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

Many hospitals have established a separate fast-track (FT) service line dedicated to patients with less urgent care needs to improve emergency department (ED) operational efficiency. However, so far, hospitals have not yet established consistent guidelines for determining which patients should be routed to the FT, possibly due to the lack of a comprehensive understanding of how FT routing decisions impact patient outcomes. Using data from three Canadian hospitals, we find that FT routing decisions are not purely clinical-driven, and ED operational status related to congestion is also associated with FT routing decisions. We then use an instrumental variable approach to quantify the impact of the FT routing decisions on patient outcomes (i.e, ED length of stay and revisit rate). We find that being routed into FT can improve patient access to emergency care by reducing the average ED length of stay by 25.9%. However, surprisingly, we uncover a hidden unintended consequence on the quality of care: being routed into the FT can lead to a 7.4% increase in the 48-hour revisit rate for high-complexity patients and a 3.0% (2.8%) increase in the 48-hour (72-hour) revisit rate for middle-complexity patients. To balance the trade-off between fast care access and quality assurance, we propose a multi-class queueing model to derive the optimal routing policy. Using a data-calibrated simulation, we compare the performance of several routing policies and find that the optimal state-dependent routing policy can lead to a 4.72% reduction in the 48-hour patient revisits and a 22.2% reduction in the average patient waiting time, compared to the current routing policy used by our study hospitals. Interestingly, the popularly adopted triage-score-based policy, which routes all (and only) patients of triage levels 4 and 5 to the FT, performs the worst among all policies under consideration.

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 which, 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 towards fast care delivery and rapid discharge for patients with less urgent conditions. FT has become more prevalent in recent years, implemented in nearly 80% of academic EDs in the US (Liu et al. 2013).

Upon arriving at an ED, a patient who does not have life-threatening conditions is first triaged by the nursing staff, who (i) assign the patient a triage score that indicates the urgency level of the patient's care needs and (ii) route the patient either into the main ED area, where most patients are treated, or into 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 into the FT. Hence, EDs currently make FT routing decisions at their own discretion. Some EDs adopt flexible routing protocols, 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. 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 FT. Such a policy is simple and easy to implement, but it is rigid. As a result, triage nurses may deviate from established protocols to gain some flexibility. Evidence from the emergency medicine community (Bjørn and Rødje 2008) shows that under this strict procedure, when the main ED area is crowded, triage nurses may intentionally assign patients higher triage scores (i.e., lower acuity levels) than under less crowded situations, so as to "legally" route patients to the FT area to reduce the overall ED congestion (a phenomenon usually referred to as *triage drift*). Therefore, it is inherently important to establish flexible and evidence-based policies to guide FT routing decisions (Peck and Kim 2010).

To devise new routing policies, it is crucial to understand (i) which factors are associated with triage nurses' routing decisions and (ii) how the FT routing decisions affect patient outcomes. Intuitively, patient characteristics (e.g., age and gender) and medical conditions (e.g., triage score and chief complaints) are key factors associated with FT routing decisions, as the FT area is dedicated to patients with less urgent and less complex medical conditions. However, since the FT area is a rigidly separated area, the mismatch of demand and capacity in both the FT and main areas can occur if the routing decisions are purely based on clinical factors, which leads to operational inefficiency. Particularly in a congested system, the workloads 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 go to the main or FT area, may consider the ED operational conditions in the FT routing decisions to better match healthcare resources with demands. Yet, the impact of such routing decisions on patient outcomes is unclear.

Indeed, recent studies on healthcare operations management have suggested that healthcare practitioners' decision-making processes are not purely clinically driven. Other factors can also affect these decisions, such as the system congestion level, physician workload, and even facility layout; see, e.g., Kim et al. (2015), Freeman et al. (2017), Ding et al. (2019), and Meng et al. (2021). These papers investigate various decision-making processes in healthcare settings, such as patient admission and prioritization in EDs and ICUs. However, to the best of our knowledge, the decision as to whether a patient should be treated in the main ED area or the FT has not yet been studied. The objectives of this study are, therefore, threefold: (i) to investigate the relationship between ED congestion and triage nurses' FT routing decisions, (ii) to evaluate the impact of the current FT routing decisions in practice on patient outcomes, and (iii) to devise new routing policies to improve ED operational performance and patient quality of care.

To achieve our research goals, we obtain two-year patient electronic health record data from three hospital EDs in Alberta, Canada. Our dataset is unique in that the FT routing decisions in our study EDs depend not only on triage scores but also on patient and ED characteristics, enabling the investigation of our research questions. Our findings and contributions can be summarized as follows.

First, we find a positive correlation between the ED congestion measure and FT routing decisions (i.e., the likelihood of being routed into 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, which is consistent with the prior findings on other healthcare decision-making processes (see, e.g., Kim et al. 2015, Freeman et al. 2017, Ding et al. 2019, and Meng et al. 2021).

Next, we seek to understand the impact of FT routing decisions on patient outcomes. One challenge in addressing this research question is that the FT routing decision might be affected by factors that are unobservable in our data. To overcome this challenge, we utilize an instrumental variable (IV) approach with IVs related to ED operational status to quantify the causal impact of FT routing decisions on patient outcomes. Our estimation finds that being routed into the FT area can reduce a patient's LOS by 25.9% on average, equivalent to 0.91 hospital hours. This finding supports the motive of establishing the FT line, that is, to provide fast care delivery and improve emergency care access.

Third, to understand which patients can be safely treated in the FT area, we aim to categorize patients into different groups and investigate the impact of the FT routing decisions for each patient group separately. To this end, one empirical challenge is how to classify patients into different groups to reflect their heterogeneous needs in diagnosis and emergency care. Inspired by the ED practice (O'Brien et al. 2006, Kelly et al. 2007) and prior operations management literature (Saghafian et al. 2012, 2014), we classify patients based on their likelihood of admission. Based on our classification, we find that, surprisingly, being routed into the FT can lead to a 7.4% increase in the 48-hour revisit rate for high-complexity patients and a 3.0% (2.8%) increase in the 48-hour (72-hour) revisit rate for middle-complexity patients. We then explore the underlying mechanism by examining patients' diagnosis and treatment procedures in the main and FT areas. We remark

that our paper is the first to document the trade-off between fast care access and the quality of care driven by the FT routing decisions in EDs.

Finally, 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 find that the optimal state-dependent routing policy reduces the 48-hour patient revisits by 4.72% and the average patient waiting time by 22.2% over the current policy used in 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 close to 2%. However, interestingly, the popularly adopted triage-score-based policy that simply routes all patients of triage levels 4 and 5 into the FT area has the worst performance, potentially due to its inflexibility.

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 this paper, we focus on the patient routing decisions at triage in hospital EDs. Hence, our work is related to existing literature from both the emergency medicine and the operations management communities. In particular, our work is relevant to three streams of research within the operations literature: (i) the impact of routing decisions in healthcare settings; (ii) the behavior of healthcare decision makers; and (iii) the skill-based routing in service systems.

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) and Chrusciel et al. (2019). These medical papers conclude that the implementation of an FT line has a negative correlation with the average patient waiting time, LOS, and the rate at which patients left without being seen. However, there is a lack of consensus about the impact of FT on the quality of care. Specifically, the introduction of the FT line does not impact the mortality or the 72-hour revisit rate, concluded by Devkaran et al. (2009) and Sanchez et al. (2006). In contrast, Ieraci et al. (2008) find a statistically significant increase in the 48-hour revisit rate for non-admitted patients and Chrusciel et al. (2019) observe a rise in the 30-day readmission rate after the implementation of the FT. Our work is relevant in that we also study the relationship between FT routing decisions and patient outcomes. However, these prior medical studies focus more on correlation than causality, whereas our work considers potential endogeneity issues and establishes the causal link via an IV approach. Moreover, we investigate the heterogeneous effects for patients of different complexity levels and propose new routing policies based on our empirical findings, which further differentiate our work from the emergency medicine literature.

In the operations management literature, our work is most relevant to studies that empirically examine the impact of routing decisions in healthcare settings, including Kim et al. (2015), Chan et al. (2018), and Song et al. (2020). 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. The authors use hospital operational factors as IVs to handle potential endogeneity issues. 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 constraint on the unit designed for this patient's service needs. The authors examine the impact of off-service placement on patient outcomes so that hospital managers can make better capacity allocation-related decisions. The 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. 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 prior related work, this paper focuses on triage nurses' FT routing decisions and quantifies their impact on emergency care access and quality assurance, which adds to the existing literature.

In recent years, studies on healthcare workers' behavioral issues, especially in congested systems, have attracted increasing attention from the operations management community (KC et al. 2020). Existing evidence has 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. 2020), and patient undercoding (Powell et al. 2012). Our study finds a positive association between ED congestion and FT routing decisions made by triage nurses and thus is relevant to this stream of literature. On the other hand, our study focuses on quantifying the impact of the FT routing decisions on patient outcomes and devising new routing policies, which differs from the literature on workload-induced behaviors.

Finally, 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 which 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).

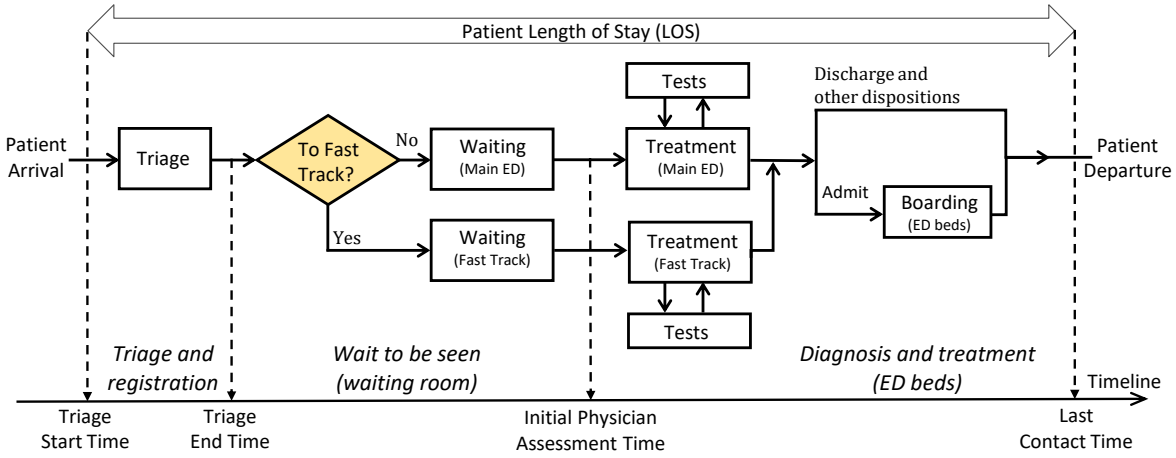
3. Study Setting and Data

In this section, we describe the ED patient flow process and the data used to support our analysis. Section 3.1 describes the setting of our empirical analysis, Section 3.2 presents the details of our data, and Section 3.3 discusses the choice of variables in the estimation.

3.1. Patient Flow

We first describe the patient flow process in our collaborator hospitals. The three 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 triaged into five levels, following the CTAS protocol. The timestamps at the start and end of the triage process are referred to as triage start time and triage end time, respectively. The time duration between triage start and end is referred to as the *triage time*. After assigning triage scores, triage nurses route patients into either the main ED or the FT area, which are two separate treatment areas with separate medical facilities and dedicated care teams. They share the same pool of attending physicians and have similar configurations except that the FT area contains fewer beds and physicians. During the study period, the FT line operates 10 hours every day in the three EDs. The average daily arrivals to the three EDs are 178.9, 194.1, and 183.7; the average daily traffic to the FT areas of the three EDs are 33.4, 34.1, and 41.7.

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 ED *length of stay (LOS)*.

3.2. Data Description and Cleaning

Our data contain patient visit records from the EDs of three urban hospitals in Alberta, Canada, from August 2013 to July 2015, involving a total of 406,219 visits from 246,420 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, gender, chief complaint, and triage score) and the details of their ED visits.

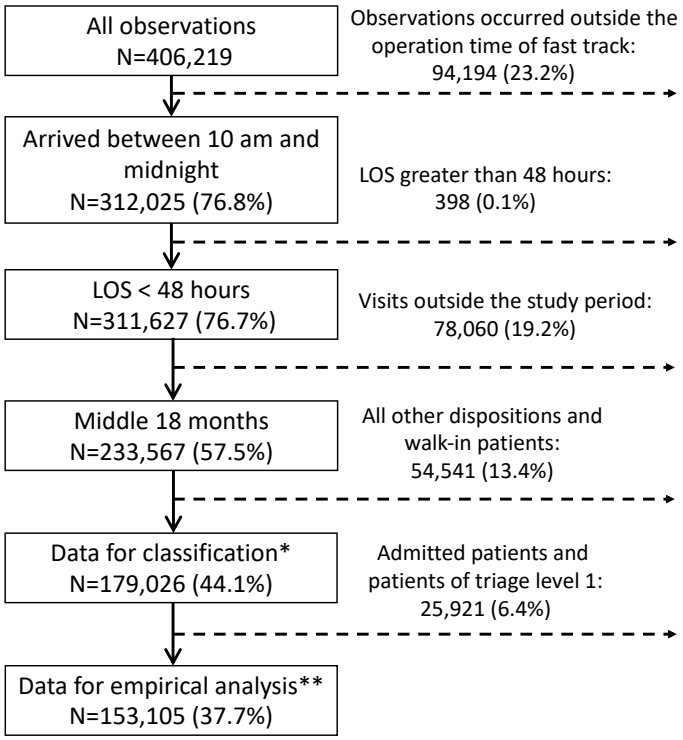
To proceed with our empirical analysis, we start the data selection process by first excluding patient visits that had occurred outside the FT operation time, which leaves us with a total of 312,025 observations. In addition, we remove observations with an LOS greater than 48 hours, as those extreme cases could bias our results (Song et al. 2015). Note that only 398 observations are identified with LOS greater than 48 hours in our data. Next, we restrict our sample to visits during the 18 months in the middle of our study period (from November 1, 2013 to April 1, 2015) to avoid censored estimates (Kim et al. 2015, Song et al. 2020, Chan et al. 2018). In this step, we remove 78,060 observations (19.2% of the original data). We further restrict our data to (i) walk-in patients only because the process of triage and treatment is more standardized for this type of patients (Batt and Terwiesch 2016); (ii) patients with dispositions of admission and discharge only (which removes dispositions such as left without being seen, left against medical advice, transferred, and so on). These two steps leave us with 179,026 observations, which are used for the patient classification in Section 4.4 and the simulation study in Section 6. For our empirical analysis, we further remove admitted patients and patients of triage level 1. We remove admitted patients because they usually stay longer than 48 hours in the hospital, so one of the key patient outcome measures in our study, i.e., the 48-hour revisit rate, is *not* a proper measure for these patients. Besides, we exclude patients of triage level 1, as their conditions are usually very urgent, requiring immediate attention (Ding et al. 2019). These two steps leave us with 153,105 observations used for our empirical analysis. Figure 2 depicts the data selection process.

3.3. Choice of Variables

This section presents the choice of variables used in our empirical analysis; see Table 1 for the summary statistics for the variables of interest. We first discuss the measures of patient outcomes.

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 $Revisit_{48h}$ was assigned the value 1 if the patient visited one of the three EDs within 48 hours after discharge and 0 otherwise. Similarly, $Revisit_{72h}$ equals 1 if the patient revisited one of the three ED in 72 hours after discharge and 0 otherwise. The 48- and 72-hour revisit rates are widely used in the healthcare literature to

Figure 2 A depiction of the data selection process.



Note. * represents the dataset used for patient classification in Section 4.4 and for the simulation input analysis in Section 6; ** represents the dataset for our empirical analysis.

measure the quality of emergency care (Ieraci et al. 2008, Trivedy and Cooke 2015, Song et al. 2015). The variable *LOS* is the time duration from a patient's arrival to the ED till his departure after the completion of ED treatment, which reflects how quickly a patient can obtain access to emergency care.

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 into the FT area and 0 otherwise. In the following, we discuss a set of variables for the operational characteristics and a set for the patient characteristics.

The key operational characteristics of interest are measures of the system congestion levels in the main ED area, the FT area, and the entire ED, denoted by *MainCongestion*, *FTCongestion*, and *EDCongestion*, respectively. They are calculated as the ratio of the physician workload to the physician capacity in the respective treatment area. The physician workload is calculated as the number of patients waiting to be seen and currently being treated in this area divided by the number of physicians on duty in the same area. The physician capacity is defined as the 95th percentile of the distribution of the physician workload in the treatment area. We use the 95th percentile instead of the maximum to avoid observations under extreme situations (Kim et al. 2015, Li et al. 2021). Hence, the congestion level by our definition measures the extent to which a physician's current workload takes up her service capacity. The average values for

Table 1 Summary statistics of key variables.

	Main area				Fast-track			
Variables	Mean	SD	Min	Max	Mean	SD	Min	Max
Patient Outcomes								
$Revisit_{48h}$ (%)	6.34	24.37	0	1	3.89	19.33	0	1
$Revisit_{72h}$ (%)	7.65	26.57	0	1	4.81	21.40	0	1
LOS (in hours)	4.18	2.81	0.00	35.27	2.74	1.76	0.04	23.62
Operational Characteristics								
EDCongestion	0.78	0.15	0.13	1.34	0.79	0.15	0.14	1.33
MainCongestion	0.78	0.15	0.13	1.34	0.79	0.15	0.15	1.34
FTCongestion	0.60	0.26	0	1.82	0.59	0.26	0	1.82
AvgWeightedTreatOccAll	0.54	0.22	0	1.30	0.47	0.23	0	1.25
HospWaitTime (in hours)	1.68	1.33	0	17.64	1.49	1.10	0	9.97
Triage time (in minutes)	4.89	2.14	0.45	44.87	4.16	1.85	0.63	46.22
Patient Characteristics								
Gender (Male = 1)	0.42	0.49	0	1	0.56	0.50	0	1
Age group (%)								
0 to 25 years	20.68	40.50	0	1	23.25	42.25	0	1
25 to 40 years	31.56	46.48	0	1	29.36	45.54	0	1
40 to 55 years	22.18	41.55	0	1	22.58	41.81	0	1
55 to 70 years	15.80	36.48	0	1	16.45	37.07	0	1
Over 70 years	9.78	29.70	0	1	8.36	27. 67	0	1
Triage level (%)								
CTAS 2	0.34	0.47	0	1	0.16	0.37	0	1
CTAS 3	0.44	0.50	0	1	0.38	0.49	0	1
CTAS 4	0.17	0.37	0	1	0.33	0.47	0	1
CTAS 5	0.06	0.24	0	1	0.13	0.34	0	1
Instrumental Variable								
MEAdjBusyRatio	1.13	0.07	0.72	1.38	1.13	0.07	0.71	1.38
Observations	111,053				42,052			

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

the *MainCongestion*, *FTCongestion*, and *EDCongestion* are 0.578, 0.283, and 0.577, respectively. In the robustness check in Section 5.4, we consider an alternative measure without the adjustment of the number of physicians on duty. Thus, instead of the physician capacity and physician workload, we measure the area capacity and area workload. Specifically, the alternative congestion levels are the ratios of the area workload to the area capacity in the corresponding treatment areas. In addition to the system-level operational metrics, we include two patient-level operational characteristics: the patient waiting time and triage time.

We include the following patient characteristics: age, gender, triage level, and chief complaints. To account for the possible nonlinear effect of age, we use categorized age groups instead of numerical values. We use the triage level to control for the urgency level of a patient. We also control the heterogeneity in patient health conditions within the same triage level using the 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 number of levels, especially for complaints with few observations, we follow the chief complaint classification protocol in Grafstein et al. (2003) to group the 170 complaints into 18 major categories.

4. Econometric Model

In this section, we describe the econometric model and identification strategy used in this paper. Section 4.1 presents our baseline econometric models. Due to the empirical challenge caused by endogeneity issues, Section 4.2 outlines our identification strategy with instrumental variables (IVs). Section 4.3 introduces two econometric models (with IVs) for various patient outcome variables. Finally, Section 4.4 discusses the patient classification method for our subgroup analysis.

4.1. Baseline Econometric Models

In this section, we consider three baseline econometric models: a probit model to examine factors associated with the FT routing decisions, another probit model to measure the impact of FT routing on the 48- and 72-hour revisit rates, and a log-linear regression model to measure the impact of the FT routing decision on patient *LOS*. To start with, we consider the following probit model to investigate which factors are associated with the FT routing decisions:

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

Next, we want to understand the impact of FT routing decisions on patient outcomes. The best way to quantify the impact on *Revisit* (representing both the 48- and 72-hour revisit rates) and *LOS* is through field experiments by randomly assigning patients into either the main or FT area. However, this method is impracticable for various reasons, including ethical concerns, so, instead, we use the retrospective observational data to answer this question. To estimate the causal impact of the FT routing decision on patient outcomes, we consider the following probit and log-linear regression models:

$$Revisit_i = \begin{cases} 1 & \text{if } \hat{\beta} \mathbf{X_i} + \hat{\gamma} FT_i + \hat{\omega}_h + \hat{\tau}_m + \hat{\theta}_t + \hat{\varepsilon}_i > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

$$\log(LOS_i) = \tilde{\beta} \mathbf{X_i} + \tilde{\gamma} FT_i + \tilde{\omega}_h + \tilde{\tau}_m + \tilde{\theta}_t + \tilde{\varepsilon}_i. \tag{3}$$

Note that i represents a patient in all three baseline models (1), (2) and (3). The vector $\mathbf{X_i}$ includes the age group, gender, chief complaint, triage score, and triage time of patient i. The variables ω_h, $\hat{\omega}_h$, and $\tilde{\omega}_h$ represent the hospital fixed effect; τ_m, $\hat{\tau}_m$, and $\tilde{\tau}_m$ represent the month-year fixed effect; θ_t, $\hat{\theta}_t$, and $\tilde{\theta}_t$ represent the weekday fixed effect; the random error terms ε_i, $\hat{\varepsilon}_i$, and $\tilde{\varepsilon}_i$ follow normal distributions.

One may choose to estimate Model (2) via the maximum likelihood estimation (MLE) and Model (3) via ordinary least squares, and then interpret the estimated parameters $\hat{\gamma}$ and $\tilde{\gamma}$ as the causal effects of being routed to FT on *Revisit* and *LOS*, respectively. However, such an approach ignores that the FT routing decisions may be endogenous due to factors that were observed by triage nurses during triage but unobservable in our

data, such as patient mental state and level of pain. The 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). Hence, we need to 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., KC and Terwiesch (2012), Kim et al. (2015), Chan et al. (2018), and Song et al. (2020). Specifically, following Song et al. (2020), we propose the ratio between *MainCongestion* and *EDCongestion*, denoted by $MEBusyRatio_i$, as our IV. This continuous variable characterizes the relative congestion condition between the main area and the entire ED at the triage start time (see the summary statistics in Table 1).

Next, we discuss the validity of this IV. We start with the inclusion condition. 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 correlation between the relative congestion condition of the main area and FT routing decisions. We further validate this inclusion condition statistically through the first-stage regression results (see the full estimation results in Tables 10–13 in the appendix). We observe that the coefficients for $MEBusyRatio_i$ in all the 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. This result indicates that our identification is not weak.

Finally, we consider the exclusion condition, i.e., the busyness ratio $MEBusyRatio_i$ affects patient outcomes only through the FT routing decision. We note that one potential connection between the busyness ratio and patient outcomes (other than through the FT routing decision) is through either the area occupancy level or the patient complexity level. Hence, we control for the occupancy level during patient i's visit ($AvgOccTreated_i$) in the area to which the patient is routed. Following similar ideas in Kim et al. (2015), Chan et al. (2018), and Song et al. (2020), we define the occupancy level, denoted by $AvgOccTreated_i$, as the time-averaged number of patients during the period when patient i is under diagnosis and treatment, which is calculated as the total time of all patients staying in this area during patient i's stay in this area divided by the length of that period. Therefore, by controlling for the occupancy level, we can block the path between the busyness ratio $MEBusyRatio_i$ and patient outcomes. As a result, $MEBusyRatio_i$ can only affect

patient outcomes through the FT routing decision. In addition to the area occupancy level, we also control for a patient's complexity level. Note that the triage level itself may not be enough to measure a patient's complexity level, which considers both patient clinical urgency and resource needs (see more discussions on this in Section 4.4). Moreover, ED operational factors may also affect the assignment of a patient's triage level (Chen et al. 2020b). We, therefore, approximate the unobserved patient complexity levels that were embedded in the error term by controlling for the observed patient and operational characteristics, such as age group, chief complaint, and triage time, which are the same set of variables used for the patient classification in Section 4.4.

4.3. Estimation

In this section, we describe our estimation approaches for patient outcomes. We consider two types of patient outcomes: patient revisit rate (a binary variable) and *LOS* (a continuous variable). The treatment variable here is the FT routing decision, which is a binary variable (i.e., 1 stands for being routed to the FT and 0 stands for being routed to the main area). The FT routing decision variable is endogenous, as we discussed previously. Hence, we consider two separate models for the continuous and binary outcome variables with an endogenous binary treatment variable.

4.3.1. Continuous Patient Outcome Variable The *LOS* is a continuous patient outcome variable, whereas the endogenous treatment variable, i.e., the FT routing decision, is binary. Hence, we cannot directly apply a standard two-stage least square (2SLS) approach, given the binary endogenous variable. As a result, we choose a latent variable approach to jointly estimate the probit model on FT routing decisions and the linear outcome model (Maddala 1986 and Heckman 1977). This approach explicitly models the correlation between the unobservables that affect the endogenous treatment variable and the outcomes as follows:

$$FT_i^* = \beta \mathbf{X_i} + \alpha MEBusyRatio_i + \omega_h + \tau_m + \theta_t + \varepsilon_i, \tag{4}$$

$$FT_i = \mathbb{1}\{FT_i^* > 0\}, \tag{5}$$

$$\log(LOS_i) = \beta' \mathbf{X_i} + \gamma FT_i + \delta AvgOccTreated_i + \omega_h' + \tau_m' + \theta_t' + \xi_i, \tag{6}$$

where FT_i^* is the latent variable associated with the binary treatment variable FT_i, and ε_i and ξ_i are the error terms for the FT routing model and the patient outcome model, respectively. 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. The vector $\mathbf{X_i}$ includes age group, gender, chief complaint, triage level, and triage time. We control for the average treatment area utilization level ($AvgOccTreated_i$) to account for hospital resource utilization (busyness) during patient i's diagnosis and treatment. Finally, we control for the hospital fixed effects (ω_h, ω_h'), the month-year fixed effects (τ_m, τ_m'), and weekday fixed effects (θ_t, θ_t'), and cluster the standard errors by the ID of the physician who initially treated the patient. We jointly estimate the routing decision and outcome equations through the full

maximum likelihood estimation (FMLE). The dependent variable *LOS* here is log-transformed due to the skewness concern of its distribution.

4.3.2. Binary Patient Outcome Variables We next consider the outcome variable $Revisit_i$, which is binary and represents both the 48-hour revisit ($Revisit_{48h}$) and 72-hour revisit ($Revisit_{72h}$) for patient i. Hence, both the outcome variable and the FT routing decision variable are binary. We employ the recursive bivariate probit model (Maddala 1986, Greene 2018) to estimate the effect of the endogenous binary variable (FT_i) on binary outcome variable ($Revisit_i$). The biprobit model contains two probit models with correlated error terms as follows:

$$FT_i^* = \beta \mathbf{X_i} + \alpha MEBusyRatio_i + \omega_h + \tau_m + \theta_t + \varepsilon_i, \tag{7}$$

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

$$Revisit_i^* = \beta' \mathbf{X_i} + \gamma FT_i + \delta AvgOccTreated_i + \eta WaitTime_i + \omega_h' + \tau_m' + \theta_t' + \xi_i, \tag{9}$$

$$Revisit_i = \mathbb{1}\{Revisit_i^* > 0\}, \tag{10}$$

where FT_i^* is the latent variable associated with the binary treatment variable FT_i; $Revisit_i^*$ is the latent variable associated with the binary outcome $Revisit_i$; 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. All other control variables are the same as those described in Section 4.3.1. We further control for a patient's waiting time, denoted by $WaitTime_i$. Finally, we cluster the standard errors by the name of the physician who performed the initial assessment and estimate the model through FMLE.

4.4. Patient Classification by Complexity Levels

As mentioned earlier, the FT area is designed to treat patients with less urgent and less complex health issues with the goal of delivering care more quickly. However, triage nurses may consider both clinical and operational factors when making routing decisions; for example, in our study EDs, triage nurses may have considered the ED operational status when making FT routing decisions. Therefore, patients with more complex conditions (who would be routed to the main area under less congested situations) may have been routed to the FT area when the main area is highly congested. It is thus unclear whether there are any hidden unintended consequences. In this section, we discuss patient classification, which enables us to later quantify the heterogeneous effects of FT routing decisions on different patient complexity groups to shed light on the hidden consequences.

A fundamental step that must be defined in order to conduct patient classifications is the procedure for measuring the complexity of a patient's health issues. One immediate option may be to equate urgency levels to complexity levels and route all (and only) patients of triage levels 4 and 5 into the FT. Indeed, this has

been a popular practice in many hospital EDs (Peck and Kim 2010, Arya et al. 2013, Song et al. 2015, Ding et al. 2019), despite the lack of evidence to support this practice (Peck and Kim 2010). However, diagnosing and treating a low-urgency patient is not necessarily easier and faster than diagnosing and treating a patient with a higher urgency level (Ieraci et al. 2008). Moreover, resource needs (such as diagnostic tests and expected need for admission) and nursing requirements must also be considered when classifying patients by complexity.

In this regard, a patient streaming strategy based on predicted disposition (admission to hospital vs. discharge from the ED) has been found to be successful by ED practitioners (O'Brien et al. 2006, Kelly et al. 2007). In the operations management 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 as the outcome variable and estimate the following probit model based on patient information collected at triage:

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

where $\mathbf{X_i^p}$ is a vector of patient characteristics including triage level, age group, gender, and chief complaint, and ϕ_i is the unobserved component. We construct patient complexity classes by partitioning the fitted probability of admission, denoted as $\hat{M}_i$, which is calculated by $\hat{M}_i = \Phi(\hat{\beta}_p \mathbf{X_i^p})$, where $\hat{\beta}_p$ is the estimated β_p. 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. The fitted probability distribution for patients routed to the main area and FT is shown in Figure 6 in the appendix. We observe that most patients with a high $\hat{M}_i$ were routed to the main area, whereas most patients with a low $\hat{M}_i$ were routed to the FT area. Nevertheless, we observe a few patients with a high $\hat{M}_i$ who were routed to the FT area and vice versa. 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.

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 middle-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 $\hat{M}_i$. The criterion for choosing these thresholds is that the values of $\hat{M}_i$ for patients of the same complexity class should be close, while for patients in different complexity levels should be distant. As a result, we choose t_1 and t_2 to be the 35th and 75th percentiles of $\hat{M}_i$, respectively. Tables 8 and 9 in the appendix present the summary statistics of patient characteristics and patient outcomes (i.e., the revisit rates and LOS), respectively, for patients of the three complexity classes. Based on this classification, over 93% of patients in the high-complexity class were routed into the main area, and over 53% of patients in the

low-complexity class were routed into the FT area (see Table 7 in the appendix). We can then estimate the heterogeneous effects of the FT routing decision on patient outcomes by applying the latent variable model and the IV discussed in the previous section to each of the three complexity classes. We also run robustness checks with alternative choices for t_1 and t_2, and our empirical results remain consistent.

5. Estimation Results

In this section, we present our estimation results. Section 5.1 provides the main results on the factors associated with FT routing decisions and the impact of routing decisions on patient outcomes. Section 5.2 outlines our examination of the heterogeneous impact of the FT routing decisions on different patient complexity groups. Section 5.3 includes a discussion on potential drivers of the observed effects of the FT routing decisions, and Section 5.4 presents the robustness checks.

5.1. Estimation Results for All Patients

We start our discussion with the main results related to the factors associated with FT routing decisions and the impact of routing decisions on patient outcomes using all patient visit data.

Observation 1. *FT routing decisions are not purely clinical-driven, and ED operational status related to congestion is also associated with FT routing decisions.*

Using the probit model (1) introduced in Section 4.1, we present the full estimation results on FT routing decisions in the second column of Table 10 in the appendix. We first find that clinical factors, such as age group, triage level, gender, and triage time, are associated with FT routing decisions. Interestingly, we also find that the coefficient of $MEBusyRatio_i$ is 0.070 (p-value < 0.001), 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. This suggests that FT routing decisions are not purely clinical-driven, ED operational status related to congestion is also important in the decision-making process.

Observation 2. *Being routed to the FT area reduces the average LOS, whereas its impact on ED revisit rate is statistically insignificant.*

We present our empirical results using all patient visit record data in Table 2. In particular, we show results both with and without IV to illustrate the potential estimation bias without IV. In Table 2, we present the average marginal effect (AME) for each outcome variable when the patient is routed to the FT area. The first two rows show the effect of being routed to FT on the 48- and 72-hour revisit rates, respectively. The last row shows the effect on *LOS*. Table 10 in the appendix presents the full estimation results.

We discover that being routed to the FT area reduces the average *LOS* (i.e., a negative coefficient of -0.391 with p-value < 0.01), consistent with the medical literature that FT improves care access (Sanchez et al. 2006, Devkaran et al. 2009, Chrusciel et al. 2019). Specifically, being routed to the FT leads to an average reduction of 0.91 hospital hours in *LOS*, i.e., a 25.9% reduction. This is calculated as the difference

between the predicted *LOS* when patients were routed to the main area and when patients were routed to the FT area (i.e., $3.51 - 2.60 = 0.91$), similar to Chan et al. (2018). Next, we do not find statistically significant effects on $Revisit_{48h}$ or $Revisit_{72h}$ using all patient visit data, which is consistent with the findings reported by Devkaran et al. (2009). In contrast, Ieraci et al. (2008) and Chrusciel et al. (2019) find a statistically significant increase in the 48-hour and 30-day revisit rates, respectively, with the implementation of FT.

We also present the estimated correlation $\rho(SE)$ between the error terms of the routing decision equation and the outcome equation, after which we perform the likelihood ratio test "Test $\rho = 0$" that compares the log-likelihood of our full model with the sum of log-likelihood of two separate models. Column "Test $\rho = 0$" presents the p-values of the likelihood ratio test results. Similar to the Hausman test, this likelihood ratio test checks the exogeneity of a dummy independent variable with a dummy dependent variable (Knapp and Seaks 1998). The likelihood ratio test results for the 48- and 72-hour revisits do not statistically indicate the endogeneity issue for the FT routing decision. Also, the results with and without IV are quite similar. However, the likelihood ratio test strongly indicates the endogeneity issue for the dependent variable *LOS* (p-value = 0.001), which leads to a negative bias for the estimation (i.e., -0.209 vs. -0.391).

Table 2 The average marginal effect (AME) of the fast-track routing decision on patient outcomes for all patients and for high-complexity, middle-complexity, and low-complexity patients.

	With IV			Without IV
Outcome variables	AME (SE)	ρ (SE)	Test $\rho = 0$	AME (SE)
All patients ($N = 153,105$)				
$Revisit_{48h}$	0.002 (0.008)	0.002 (0.038)	0.953	0.002 (0.003)
$Revisit_{72h}$	-0.001 (0.009)	0.009 (0.037)	0.815	0.001 (0.003)
$\log(LOS)$	-0.391*** (0.062)	0.148*** (0.045)	0.001	-0.209*** (0.014)
High-complexity patients ($N = 38,097$)				
$Revisit_{48h}$	0.074* (0.040)	-0.171** (0.086)	0.050	0.013* (0.008)
$Revisit_{72h}$	0.054 (0.041)	-0.112 (0.093)	0.230	0.013 (0.008)
$\log(LOS)$	-0.384*** (0.105)	0.011 (0.071)	0.881	-0.370*** (0.022)
Middle-complexity patients ($N = 61,252$)				
$Revisit_{48h}$	0.030** (0.014)	-0.090** (0.044)	0.041	0.007* (0.004)
$Revisit_{72h}$	0.028** (0.014)	-0.083** (0.041)	0.041	0.005 (0.004)
$\log(LOS)$	-0.423*** (0.047)	0.154*** (0.032)	0.000	-0.229*** (0.018)
Low-complexity patients ($N = 53,756$)				
$Revisit_{48h}$	-0.014 (0.012)	0.095 (0.086)	0.274	-0.002 (0.002)
$Revisit_{72h}$	-0.006 (0.012)	0.008 (0.077)	0.919	-0.004* (0.003)
$\log(LOS)$	-0.669*** (0.060)	0.413*** (0.045)	0.000	-0.151*** (0.014)

Notes. Standard errors (SEs) clustered by the name of the physician who performed 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. Heterogeneity Among Complexity Classes

Our estimation results on all patient visit data reveal that FT routing can reduce patient LOS but does not affect the quality of care (measured by revisit rates), which seems to align with the key objective of initiating an FT line. However, we also uncover a positive correlation between the ED congestion level and the likelihood of being routed to the FT area. Therefore, a follow-up issue to be examined is whether any hidden consequences exist for patients who were treated in FT but would have been routed to the main area in a less congested ED. To address this question, we investigate the heterogeneous effects of the FT routing decisions on different patient complexity classes. A summary of the main results is shown in Table 2. The full estimation results are shown in Tables 11–13 in the appendix.

Observation 3. *Being routed to FT (i) reduces the average LOS for patients in each complexity class; (ii) increases the 48-hour revisit rate for high- and middle-complexity patients; and (iii) increases the 72-hour revisit rate for middle-complexity patients.*

High-Complexity Patients As Table 2 indicates, being routed to FT reduces the LOS for high-complexity patients (a negative coefficient -0.384, p-value < 0.01). Specifically, being routed to the FT leads to an average reduction of 1.11 hours in *LOS* (a 25.5% reduction), which is the difference between the predicted *LOS* of high-complexity patients when routed to the main area and when routed to the FT area (i.e., $4.35 - 3.24 = 1.11$). Next, in contrast to Observation 2, we find that being routed to the FT area hurts the quality of care for high-complexity patients in terms of increasing their revisit rates (a positive coefficient 0.074, p-value < 0.1). In other words, being routed to the FT increases the 48-hour revisit rate by 7.4%. On the other hand, we find from the full estimation results in Table 11 that waiting for an additional hour only increases the 48-hour revisit rate by 0.3%. Finally, the p-value of the likelihood ratio test for the 48-hour revisit rate is 0.050, indicating a strong endogeneity issue, which also explains the difference between the results with IV and without IV. In addition, we do not find statistically significant evidence of the endogeneity issue for $Revisit_{72h}$ and *LOS*.

Middle-Complexity Patients We first observe that being routed to FT reduces the LOS for middle-complexity patients (a negative coefficient -0.423, p-value < 0.01). Specifically, being routed to the FT leads to an average reduction of 0.43 hours in *LOS* (a 27.7% reduction), calculated as the difference between the predicted *LOS* of middle-complexity patients when routed to the main area and when routed to the FT area (i.e., $3.54 - 2.56 = 0.98$). Next, we find positive and statistically significant effects of being routed to FT on both 48- and 72-hour revisit rates for the middle-complexity patients. Specifically, being routed to FT increases both the 48- and 72-hour revisit rates by around 3%. Finally, comparing the results with and without IV, we again observe a negative bias introduced by the omitted variables, and the likelihood ratio test indicates the existence of a strong endogeneity issue across all the outcome variables.

Low-Complexity Patients Similar to the previous results, we find that being routed to FT reduces the LOS for low-complexity patients (a negative coefficient -0.669, p-value < 0.01). Specifically, being routed

to the FT leads to a 1.31-hour reduction in *LOS* (an 40.1% reduction), i.e., the difference between the predicted *LOS* of low-complexity patients when routed to the main area and when routed to the FT area (i.e., 3.27 – 1.96 = 1.31). We do not find statistically significant effects of the FT routing decisions on the 48- or the 72-hour revisit rates for the low-complexity patients. This finding supports the essential goal of introducing the FT line, that is, to treat low-complexity patients faster and improve operational efficiency without compromising the quality of care. Finally, comparing the results with and without IV, we observe a negative bias introduced by the omitted variables for the outcome variable *LOS*. As for the 48- and 72-hour revisit rates, the p-values for the likelihood ratio test do not indicate an endogeneity issue, and the results are similar with and without IV.

5.3. Discussion on the Mechanism

Our results, as presented in Sections 5.1 and 5.2, show that being routed to the FT area significantly reduces the LOS for all patient classes. However, there are hidden consequences for patients of high- and middle-complexity levels; that is, being routed to FT increases their revisit rates. To understand the potential drivers of this finding, we explore the differences in the diagnosis and treatment process between patients treated in the main area and those treated in the FT area. In particular, we examine the following factors which are available in our data: the number of lab tests, the number of medications, and the number of diagnostic images (CT scans and X-ray tests). These factors capture the complexity of a patient's diagnosis and treatment process after being routed to a particular treatment area. Since all these factors are count variables, we employ a negative binomial model to estimate the effect of the FT routing decision while controlling for age group, gender, chief complaint, triage score, triage time, hospital fixed effect, month-year fixed effect, and weekday fixed effect; see the main estimation results in Table 3.

Table 3 Estimation results on the impact of the fast-track routing decisions on the numbers of lab tests, medication orders, CT (computerized tomography) scans, and X-ray tests for patients of different complexity levels.

	Observations	Lab Tests	Medications	CT Scans	X-ray Tests
High-complexity patients	38,097	-1.14***	-0.53***	-0.63***	-0.00
		(0.03)	(0.03)	(0.08)	(0.03)
Middle-complexity patients	61,252	-1.00***	-0.36***	-0.69***	0.15***
		(0.02)	(0.02)	(0.05)	(0.02)
Low-complexity patients	53,756	-1.03***	-0.27***	-0.51***	0.16***
		(0.03)	(0.02)	(0.06)	(0.02)

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

As Table 3 illustrates, we find that being routed to the FT area reduces the number of lab tests, the number of medications, and the number of CT scans for patients of all complexity classes, implying that physicians tend to simplify the diagnosis and treatment process for patients in the FT area. However, the FT routing

decision increases the number of X-ray tests for middle- and low-complexity patients, whereas its impact on the X-ray orders for high-complexity patients is statistically insignificant. We conjecture that physicians working at FT may tend to order fewer CT scans, which produce detailed, high-quality images, and order more X-ray tests, which are less expensive, faster, but convey less detail than CT scans. Physicians working at FT understand that the purpose of setting up an FT line is for the fast delivery of care for less complex and less urgent patients. Hence, they may get used to speeding up the patient flow process by simplifying patient diagnostic procedures. Consequently, the quality of care may be compromised for high- and middle-complexity patients who were routed to the FT area. On the other hand, low-complexity patients are often easier to diagnose and are much less likely to revisit the ED (lower 48- and 72-hour revisit rates), compared to high- and middle-complexity patients (see Table 9). Hence, the quality of care for low-complexity patients treated in the FT area is not affected.

5.4. Robustness Check

In this section, we present our robustness checks. Table 4 presents the robustness check results for all patients, and Table 5 shows the results for patients from different complexity classes.

We first consider alternative IVs. The IV used in our model is computed as the relative congestion level between the main area and the entire ED. As mentioned previously, the congestion level in a particular area is computed as the number of patients handled per physician in that particular area divided by the physician capacity. In this way, we adjust the congestion level by the number of physicians in a particular area. However, from triage nurses' perspective, the congestion level can be purely determined by the number of patients in a particular area, which is directly observable. Hence, we consider an alternative IV that does not adjust the number of physicians working in the corresponding area and uses the area capacity instead of the physician capacity. In other words, the new congestion level is calculated as the total number of patients in that area divided by the area capacity. See Section 3.3.2 for more details. Panel (1) of Tables 4 and 5 show the estimation results using this alternative IV, which are consistent with our main estimation.

Next, nurses may have used the past congestion information to inform current routing decisions due to the delayed congestion effect. Therefore, we consider alternative IVs using the relative congestion level measure of 30 minutes, 1 hour, and 2 hours before the triage start time. Panels (2), (3), and (4) in Tables 4 and 5 show the estimation results with these alternative IVs: all results are consistent with our main estimation.

We next consider alternative cutoffs to partition the patient complexity classes. In our main analysis, the cutoffs t_1 and t_2 are the 35th and 75th percentiles, respectively. To show the robustness of our results, we vary t_1 and t_2. In particular, we test four pairs of the thresholds (t_1, t_2), including the 30th and 45th percentiles for t_1, the 65th and 85th percentiles for t_2, and their combinations. The results are shown in panels (5), (6), (7), and (8), respectively, in Tables 4 and 5. We find that the results are consistent with our main estimation.

Finally, we remove extreme observations with a triage time longer than 17 minutes (i.e., 99.9% percentile), see panel (9) in Tables 4 and 5. The results are again consistent with our main estimation.

Table 4 Average marginal effect of FT routing on patient outcomes for the nine robustness checks (all patients).

All patients								
$Revisit_{48h}$	$Revisit_{72h}$	log(LOS)	$Revisit_{48h}$	$Revisit_{72h}$	log(LOS)	$Revisit_{48h}$	$Revisit_{72h}$	log(LOS)
Panel (1)			Panel (2)			Panel (3)		
0.00	-0.00	-0.50***	0.00	-0.00	-0.36***	-0.00	-0.00	-0.32***
(0.01)	(0.01)	(0.06)	(0.01)	(0.01)	(0.06)	(0.01)	(0.01)	(0.06)
Panel (4)			Panel (5)			Panel (6)		
-0.00	-0.01	-0.27***	0.00	-0.00	-0.39***	0.00	-0.00	-0.39***
(0.01)	(0.01)	(0.06)	(0.01)	(0.01)	(0.06)	(0.01)	(0.01)	(0.06)
Panel (7)			Panel (8)			Panel (9)		
0.00	-0.00	-0.39***	0.00	-0.00	-0.39***	0.00	-0.00	-0.39***
(0.01)	(0.01)	(0.06)	(0.01)	(0.01)	(0.06)	(0.01)	(0.01)	(0.06)

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

Table 5 Average marginal effect of FT routing on patient outcomes for the nine robustness checks (subgroups).

	High-complexity patients			Middle-complexity patients			Low-complexity patients		
	$Revisit_{48h}$	$Revisit_{72h}$	log(LOS)	$Revisit_{48h}$	$Revisit_{72h}$	log(LOS)	$Revisit_{48h}$	$Revisit_{72h}$	log(LOS)
Panel (1)	0.07*	0.05	-0.49***	0.03**	0.03**	-0.51***	-0.01	-0.01	-0.78***
	(0.04)	(0.04)	(0.13)	(0.01)	(0.01)	(0.05)	(0.01)	(0.01)	(0.05)
Panel (2)	0.07*	0.05	-0.39***	0.03**	0.02*	-0.41***	-0.01	-0.01	-0.62***
	(0.04)	(0.04)	(0.11)	(0.01)	(0.01)	(0.05)	(0.01)	(0.01)	(0.06)
Panel (3)	0.07*	0.05	-0.39***	0.03*	0.02*	-0.40***	-0.02	-0.00	-0.55***
	(0.04)	(0.04)	(0.11)	(0.01)	(0.01)	(0.05)	(0.01)	(0.01)	(0.07)
Panel (4)	0.07*	0.05	-0.39***	0.03*	0.02	-0.38***	-0.02	-0.01	-0.43***
	(0.04)	(0.04)	(0.11)	(0.01)	(0.01)	(0.05)	(0.01)	(0.01)	(0.08)
Panel (5)	0.22*	0.20	-0.51**	0.04***	0.04***	-0.37***	-0.01	-0.01	-0.67***
	(0.12)	(0.14)	(0.23)	(0.01)	(0.01)	(0.05)	(0.01)	(0.01)	(0.06)
Panel (6)	0.05**	0.03	-0.38***	0.03*	0.03*	-0.43***	-0.01	-0.01	-0.67***
	(0.02)	(0.02)	(0.09)	(0.01)	(0.01)	(0.05)	(0.01)	(0.01)	(0.06)
Panel (7)	0.07*	0.05	-0.38***	0.03*	0.03	-0.47***	-0.00	0.00	-0.63***
	(0.04)	(0.04)	(0.11)	(0.02)	(0.02)	(0.06)	(0.01)	(0.01)	(0.05)
Panel (8)	0.07*	0.05	-0.38***	0.02	0.01	-0.41***	-0.01	-0.00	-0.67***
	(0.04)	(0.04)	(0.11)	(0.01)	(0.01)	(0.05)	(0.01)	(0.01)	(0.07)
Panel (9)	0.08*	0.06	-0.37***	0.03*	0.02*	-0.42***	-0.01	-0.01	-0.66***
	(0.04)	(0.04)	(0.11)	(0.01)	(0.01)	(0.05)	(0.01)	(0.01)	(0.06)

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

6. Evaluation of Alternative Fast-Track Routing Policies

In previous sections, we have investigated the impact of FT routing decisions on patient outcomes. Next, we propose a multi-class routing problem with two parallel queues to study the optimal routing policy. Specifically, we model the problem using the Markov decision process (MDP) and leverage our estimation results from Section 5 to calibrate the model parameters. We draw insights from the structure of the optimal policy and propose heuristic routing policies, the performances of which are evaluated 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-, middle-, 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 and a queue with infinite capacity. At station j, the service times (*diagnosis and treatment time*) are independent and identically distributed exponential random variables with a mean of $1/\mu_j$, $j = 1, 2$, for all patients. In our study hospitals, a single physician is scheduled to work in the FT area when it is open. However, multiple physicians could be working in parallel in the main area. We argue that the single server assumption is reasonable since physicians are almost always busy, and we use a super server to represent the joint efforts of all physicians. We further assume that patients are served on a first-come-first-served (FCFS) basis at each station.

Upon arrival, a patient will be routed into one of the two queues by the decision maker (i.e., triage nurses), where he or she waits to be seen. If a patient of class i is routed into queue j, a cost $r_{ij}(t)$ will be incurred after the completion of service at station j, $j = 1, 2$, if the patient has 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 2). The cost also depends on the patient's waiting time, as shown by our empirical results (see Tables 11–13), 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 type to a certain extent. In our simulation study, we explicitly account for patient characteristic information when estimating the cost term $r_{ij}(t)$. The objective of the decision maker is then 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 into any queue to keep our model general. However, as we show later in Figure 4, the optimal policy almost never routes any patient of high-complexity level into 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 area 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 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, respectively, routing the patient to the main area and the FT area.

Let $V_t(\boldsymbol{x}, \pi)$ 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 $\Gamma = \sum_{i=1}^{3} \lambda_i + \sum_{i=1}^{2} \mu_i$. 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, and let $\boldsymbol{e_1} \equiv (1, 0)$ and $\boldsymbol{e_2} \equiv (0, 1)$. Then, the Bellman's 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

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

where $\mathbb{1}_{\{x_i \geq 1\}} = 1$ indicates $x_i \geq 1$, and $\mathbb{1}_{\{x_i \geq 1\}} = 0$ 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 are 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 results would better reflect the 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 study. 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 in (4) 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:

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

where ξ_i is the unobserved patient information, γ_i is the coefficient for the routing decision dummy variable, and h_i is the cost per unit time a class i patient waits in the system. The noise term ε_i conditioning on ξ_i follows a normal distribution with a mean of $\rho\xi_i$ and a variance of $1 - \rho^2$, where ρ is the correlation coefficient of ε_i and ξ_i.

6.2. Simulation Design, Input Modeling, and Validation

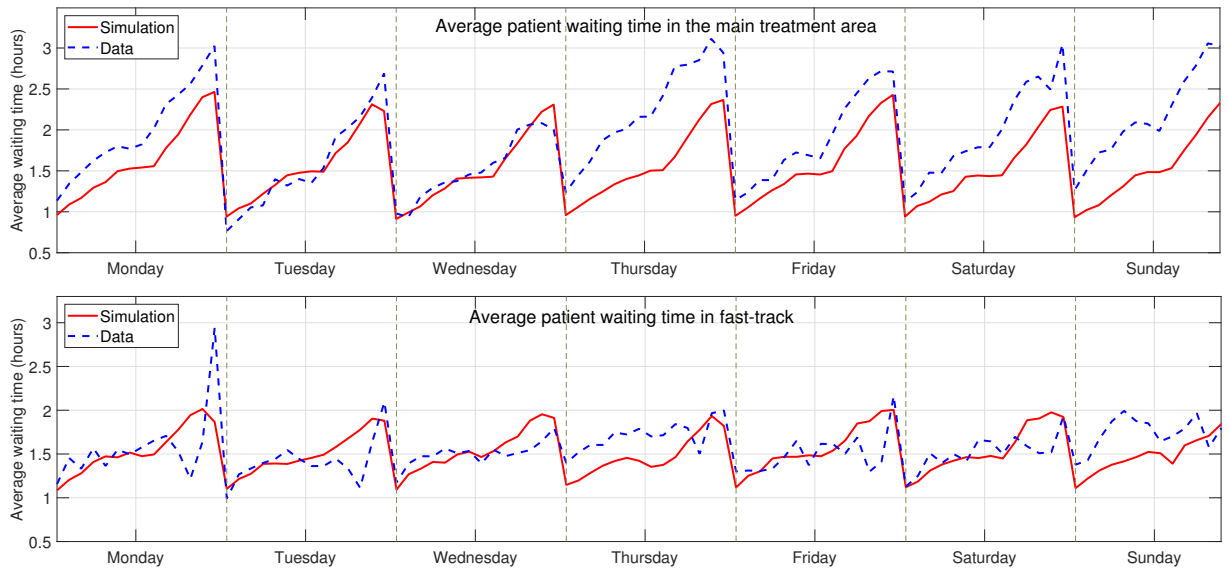
We use a discrete event simulation model to simulate the ED patient flow as a multi-class queue with two parallel stations. Upon the arrival of a new patient, the patient will be classified into one of the three complexity classes based on the approach in Section 4.4 and then routed into one of the two queues according to a specific routing policy. The patient stays in the queue until the completion of his or her service and then leaves the system. The objective of the simulation is to compare different routing policies, which are defined in Section 6.3.

The input modeling is based on patient visit data collected between November 2013 and April 2015. Our descriptive statistical analysis shows that the arrival process depends on both time of day and day of the week. Hence, we choose one week as a cycle and treat the data of each week as samples from the stochastic process underlying the true arrival process. We relax the stationary assumption on the arrival process and generate the inter-arrival times in a non-parametric manner. More specifically, we first collect from the data all the inter-arrival times indexed by time of day and day of the week. Then, depending on the current time of day and day of the week of the simulation clock, we sample replacements from the corresponding set of inter-arrival arrival times (bootstrap). Next, we randomly sample a patient in the same manner and assign this patient's profile (age, gender, triage level, and so on) to the newly arrived patient in the simulation. We follow the approach in Section 4.4 to determine the patient's complexity class. The service times at both stations are assumed to follow exponential distributions with time-dependent rates estimated from the data. The FT areas in our study hospitals operate from 10 a.m. to midnight, and 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 he or she leaves work. However, all patients waiting in the FT area are moved to the main area instantaneously.

In our data, most admitted patients were treated in the main ED area. Moreover, the 48- and 72-hour revisit rates are *not* appropriate outcome measures for admitted patients because they usually stay in the hospital for longer periods of time than 48 or 72 hours. Hence, admitted patients will not generate any revisit cost and, thus, were excluded from the data for our empirical study. They will, however, affect the waiting time (thus revisit cost) of discharge patients. In the simulation, we include admitted patients and assume that they will be treated in the same area as they are in the data.

We first validate our simulation model with data. We begin the simulation with an empty ED and run 50 replications with a replication length of 365 days. The routing policy used in the simulation is based on the estimated routing policy, using data from our study EDs (see Section 6.3 for details). For each replication, we identify the first 30 days as the warm-up period using Welch's graphical method (Law and Kelton 2000). We then use the output to calculate the time-dependent average waiting time in the main area and in the FT area. The weekly average patient waiting times from the simulation and from the data are shown in Figure

Figure 3 Comparison of the average patient waiting times from simulation and from data.



3, which provide evidence that our simulation model captures the trend of the average waiting time from the data reasonably well.

6.3. Fast-Track Routing Policies

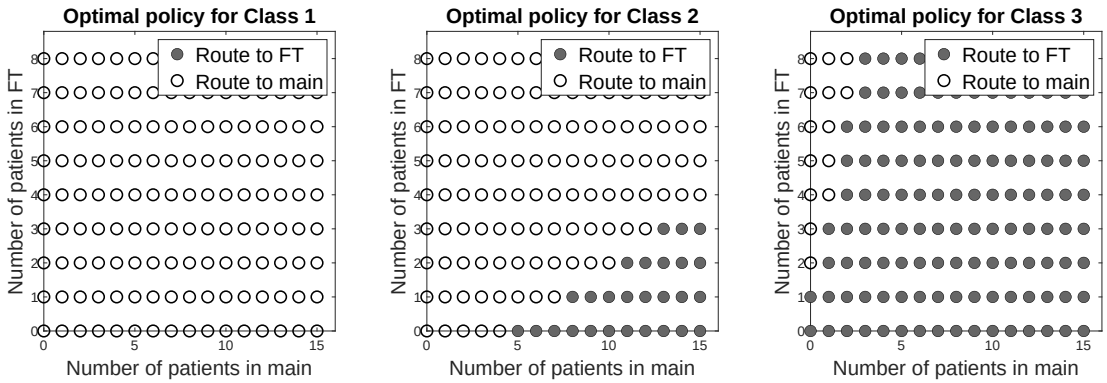
In this section, we compare five FT routing policies through simulations. We first describe the policies of interest explicitly.

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, gender, and so on; and x_1 and x_2 are the number of patients waiting in the main area 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 from 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. However, we include it as a heuristic policy in the policy comparison.

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 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. The dynamic routing mainly applies to Class 2 (i.e., middle-complexity) patients. Specifically, when the main area becomes crowded, it is optimal to route more patients of Class 2 to the FT area to reduce their waiting times, which also eases the congestion level in the main area. Motivated by the structure of the optimal routing policy and the insights noted, we propose

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



the following static routing policies, which are easier to implement because they are state-independent and do not require solving an MDP.

Static Routing Policy (SP): According to the static routing policy, 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. In the simulation study, the percentiles are chosen from $\{25, 30\}$. Hence, the corresponding static routing policies are denoted as SP-25 and SP-30.

Triage-Score-Based Routing Policy (TP): In the simulation study, we also consider the routing policy that routes (i) patients of triage levels 4 and 5 to the FT area, and (ii) patients of triage levels 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 a patient arrival process under different routing policies. We run the simulation under each policy for 50 replications, each replication with a length of 365 days. For each replication, we identify the first 30 days 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 6 shows the average waiting time and the 48-hour patient revisits and their corresponding 95% confidence intervals for the five routing policies under consideration. 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 6 (see a graphical comparison in Figure 5), from which we make the following observations.

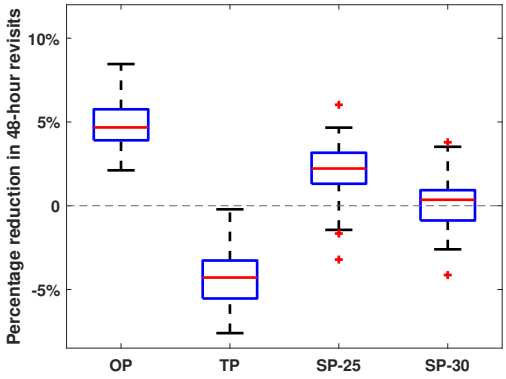
Observation 4. *The state-dependent policy OP performs the best among all the routing policies in terms of reducing both the 48-hour patient revisits and the average patient waiting time.*

Table 6 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	5693 ± 12	5424 ± 12	5950 ± 23	5576 ± 12	5681 ± 12
Reduction in 48-hour patient revisits (%)		4.72 ± 0.27	-4.52 ± 0.50	2.05 ± 0.40	0.21 ± 0.32
Average waiting time (hours)					
All patients	1.58 ± 0.03	1.23 ± 0.01	1.73 ± 0.05	1.68 ± 0.03	1.75 ± 0.03
Patients in main area	1.54 ± 0.03	0.99 ± 0.01	1.69 ± 0.06	1.61 ± 0.04	1.19 ± 0.02
Patients in FT area	1.70 ± 0.03	1.86 ± 0.03	1.85 ± 0.03	1.86 ± 0.04	3.04 ± 0.07

Notes: The calculation of the 48-hour patient revisits is based on the total number of discharged patients during FT opening hours for the three EDs in 18 months (i.e., 116,899).

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



The percentage reduction in the 48-hour patient revisits by Policy OP over the current routing policy used in our study EDs, i.e., Policy CP, can be 4.72%. At the same time, Policy OP reduces the average waiting times of all patients by 22.2%, 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. Our simulation results show that 27.65% patients are routed to the FT area under Policy OP, whereas the FT area treats 25% patients under Policy CP.

Observation 5. *Policy SP-25 reduces the 48-hour patient revisits but increases the average patient waiting time.*

The static routing policy SP-25 is interesting because, under this policy, the same percentage of patients are routed into FT as under Policy CP. However, SP-25 can reduce the 48-hour patient revisits over CP by close to 2%, which implies that our patient classification can pick out the "right" patients to be routed to FT to reduce revisits and improve patient outcomes. SP-25 is similar to TP in that it does not take the ED congestion information into the FT routing decision. This also explains the longer average waiting time under SP-25 compared to CP, under which triage nurses consider ED congestion in the FT routing decision.

Observation 6. *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 8.8% and 4.3%, 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 since SP-25 can pick out the "right" patients who are safer to be treated in the FT area. Moreover, the average waiting time under TP is the longest among all five policies under consideration (the difference in waiting time between TP and SP-30 is statistically insignificant).

To summarize, 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 dynamical routing policy benefits from the server pooling effect, which, to a certain extent, makes up the "anti-pooling" deficit from setting up the FT line 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 Policy OP and Policy SP-25 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 FT routing decisions on patient outcomes using data from three 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-hour revisit rate increases by 7.4% for high-complexity patients and by 3.0% for middle-complexity patients. Therefore, we advise caution since it has unintended consequences on the quality of care, especially for patients with more complex care needs. We believe this is the first work that provides empirical evidence quantifying the causal effects of FT routing decisions on various patient outcomes based on their complexity groups and documenting the trade-off

between care access and quality of care. Being aware of this important trade-off, 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 at triage levels 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 the access to emergency care and the quality of care when making the FT routing decisions.

As more hospitals have implemented FT lines 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 three Canadian EDs where the physician scheduled to work in the FT area has similar training to physicians working in the main ED area. While we believe many EDs have similar settings to ours, we note that the staffing of FT lines 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 into FT. Therefore, our results may not be directly applied to those hospitals, and it would be valuable to conduct analysis 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 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, Andrew, Paul Parks, Alan Drummond, Brian H Rowe, Howard J Ovens. 2013. Emergency department overcrowding and access block. *Canadian Journal of Emergency Medicine* **15**(6) 359–370.

Arya, Rajiv, Grant Wei, Jonathan V McCoy, Jody Crane, Pamela Ohman-Strickland, Robert M Eisenstein. 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, Robert J, Christian Terwiesch. 2016. Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Science* **63**(11) 3531–3551.

Bjørn, Pernille, Kjetil Rødje. 2008. Triage drift: a workplace study in a pediatric emergency department. *Computer Supported Cooperative Work (CSCW)* **17**(4) 395–419.

Burt, Catharine W, Linda F McCaig. 2006. Staffing, capacity, and ambulance diversion in emergency departments, United States, 2003-04. *Adv Data* **376** 1–23.

Chan, Carri W, Linda V Green, Suparerk Lekwijit, Lijian Lu, Gabriel Escobar. 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.

Chen, Jinsheng, Jing Dong, Pengyi Shi. 2020a. A survey on skill-based routing with applications to service operations management. *Queueing Systems* 1–30.

Chen, Wanyi, Benjamin Linthicum, Nilay Tanik Argon, Thomas Bohrmann, Kenneth Lopiano, Abhi Mehrotra, Debbie Travers, Serhan Ziya. 2020b. The effects of emergency department crowding on triage and hospital admission decisions. *The American Journal of Emergency Medicine* **38**(4) 774–779.

Chrusciel, Jan, Xavier Fontaine, Arnaud Devillard, Aurélien Cordonnier, Lukshe Kanagaratnam, David Laplanche, Stéphane Sanchez. 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).

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

Devkaran, Subashnie, Howard Parsons, Murray Van Dyke, Jonathan Drennan, Jaishen Rajah. 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, Yichuan, Eric Park, Mahesh Nagarajan, Eric Grafstein. 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, Jing, Pengyi Shi, Fanyin Zheng, Xin Jin. 2019. Off-service placement in inpatient ward network: Resource pooling versus service slowdown. *Working paper* .

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

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

Grafstein, Eric, Bernard Unger, Michael Bullard, Grant Innes, et al. 2003. Canadian emergency department information system (CEDIS) presenting complaint list (version 1.0). *Canadian Journal of Emergency Medicine* **5**(1) 27–34.

Greene, William H. 2018. *Econometric analysis*. Prentice Hall, Englewood Cliffs, NJ.

Guttmann, Astrid, Michael J Schull, Marian J Vermeulen, Therese A Stukel. 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**.

Heckman, James J. 1977. Dummy endogenous variables in a simultaneous equation system. Tech. rep., National Bureau of Economic Research.

Helm, Jonathan E, Mark P Van Oyen. 2014. Design and optimization methods for elective hospital admissions. *Operations Research* **62**(6) 1265–1282.

Ieraci, Susan, Erol Digiusto, Paul Sonntag, Linda Dann, Debra Fox. 2008. Streaming by case complexity: evaluation of a model for emergency department fast track. *Emergency Medicine Australasia* **20**(3) 241–249.

Ingolfsson, Armann, Elvira Akhmetshina, Susan Budge, Yongyue Li, Xudong Wu. 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, Diwas Singh, Stefan Scholtes, Christian Terwiesch. 2020. Empirical research in healthcare operations: past research, present understanding, and future opportunities. *Manufacturing & Service Operations Management* **22**(1) 73–83.

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

KC, Diwas Singh, Christian Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* **14**(1) 50–65.

Kelly, Anne-Maree, Michael Bryant, Lisa Cox, Damien Jolley. 2007. Improving emergency department efficiency by patient streaming to outcomes-based teams. *Australian Health Review* **31**(1) 16–21.

Kim, Song-Hee, Carri W Chan, Marcelo Olivares, Gabriel Escobar. 2015. ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science* **61**(1) 19–38.

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

Knapp, Laura Greene, Terry G Seaks. 1998. A hausman test for a dummy variable in probit. *Applied Economics Letters* **5**(5) 321–323.

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

Law, Averill M, W David Kelton. 2000. *Simulation modeling and analysis*. 3rd ed. McGraw-Hill New York.

Li, Wenhao, Zhankun Sun, L. Jeff Hong. 2021. Who is next: Patient prioritization under emergency department blocking. *Operations Research, Forthcoming* .

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

Liu, Shan W, Azita G Hamedani, David FM Brown, Brent Asplin, Carlos A Camargo Jr. 2013. Established and novel initiatives to reduce crowding in emergency departments. *Western Journal of Emergency Medicine* **14**(2) 85.

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

Lu, Lauren Xiaoyuan, Susan Feng Lu. 2018. Distance, quality, or relationship? Interhospital transfer of heart attack patients. *Production and Operations Management* **27**(12) 2251–2269.

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

Maddala, Gangadharrao S. 1986. *Limited-dependent and qualitative variables in econometrics*. 3, Cambridge university press.

Meng, Lesley, Robert J Batt, Christian Terwiesch. 2021. The impact of facility layout on service worker behavior: An empirical study of nurses in the emergency department. *Manufacturing & Service Operations Management* .

O'Brien, Debra, Aled Williams, Kerrianne Blondell, George A Jelinek. 2006. Impact of streaming "fast track" emergency department patients. *Australian Health Review* **30**(4) 525–532.

Peck, Jordan S, Sang-Gook Kim. 2010. Improving patient flow through axiomatic design of hospital emergency departments. *CIRP Journal of Manufacturing Science and Technology* **2**(4) 255–260.

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

Saghafian, Soroush, Wallace J Hopp, Mark P Van Oyen, Jeffrey S Desmond, Steven L Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* **60**(5) 1080–1097.

Saghafian, Soroush, Wallace J Hopp, Mark P Van Oyen, Jeffrey S Desmond, Steven L Kronick. 2014. Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management* **16**(3) 329–345.

Samiedaluie, Saied, Beste Kucukyazici, Vedat Verter, Dan Zhang. 2017. Managing patient admissions in a neurology ward. *Operations Research* **65**(3) 635–656.

Sanchez, Miquel, Alan J Smally, Robert J Grant, Lenworth M Jacobs. 2006. Effects of a fast-track area on emergency department performance. *The Journal of Emergency Medicine* **31**(1) 117–120.

Soltani, Mohamad, Robert Batt, Hessam Bavafa, Brian Patterson. 2020. Does what happens in the ED stay in the ED? the effects of emergency department physician workload on post-ED care use. *Working paper* .

Song, Hummy, Anita L Tucker, Ryan Graue, Sarah Moravick, Julius J Yang. 2020. Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science* **66**(9) 3825–3842.

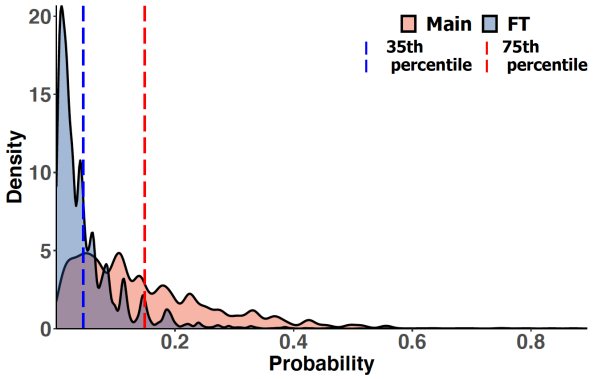
Song, Hummy, Anita L Tucker, Karen L Murrell. 2015. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* **61**(12) 3032–3053.

Trivedy, Chetan R, Matthew W Cooke. 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.

Wooldridge, Jeffrey M. 2012. *Introductory econometrics: A modern approach*. South-Western Cengage Learning.

Appendix A: Figures

Figure 6 **Patient Complexity Classification**



Appendix B: Tables

Table 7 **Routing statistics for patients of different complexity classes**

	High-complexity patients		Middle-complexity patients		Low-complexity patients	
	Count	Percentage	Count	Percentage	Count	Percentage
Main area	35,727	93.78%	50,099	81.79%	25,227	46.93%
Fast-track	2,370	6.22%	11,153	18.21%	28,529	53.07%

Table 8 **Summary statistics for patients of different complexity classes**

	High-complexity patients				Middle-complexity patients				Low-complexity patients			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Age (years)	57.22	18.48	0	101.8	40.70	17.63	0.0	104.2	31.36	15.19	0.0	95.7
Gender (Male%)	48.86	49.99	0	1	43.51	49.58	0	1	45.43	49.79	0	1
Triage level (%)												
CTAS 2	56.85	49.53	0	1	28.72	45.25	0	1	9.34	29.10	0	1
CTAS 3	37.55	48.43	0	1	52.23	49.95	0	1	34.00	47.37	0	1
CTAS 4	4.49	20.72	0	1	15.47	36.16	0	1	38.98	48.77	0	1
CTAS 5	1.10	10.44	0	1	3.58	18.58	0	1	17.68	38.15	0	1

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

Table 9 **Summary statistics for patient outcomes of different complexity classes**

	High-complexity patients		Middle-complexity patients		Low-complexity patients	
	Main	Fast-track	Main	Fast-track	Main	Fast-track
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
$Revisit_{48h}$ (%)	6.91 (25.37)	6.75 (25.10)	7.14 (25.76)	5.39 (22.58)	3.94 (19.46)	3.06 (17.22)
$Revisit_{72h}$ (%)	8.51 (27.91)	8.31 (27.61)	8.41 (27.75)	6.60 (24.83)	4.90 (21.59)	3.82 (19.17)
LOS (hours)	5.08 (3.25)	3.18 (2.02)	4.03 (2.58)	3.03 (1.90)	3.21 (2.12)	2.59 (1.66)

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

Table 10 Complete estimation results based on visit records of all patients. The first column shows the results of Model (1); other columns show the estimation results on patient outcome variables (with IV).

	Probit	48-hour revisit		72-hour revisit		Length of stay	
	FT	*FT*	$Revisit_{48h}$	*FT*	$Revisit_{72h}$	*FT*	log(*LOS*)
MEAdjBusyRatio	0.070***	0.070***		0.070***		0.085***	
	(0.005)	(0.009)		(0.009)		(0.011)	
Age group (Base=0–25 years)							
25–40 years	-0.042***	-0.042***	0.131***	-0.042***	0.146***	-0.040***	0.079***
	(0.013)	(0.013)	(0.018)	(0.013)	(0.017)	(0.013)	(0.005)
40–55 years	-0.030**	-0.030**	0.084***	-0.030**	0.096***	-0.026*	0.157***
	(0.013)	(0.015)	(0.019)	(0.015)	0.017)	(0.015)	(0.007)
55–70 years	-0.046***	-0.046**	0.081***	-0.046**	0.109***	-0.040**	0.195***
	(0.015)	(0.018)	(0.021)	(0.018)	(0.020)	(0.018)	(0.007)
> 70 years	-0.158***	-0.158***	0.136***	-0.158***	0.167***	-0.148***	0.282***
	(0.018)	(0.020)	(0.026)	(0.020)	(0.024)	(0.020)	(0.010)
Triage level (Base=CTAS 2)							
CTAS 3	0.329***	0.329***	-0.060***	0.329***	-0.076***	0.318***	-0.056***
	(0.012)	(0.018)	(0.014)	(0.018)	(0.012)	(0.019)	(0.006)
CTAS 4	0.534***	0.534***	-0.186***	0.534***	-0.203***	0.519***	-0.144***
	(0.014)	(0.030)	(0.019)	(0.030)	(0.018)	(0.030)	(0.010)
CTAS 5	0.472***	0.472***	-0.248***	0.472***	-0.247***	0.457***	-0.219***
	(0.018)	(0.039)	(0.023)	(0.039)	(0.022)	(0.040)	(0.012)
Gender (Male=1)	0.160***	0.160***	-0.055***	0.160***	-0.036***	0.159***	-0.022***
	(0.009)	(0.011)	(0.013)	(0.011)	(0.012)	(0.011)	(0.004)
Hospital (Base=ED A)							
ED B	-0.534***	-0.534***	0.161***	-0.534***	0.141***	-0.539***	0.127***
	(0.012)	(0.030)	(0.018)	(0.030)	(0.018)	(0.030)	(0.018)
ED C	-0.117***	-0.117	0.136***	-0.117	0.118***	-0.120	-0.097***
	(0.012)	(0.133)	(0.021)	(0.133)	(0.019)	(0.132)	(0.026)
Triage Time	-0.119***	-0.119***	0.006	-0.119***	0.008	-0.120***	0.035***
	(0.005)	(0.008)	(0.006)	(0.008)	(0.006)	(0.008)	(0.003)
AvgOccTreated			0.049***		0.058***		0.332***
			(0.008)		(0.007)		(0.008)
WaitTime			0.048***		0.045***		
			(0.007)		(0.006)		
FT			0.019		-0.011		-0.391***
			(0.076)		(0.075)		(0.062)
N	153,105	153,105		153,105		153,105	

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

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

Table 11 Complete 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*)
MEAdjBusyRatio	0.059***		0.058***		0.059***	
	(0.012)		(0.012)		(0.012)	
Age group (Base=0–25 years)						
25–40 years	-0.048	0.147**	-0.044	0.167***	-0.042	0.117***
	(0.112)	(0.066)	(0.111)	(0.060)	(0.112)	(0.025)
40–55 years	-0.158	0.036	-0.155	0.048	-0.155	0.223***
	(0.114)	(0.065)	(0.114)	(0.061)	(0.114)	(0.025)
55–70 years	-0.294***	-0.065	-0.292**	-0.033	-0.292**	0.272***
	(0.114)	(0.066)	(0.113)	(0.061)	(0.114)	(0.028)
> 70 years	-0.338***	0.013	-0.337***	0.043	-0.340***	0.385***
	(0.121)	(0.073)	(0.120)	(0.069)	(0.121)	(0.029)
Triage level (Base=CTAS 2)						
CTAS 3	0.292***	-0.088***	0.293***	-0.096***	0.293***	-0.113***
	(0.032)	(0.026)	(0.032)	(0.023)	(0.032)	(0.010)
CTAS 4	0.619***	-0.183***	0.618***	-0.194***	0.618***	-0.300***
	(0.056)	(0.063)	(0.056)	(0.058)	(0.055)	(0.022)
CTAS 5	0.839***	-0.334***	0.840***	-0.259**	0.843***	-0.382***
	(0.090)	(0.122)	(0.090)	(0.113)	(0.090)	(0.045)
Gender (Male=1)	0.116***	-0.015	0.116***	0.014	0.115***	-0.004
	(0.025)	(0.022)	(0.025)	(0.023)	(0.025)	(0.007)
Hospital (Base=ED A)						
ED B	-0.062	0.108***	-0.061	0.104***	-0.059	0.090***
	(0.041)	(0.031)	(0.041)	(0.028)	(0.041)	(0.018)
ED C	-0.108	0.122***	-0.107	0.111***	-0.107	-0.106***
	(0.132)	(0.031)	(0.132)	(0.028)	(0.133)	(0.027)
Triage Time	-0.150***	0.009	-0.150***	0.009	-0.150***	0.050***
	(0.014)	(0.011)	(0.014)	(0.011)	(0.015)	(0.004)
AvgOccTreated		-0.016		0.000		0.325***
		(0.013)		(0.013)		(0.010)
WaitTime		0.022**		0.023**		
		(0.010)		(0.010)		
FT		0.439**		0.303		-0.384***
		(0.190)		(0.197)		(0.105)
N	38,097		38,097		38,097	

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

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

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

	48-hour revisit		72-hour revisit		Length of stay	
	FT	$Revisit_{48h}$	*FT*	$Revisit_{72h}$	*FT*	log(*LOS*)
MEAdjBusyRatio	0.079***		0.079***		0.087***	
	(0.011)		(0.011)		(0.011)	
Age group (Base=0-25 years)						
25-40 years	0.078***	0.023	0.078***	0.046**	0.076***	0.074***
	(0.029)	(0.025)	(0.029)	(0.023)	(0.029)	(0.011)
40-55 years	0.043	-0.088***	0.043	-0.064**	0.042	0.150***
	(0.032)	(0.031)	(0.032)	(0.029)	(0.033)	(0.014)
55-70 years	0.033	-0.049	0.034	-0.005	0.035	0.201***
	(0.042)	(0.039)	(0.042)	(0.038)	(0.042)	(0.017)
> 70 years	-0.064	-0.014	-0.063	0.018	-0.063	0.274***
	(0.052)	(0.060)	(0.052)	(0.057)	(0.052)	(0.024)
Triage level (Base=CTAS 2)						
CTAS 3	0.434***	-0.018	0.434***	-0.051**	0.425***	-0.041***
	(0.026)	(0.023)	(0.026)	(0.023)	(0.027)	(0.009)
CTAS 4	0.738***	-0.087**	0.738***	-0.124***	0.728***	-0.142***
	(0.042)	(0.035)	(0.042)	(0.032)	(0.042)	(0.013)
CTAS 5	0.869***	-0.072	0.869***	-0.109**	0.860***	-0.280***
	(0.051)	(0.053)	(0.051)	(0.052)	(0.051)	(0.024)
Gender (Male=1)	0.157***	-0.104***	0.157***	-0.081***	0.152***	-0.041***
	(0.019)	(0.018)	(0.019)	(0.016)	(0.018)	(0.007)
Hospital (Base=ED A)						
ED B	-0.289***	0.214***	-0.289***	0.200***	-0.291***	0.122***
	(0.034)	(0.022)	(0.034)	(0.023)	(0.035)	(0.018)
ED C	-0.175	0.179***	-0.175	0.164***	-0.174	-0.097***
	(0.135)	(0.026)	(0.135)	(0.024)	(0.134)	(0.027)
Triage Time	-0.136***	-0.002	-0.136***	0.003	-0.140***	0.033***
	(0.010)	(0.008)	(0.010)	(0.007)	(0.010)	(0.004)
AvgOccTreated		0.021*		0.037***		0.344***
		(0.011)		(0.009)		(0.009)
WaitTime		0.069***		0.062***		
		(0.009)		(0.008)		
FT		0.219**		0.183**		-0.423***
		(0.089)		(0.081)		(0.047)
N	61,252		61,252		61,252	

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

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

Table 13 Complete 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*)
MEAdjBusyRatio	0.069***		0.069***		0.128***	
	(0.011)		(0.011)		(0.014)	
Age group (Base=0-25 years)						
25-40 years	-0.055***	0.175***	-0.055***	0.189***	-0.057***	0.052***
	(0.015)	(0.028)	(0.015)	(0.027)	(0.015)	(0.008)
40-55 years	-0.011	0.253***	-0.011	0.271***	-0.012	0.115***
	(0.019)	(0.033)	(0.019)	(0.031)	(0.019)	(0.011)
55-70 years	0.035	0.293***	0.036	0.318***	0.031	0.123***
	(0.028)	(0.050)	(0.028)	(0.042)	(0.028)	(0.016)
> 70 years	0.054	0.369***	0.055	0.504***	0.035	0.149***
	(0.083)	(0.105)	(0.083)	(0.097)	(0.080)	(0.037)
Triage level (Base=CTAS 2)						
CTAS 3	0.097***	-0.06	0.096***	-0.071	0.094***	0.066***
	(0.029)	(0.054)	(0.029)	(0.044)	(0.029)	(0.015)
CTAS 4	0.220***	-0.192***	0.220***	-0.216***	0.212***	0.017
	(0.037)	(0.055)	(0.037)	(0.047)	(0.037)	(0.018)
CTAS 5	0.110**	-0.240***	0.110**	-0.250***	0.109**	-0.045***
	(0.044)	(0.061)	(0.044)	(0.054)	(0.044)	(0.016)
Gender (Male=1)	0.211***	-0.023	0.211***	-0.024	0.204***	-0.008
	(0.017)	(0.025)	(0.017)	(0.024)	(0.017)	(0.008)
Hospital (Base=ED A)						
ED B	-0.799***	0.094**	-0.799***	0.085**	-0.785***	0.098***
	(0.032)	(0.043)	(0.032)	(0.038)	(0.032)	(0.025)
ED C	-0.101	0.090**	-0.101	0.065*	-0.115	-0.088**
	(0.140)	(0.040)	(0.140)	(0.036)	(0.136)	(0.035)
Triage time	-0.108***	0.018	-0.108***	0.019*	-0.105***	0.020***
	(0.012)	(0.013)	(0.012)	(0.011)	(0.011)	(0.005)
AvgOccTreated		0.135***		0.125***		0.322***
		(0.013)		(0.012)		(0.008)
WaitTime		0.031**		0.034***		
		(0.013)		(0.011)		
FT		-0.193		-0.063		-0.669***
		(0.155)		(0.137)		(0.060)
N	53,756		53,756		53,756	

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

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