-0.44***	0.00	0.00	-0.44***
	(0.01)	(0.01)	(0.06)	(0.03)	(0.03)	(0.06)	(0.02)	(0.02)	(0.05)	(0.01)	(0.01)	(0.09)
(6)	-0.00	-0.00	-0.25***	0.07**	0.07**	-0.30***	0.09***	0.09**	-0.47***	-0.01	-0.01	-0.44***
	(0.01)	(0.01)	(0.06)	(0.03)	(0.03)	(0.06)	(0.04)	(0.04)	(0.08)	(0.01)	(0.01)	(0.07)
(7)	-0.00	-0.00	-0.25***	0.07**	0.07**	-0.29***	0.05**	0.05*	-0.49***	-0.00	-0.00	-0.42***
	(0.01)	(0.01)	(0.06)	(0.03)	(0.03)	(0.05)	(0.03)	(0.03)	(0.08)	(0.01)	(0.01)	(0.07)
(8)	-0.00	-0.00	-0.25***	0.08**	0.07**	-0.32***	0.04**	0.04**	-0.45***	-0.00	-0.00	-0.42***
	(0.01)	(0.01)	(0.06)	(0.03)	(0.03)	(0.06)	(0.02)	(0.02)	(0.06)	(0.01)	(0.01)	(0.07)
(9)	-0.00	-0.00	-0.24***	0.07**	0.07**	-0.29***	0.06***	0.06**	-0.44***	-0.00	-0.00	-0.43***
	(0.01)	(0.01)	(0.05)	(0.03)	(0.03)	(0.05)	(0.02)	(0.03)	(0.06)	(0.01)	(0.01)	(0.07)
(10)	-0.00	-0.01	-0.25***	0.07**	0.07**	-0.31***	0.06**	0.06**	-0.45***	-0.00	-0.00	-0.43***
	(0.01)	(0.01)	(0.06)	(0.03)	(0.03)	(0.06)	(0.02)	(0.03)	(0.06)	(0.01)	(0.01)	(0.07)
(11)	-0.00	-0.00	-0.20***	0.06*	0.06	-0.30***	0.07**	0.10**	-0.34***	-0.01	-0.01	-0.19
	(0.01)	(0.01)	(0.08)	(0.04)	(0.04)	(0.06)	(0.04)	(0.04)	(0.09)	(0.02)	(0.02)	(0.16)

Notes. Panel (1): alternative IV without adjusting the number of physicians on duty; Panels (2), (3), (4): alternative IV using information that is 0.5, 1, and 2 hours before the triage start; Panels (5)–(8): alternative classification cutoffs: setting (t_1, t_2) to be the (35, 60)th, (45, 60)th, (40, 55)th, and (40, 65)th percentiles, respectively; Panel (9): alternative sample removing observations with triage time longer than 17 mins; Panel (10): alternative sample including patients of triage level 1; Panel (11): alternative model specification with controls on comorbidity. Re_{48h} stands for $Revisit_{48h}$ and Re_{72h} stands for $Revisit_{72h}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 14 Robustness check with 7-day revisits.

All patients	High-complexity	Medium-complexity	Low-complexity
-0.01	0.05*	0.04	-0.01
(0.01)	(0.03)	(0.03)	(0.01)

Notes. Standard errors clustered by the name of the physician who performed the initial assessment are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$