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

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

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

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

Problem definition: This work aims to examine the role of emergency department (ED) operational status related to congestion in fast-track (FT) routing decisions and the subsequent effects on patient outcomes. **Methodology/results:** In this paper, we utilize a two-year dataset from two hospital EDs in Alberta, Canada, and adopt an instrumental variable (IV) approach to examine the effects of FT routing decisions on patient outcomes. Based on the empirical findings, we utilize a data-calibrated simulation to compare the performance of different routing policies. First, our study reveals that FT routing decisions are not purely clinical-driven, and ED operational status related to congestion is also associated with FT routing decisions. Second, we find that being routed to FT can improve emergency care efficiency by reducing the average ED length of stay (LOS). However, this efficiency improvement comes at the cost of potential quality decline. In particular, being routed to the FT leads to an 8.2% increase in the 48-hour revisit rate for the high-complexity group and a 2.3% increase for the medium-complexity group. Third, we delve into the mechanisms behind observed patient outcomes and find that physicians in the FT area may prioritize expediting patient flow by simplifying patient diagnosis and treatment procedures. Consequently, the quality of care may be compromised for high- and medium-complexity patients. Finally, our simulation findings highlight the importance of selecting the "right" patients to be routed to the FT unit. To this end, the complexity-based classification method and dynamic routing policies emerge as promising avenues. **Managerial implications:** Our findings call for immediate attention from healthcare practitioners to carefully balance the trade-off between emergency care efficiency and quality, emphasizing the necessity of selecting the "right" patients for routing.

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

1. Introduction

Emergency department (ED) congestion has been observed in many hospitals across the world and poses critical challenges to both healthcare practitioners and policy makers. According to the National Center for Health Statistics, 40%–50% of US hospitals have experienced ED congestion (Burt and McCaig 2006). As a result, patients have to spend hours in the waiting area, leading to an increased risk of cross-infection, mortality, and patient readmission (Guttmann et al. 2011). Hence, a crowded ED is more than a nuisance; it is a threat to both individual patients and overall public health (Maa 2011). Many strategies have been proposed to regulate patient flow and reduce ED congestion. Among these, fast-track (FT) has been highlighted by

the American College of Emergency Physicians (ACEP) as a high-impact initiative (Liu et al. 2013). In particular, FT is a separate ED area that provides dedicated pathways aimed toward fast care delivery and rapid discharge for patients with less urgent conditions, which becomes more prevalent in recent years and has been implemented by nearly 80% of academic EDs in the US (Liu et al. 2013).

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

Moreover, so far, hospitals have not yet established consistent guidelines for determining which patients should be routed to the FT, which might be due to the lack of a more comprehensive understanding of how FT routing decisions impact patient outcomes as we discussed earlier. Upon arriving at an ED, a patient who does not have life-threatening conditions is first triaged by the nursing staff, who (i) assigns the patient a triage score that indicates the urgency level of the patient's care needs and (ii) routes the patient to either the main ED area, where most patients are treated, or to the FT area. Standard protocols have been established for assigning triage scores, such as the Canadian Triage and Acuity Scale (CTAS), the most commonly used triage protocol in Canada, and the Emergency Severity Index (ESI), the algorithm commonly adopted in the US. Both protocols are five-point scoring systems (1 to 5) with smaller numbers indicating higher levels of urgency. However, neither of these protocols specifies the type of patients that should be routed to the FT. Hence, EDs currently make FT routing decisions at their own discretion. Some EDs adopt a flexible routing policy, under which triage nurses make routing decisions based on both triage scores and other patient and

ED factors (which is the practice in our study hospitals). On the other hand, many EDs simply implement triage-score-based routing policies. Specifically, it has been observed in both American (Peck and Kim 2010, Arya et al. 2013, Song et al. 2015) and Canadian EDs (Ding et al. 2019, Al Darrab et al. 2006) that all (and only) patients of triage levels 4 and 5 are routed to the FT. Such a policy is simple and easy to implement, but it is rigid and may lead to a mismatch between demand and supply in different treatment areas. Meanwhile, flexible policies could route patients with similar clinical conditions to different ED areas under different congestion conditions, of which the consequence is not clear. Therefore, it is inherently important to establish managerial implications to guide FT routing decisions (Peck and Kim 2010), which is the third goal of this study.

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

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

Second, we find that being routed to FT can improve emergency care efficiency by reducing the average ED LOS and LWBS rates. This finding supports the merit of establishing the FT area, that is, to provide fast care delivery and improve emergency care efficiency. However, we also find that being routed to the FT can lead to an 8.2% increase in the 48-hour revisit rate for the high-complexity group and a 2.3% increase for the medium-complexity group. These findings uncover an important trade-off between care efficiency and quality assurance.

Third, we delve into the mechanisms behind observed patient outcomes. Our analysis reveals that routing patients to the FT area leads to reduced wait times for all patient groups and a decrease in treatment time specifically for medium-complexity patients. Additionally, being routed to the FT area reduces the number of lab tests, prescribed medications, and CT scans for patients across all complexity levels. These findings suggest that FT physicians tend to simplify the diagnosis and treatment process for patients in the FT area. Consequently, quality of care may be compromised for high- and medium-complexity patients in the FT area. In contrast, low-complexity patients are easier to diagnose, minimizing the impact of simplified diagnostic procedures in the FT area on their care quality.

Finally, drawing from our empirical results, we employ a data-calibrated simulation to assess the performance of different routing policies, ranging from static policies based on our complexity-based classification method or solely on triage scores for patient selection, to dynamic policies that balance the trade-off between emergency care efficiency and quality. Our simulation results emphasize the importance of selecting the "right" patients to be routed to the FT unit. In this regard, the complexity-based classification method and dynamic routing policies emerge as promising avenues.

2. Literature Review

In recent years, studies on healthcare worker behaviors, especially in congested systems, have attracted growing attention from the operations management (OM) community (KC et al. 2020a). Existing studies have shown that healthcare workers respond to system congestion and heavy workload by varying their behavior and rationing decisions, which leads to, among others, accelerated service (KC and Terwiesch 2009, Long and Mathews 2018), compromised patient safety (Kuntz et al. 2015), early task initiation (Batt and Terwiesch 2016), higher referral rates (Freeman et al. 2017), increased post-ED care utilization (Soltani et al. 2022), biased admission decisions (Kim et al. 2020), nurse absenteeism (Green et al. 2013), easier task selection (KC et al. 2020b), elevated incidence of unnecessary hospital admissions (Freeman et al. 2021), and patient undercoding (Powell et al. 2012). In related ED studies, Batt and Terwiesch (2015) explore the behavior of queue abandonment in EDs, Sun et al. (2020) examine the role of telemedicine in alleviating ED congestion, Li et al. (2023) study the impact of ED blocking on patient prioritization, and Feizi et al. (2023) investigate the effects of batch admissions on patients' boarding times and the productivity of ED physicians. Our study contributes to this stream of literature by uncovering a positive relationship between ED congestion and FT routing decisions. Therefore, despite that the purpose of FT is to provide fast care delivery to patients with less urgent conditions, FT routing decisions are not purely clinical-driven, and operational factors related to ED congestion are also critical in making the decision.

Consequently, our work is closely related to studies that empirically examine the impact of routing decisions in healthcare settings (e.g., Kim et al. 2015, Chan et al. 2018, and Song et al. 2020). In particular, Kim et al. (2015) investigate the impact of the routing decisions (i.e., admission or denied admission) to a hospital's intensive care unit (ICU) on patient outcomes. By quantifying the cost of denied ICU admission, they provide a simulation framework to compare various admission strategies. Chan et al. (2018) empirically estimate the costs and benefits associated with routing patients to the general wards, ICUs, and step-down units. Song et al. (2020) study the off-service placement in hospitals, i.e., routing a patient to hospital beds designated for a different service due to capacity constraints on the unit designed for this patient's service needs. Routing decisions in other healthcare settings have also been investigated. For example, Lu and Lu (2018) probe the inter-hospital routing of heart attack patients, and Webb and Mills (2019) discuss how to increase the pre-hospital triage adoption to route patients to appropriate care providers before transport to

the ED. Interested readers can refer to Section 3.3 in KC et al. (2020a) for a review of studies on patient routing decisions in healthcare systems. Built upon earlier works, this paper contributes to the literature by studying triage nurses' FT routing decisions and examining their impact on emergency care efficiency and quality assurance.

Next, our work also relates to the literature on the speed-quality tradeoff, a crucial aspect highlighted by our empirical findings and utilized in our simulations. This trade-off has been explored in a number of contexts. Hopp et al. (2007) investigate systems featuring discretionary task completion involving a speed-quality tradeoff and compare them to systems where task completion is non-discretionary. Drawing from empirical evidence, KC and Terwiesch (2009) show that increased service speed, resulting from overload, adversely affects the quality of care. Anand et al. (2011) subsequently investigate how service providers manage the trade-off between service quality and duration, with a particular focus on its impact on service rates, pricing decisions, and customer behavior in customer-intensive services. Li et al. (2016) explore a similar speed-quality tradeoff in scenarios where customers exhibit bounded rationality. Our work builds upon this literature to examine the speed-quality tradeoff faced by triage nurses, and subsequently, the consequences of the FT routing decision on patient outcomes.

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

underlying mechanisms behind the observed outcomes, and (iv) employing a data-calibrated simulation to evaluate different routing policies and derive managerial implications that might help guide FT routing decisions.

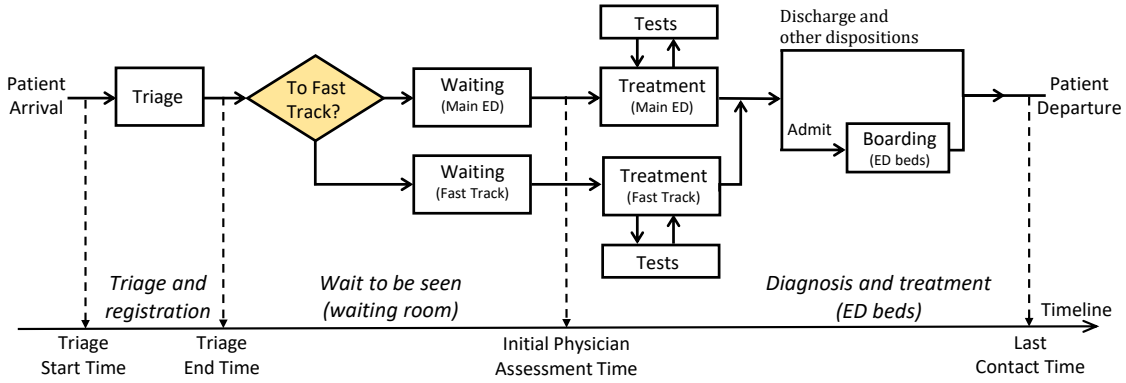
3. Study Setting and Data

This section describes our study setting and the data used to conduct our analyses. Section 3.1 describes the setting, Section 3.2 presents the details of our data, and Section 3.3 discusses the choice of key variables used in the analyses. Furthermore, the hypotheses development is presented in Appendix B.

3.1. Patient Flow Process

We first describe our setting with details on the patient flow process in our collaborator hospitals. The two EDs adopt a similar patient flow process, depicted in Figure 1. Note that our description is based on EDs in the Calgary zone of 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 ED arrival, patients are first assigned a triage score, following the CTAS protocol. The timestamps at the start and end of the triage process are referred to as triage start and end times, respectively. The time duration between triage start and end is referred to as the *triage time*, during which triage nurses assign triage scores and route patients to either the main ED or the FT area. These two are separate treatment areas with dedicated care teams; however, they share the same pool of attending physicians, send diagnostic orders to the same test center, and have similar equipment except that the FT area has fewer beds and physicians. A physician may work in the main area in one shift and the FT unit at a different time; however, each physician is dedicated to one area during any specific shift. In other words, the FT area is capable of treating patients of any urgency level except for triage level 1 patients who need to be sent to the resuscitation room. During the study period, the FT area operates 14 hours every day in the two EDs from 10 am to midnight.

After triage, patients remain in the waiting room until they are signed up by physicians, which marks the start of the *initial physician assessment*. The period between triage end and initial physician assessment is referred to as the *patient waiting time*. During this period, patients may leave the hospital without receiving treatment. This behavior is referred to as patient *left without being seen* (LWBS) in the patient records. 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. Finally, the period from the triage start time to the last contact time is referred to as the ED *length of stay* (LOS).

3.2. Data Description

Our data contain patient visit records from the EDs of two urban hospitals in Alberta, Canada, from August 2013 to July 2015. The average daily patient arrivals to the two EDs are 178.9 and 183.7, respectively; the average daily traffic to the FT areas of the two EDs are 33.4 and 41.7, respectively. There are two 7-hour shifts scheduled in the FT area and 13 shifts in the main area at both EDs during most of our study period. Hence, the two hospitals are comparable in their patient volume and service capacity. Furthermore, since the Canadian healthcare system is centralized, the two hospitals are managed by the same health authority and have adopted the same health information system, which enables the combination of the data from the two EDs. Our dataset before cleaning includes a total of 264,551 visits from 169,752 patients. Note that a patient may have visited the EDs more than once during the study period. Each observation in our data includes patient demographics (e.g., age and gender) and the details of their ED visits (e.g., chief complaint, triage score, and attending physician ID).

We now discuss our data preparation process for empirical analysis. To start with, we exclude patient visits that fall outside the FT operational hours (i.e., from midnight to 10 am), which leaves us with a total of 203,974 observations (22.9% removed) from 139,830 patients. Subsequently, we exclude observations with an LOS exceeding 48 hours, as such extreme cases could introduce bias into our results (Song et al. 2015). Note that only 299 observations in our dataset have LOS greater than 48 hours. Next, to avoid potential issues associated with censored estimates (Kim et al. 2015, Song et al. 2020, Chan et al. 2018), we remove observations from the first and last weeks of our study period, leading to the elimination of 3,595 observations (1.4% of the original data). Furthermore, we exclude patients arriving at the ED through ambulance, as these individuals typically have urgent healthcare needs requiring immediate attention from physicians. Additionally, we remove patients with dispositions labeled as "left against medical advice" and "transferred" because those patients did not receive care from their visits. However, we keep data on patients labeled as "left without being seen" to facilitate analysis of the ED LWBS outcome. These two steps leave us with 147,369 observations (19.9% of the original data removed) from 108,685 patients. Moreover, we exclude patients of triage level 1 (0.3% of the original data), as their conditions are usually very urgent, requiring immediate care. The dataset after this step (denoted as Dataset I) comprises 146,481 observations

from 108,068 patients and will be used for evaluating the impact of FT routing decisions on the ED LWBS outcome. Later in Section 5.4, we also conduct a robustness check including patients with triage score 1 and show that our main results remain consistent. Following that, we exclude patients with the disposition "left without being seen," denoted as Dataset II, which includes 142,683 observations from 105,894 patients. Dataset II is used for the estimation of the LOS outcome and patient classification in Section 4.4.

Finally, to analyze the 48-hour revisit outcome, we further exclude the admitted patients from Dataset II because the likelihood of admitted patients revisiting the ED within 48 hours is expected to be substantially lower as admitted patients receive continuing care in the inpatient unit. Consequently, combining discharged and admitted patients in the analysis may underestimate the impact of FT routing on 48-hour revisits. Additionally, our dataset does not contain hospital discharge information for admitted patients, which is essential for calculating 48-hour revisits. As a result, we exclude admitted patients and this new dataset is referred to as Dataset III which is our main dataset, consisting of 123,655 observations from 94,448 patients. We employ Dataset III to assess the patient's 48-hour revisit outcome. Furthermore, Dataset III is also used for the mechanism discussion, involving outcome variables such as treatment time, wait time, lab tests, medications, CT scans, and X-ray tests, as elaborated in Section 5.3. We have also used Dataset II to conduct the same mechanism analyses and obtained consistent results (see Tables 14 and 15 in the supplementary document).

3.3. Choice of Variables

This section presents the choice of key variables used in our empirical analyses; see Table 1 for the summary statistics of our main dataset (Dataset III). Additionally, Tables 12 and 13 in the supplementary document present the summary statistics of Dataset I and II.

3.3.1. Dependent Variables We consider three outcome measures: the 48-hour revisit rate, patient LWBS rate, and patient LOS, denoted by $Revisit_{48h}$, *LWBS*, and *LOS*, respectively. The variable $Revisit_{48h}$ equals 1 if the patient visited one of the two EDs within 48 hours after being discharged and 0 otherwise. The 48-hour revisit rates are widely used in the healthcare literature to measure the quality of emergency care (e.g., Ieraci et al. 2008, Trivedy and Cooke 2015, and Song et al. 2015). Later, we also consider a robustness check with the 72-hour revisit rate in Section 5.4 and show consistent results. Furthermore, the variable *LWBS* is assigned a value of 1 if the patient leaves the ED without being seen, and 0 otherwise. Finally, as mentioned in Section 3.1, the variable *LOS* calculates the duration from the start of the triage process to the last patient contact time.

3.3.2. Independent Variables Next, we describe the independent variables in our estimation. The key variable of interest in our study is the FT routing decision for patient i, denoted by FT_i, which equals 1 if patient i is routed to the FT area and 0 otherwise. It is worth noting that 0.3% of patients in our dataset were first routed to FT but received care in the main ED area. In our empirical analyses, we considered

Table 1 Summary statistics of key variables for Dataset III.

	Main Area				Fast-Track Area			
Variables	Mean	SD	Min	Max	Mean	SD	Min	Max
Patient Outcomes								
Revisit$_{48h}$ (%)	6.63	24.88	0	100	3.89	19.35	0	100
LOS (in hours)	4.49	2.90	0.15	40.23	2.80	1.70	0.20	25.47
Operational Characteristics								
EDCongestion	0.78	0.14	0.13	1.31	0.79	0.14	0.14	1.30
MainCongestion	0.78	0.14	0.13	1.32	0.78	0.15	0.15	1.32
FTCongestion	0.60	0.27	0.00	2.00	0.59	0.27	0.00	1.91
AvgOccTreated	0.58	0.21	0.00	1.30	0.48	0.23	0.00	1.25
WaitTime (in hours)	1.56	1.25	0.00	17.64	1.34	0.95	0.00	9.97
TreatTime (in hours)	0.46	0.31	0.02	1.45	0.39	0.28	0.02	1.44
TriageTime (in minutes)	4.50	1.88	0.62	43.43	3.82	1.62	0.70	49.22
Diagnostic Tests								
X-rayTests	0.36	0.66	0.00	11.00	0.54	0.82	0.00	7.00
CTScans	0.16	0.42	0.00	8.00	0.04	0.21	0.00	7.00
LabTests	3.82	3.41	0.00	52.00	0.57	1.72	0.00	38.00
Medications	1.92	2.44	0.00	32.00	0.62	1.15	0.00	19.00
Instrumental Variable								
MEBusyRatio	1.00	0.06	0.64	1.15	1.00	0.06	0.63	1.15
Physician Characteristics								
Workload	3.59	2.36	0.00	18.00	2.69	1.80	0.00	14.00
Patient Characteristics								
Gender (Male %)	40.83	49.15	0	100	55.43	49.70	0	100
Age group (%)								
0 to 25 years	18.02	38.44	0	100	21.51	41.09	0	100
25 to 40 years	30.80	46.17	0	100	29.64	45.67	0	100
40 to 55 years	22.18	41.54	0	100	22.14	41.52	0	100
55 to 70 years	17.03	37.59	0	100	17.12	37.67	0	100
Over 70 years	11.97	32.46	0	100	9.58	29.43	0	100
Triage score (%)								
CTAS 2	35.54	47.86	0	100	15.13	35.84	0	100
CTAS 3	44.68	49.72	0	100	37.26	48.35	0	100
CTAS 4	15.06	35.76	0	100	33.12	47.06	0	100
CTAS 5	4.73	21.22	0	100	14.49	35.20	0	100
N	85,091				38,564			

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

these patients as being routed to the main area. In what follows, we discuss a set of control variables on the system-, physician-, and patient-level operational metrics and patient characteristics.

We start with the system-level operational metric: the area occupancy level during patient i's treatment period, denoted by $AvgOccTreated_i$. Following similar ideas in Kim et al. (2015), Chan et al. (2018), and Song et al. (2020), we define the area occupancy level as the time-averaged number of patients receiving treatment in the ED (excluding the focal patient i) during patient i's treatment.[1] Next, we introduce the

[1] To explain the details of our calculation, suppose patient i receives care in the ED from 3 to 5 pm, during which 4 other patients also receive care in the main area or FT (not necessarily for 2 hours). Assume the treatment of 2 of them overlap with patient i for 2 hours, 1 of them overlaps for 1 hour (say from 1 to 4 pm), and the last one also overlaps for 1 hour (say from 4 to 5 pm). Then, $AvgOccTreated_i = (2*2+1+1)/2 = 3$.

physician workload, denoted by $Workload_i$, which is defined as the number of patients who are under the care of patient i's attending physician and have not yet received a disposition decision at the time when patient i is assigned.[2] We measure physician workload at the time of the focal patient's assignment, as this marks the crucial moment when the patient's treatment plan is established. It is important to note that all the previously dropped observations have been included in the calculation of physician workload and area occupancy level. We control $AvgOccTreated_i$ and $Workload_i$ in our estimation to block the indirect impact of our proposed IV on patient outcomes through channels other than the FT routing decision. For example, earlier work has shown that area occupancy level might adversely affect patient outcomes (Kuntz et al. 2015, Long and Mathews 2018) and physician workload could lead to physician behavioral changes that might negatively affect patient outcomes (KC and Terwiesch 2009). See detailed discussions in Section 4.2. Next, we include two patient-level operational characteristics: waiting time ($WaitTime_i$) and triage time ($TriageTime_i$). The waiting time is the period from the end of triage to the time when patient i is picked up by a physician. The triage time is the period from the triage start to end.

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

4. Econometric Model

This section describes the econometric model and identification strategy used in our paper.

4.1. Baseline Econometric Model

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

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

[2] Suppose patient i is assigned to a physician at 11 am. The physician is responsible for the care of three other patients that have not been discharged or admitted at 11 am. Then, the physician workload at the time when the focal patient is assigned is 3.

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

One may estimate Equation (1) and then interpret the estimated parameter $\tilde{\gamma}$ as the impact of being routed to FT on patient outcomes. However, such an approach ignores that the FT routing decisions may be endogenous due to factors tied to patients' clinical conditions that triage nurses considered during their decision-making process but are unobservable in our data, such as the patient's mental state and pain level. These unobserved clinical 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). To elaborate, these unobservables are likely to exhibit a negative correlation with fast-track routing decisions and adverse patient outcomes, introducing a positive bias into the estimation of the causal effect of fast-track routing. In essence, these biases could potentially lead to an underestimation of the potential adverse effects of FT routing. Next, we discuss how we address this issue in our estimation.

4.2. Instrumental Variables

To address the endogeneity issue raised in Section 4.1, we adopt an IV approach. A valid IV should satisfy two requirements: (i) inclusion condition—IVs should be correlated with the endogenous variable; and (ii) exclusion condition—IVs cannot directly affect the dependent variable except through the endogenous variable. Following the empirical healthcare literature, we consider IVs related to operational factors of the ED; see, e.g., Kim et al. (2015), Chan et al. (2018), and Song et al. (2020). Specifically, following closely Kim et al. (2015) and Song et al. (2020), we propose an ED congestion-related IV: the relative congestion level between the main area and the entire ED at patient i's triage start time, denoted by $MEBusyRatio_i$. To compute this variable, we first measure congestion levels in the main area, the FT area, and the entire ED, denoted by *MainCongestion*, *FTCongestion*, and *EDCongestion*, respectively. This area congestion measure is calculated as the area workload divided by the area capacity. In particular, the area workload is computed as the total number of patients waiting and being treated in this area. Next, the area capacity is defined as the 95th percentile of the distribution of the area workload, where we use the 95th percentile instead of the maximum to avoid observations under extreme situations (Kim et al. 2015). Note that we compute this capacity measure for each hospital separately. In general, the congestion measure here captures the extent to which the area workload takes up to its service capacity. Based on these congestion measures, we can then compute our proposed IV on the relative congestion level between the main area and the entire ED at patient i's triage time. Later in the robustness check in Section 5.4, we also consider an alternative congestion measure that adjusts the number of physicians on duty.

Next, we discuss the validity of this IV. We start with the inclusion condition. It has been shown in the earlier work that healthcare admission controllers take into account hospital congestion or utilization when

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

Moving forward, we discuss the exclusion condition, i.e., the busyness ratio $MEBusyRatio_i$ affects patient outcomes only through the FT routing decision. To start with, we note that our IV (the busyness ratio $MEBusyRatio_i$) measures the relative congestion condition at the starting time of triage. Therefore, ideally, our proposed IV only affects the FT routing decision but not patient outcomes that occurred after the treatment. However, one may argue that the relative congestion condition at the time of triage might be correlated with the overall ED congestion during the patient's treatment, thus affecting patient outcomes. Although our proposed IV is a relative measure (i.e., the relative congestion level between the main area and the entire ED), we still cannot fully rule out the possibility that this relative congestion measure might be correlated with the area congestion during the patient's treatment process. Therefore, we introduce the following two important control variables in our estimation to block the indirect impact of our proposed IV on patient outcomes through channels other than the FT routing decision. First, following Kim et al. (2015), we control for the area occupancy level ($AvgOccTreated_i$) during the focal patient's diagnosis and treatment period, which allows us to separate the impact of relative congestion on the FT routing decision from its direct impact on patient outcome. This step is important as earlier work (e.g., Kuntz et al. 2015, Long and Mathews 2018) has shown that area occupancy level might adversely affect patient outcomes. Second, similar to the control on the area occupancy level, we also control for the workload of patient i's attending physician ($Workload_i$) at the time when patient i was assigned to this physician. Following a similar logic, this control variable allows us to separate the impact of relative area congestion on the FT routing decision from its impact on physician behaviors. This is another important step because earlier work (e.g., KC and Terwiesch 2009) has shown that increased physician workload leads to their behavioral changes that might negatively affect patient outcomes. As a result, conditional on the occupancy level and the individual physician workload, the busyness ratio $MEBusyRatio_i$ can only affect patient outcomes through the FT routing decision. It is worth noting that controlling for $AvgOccTreated_i$ or $Workload_i$ helps alleviate the concern if the IV ($MEBusyRatio_i$) and $AvgOccTreated_i$ or $Workload_i$ are not highly correlated. Thus, we

check the correlations and find that the variables $MEBusyRatio_i$ and $AvgOccTreated_i$ (or $Workload_i$) do not exhibit a strong correlation, with a low correlation coefficient of only 0.052 (or -0.032).

To further validate the exogeneity of our proposed IV, we perform different analyses by regressing our IV and various ED utilization metrics against observable patient severity measures, see Table 6 in Appendix A. We do not find any statistically significant association between our proposed IV ($MEBusyRatio_i$) and the observable patient severity factors. This further supports the exclusion condition and the validity of the IV. Additionally, it is worth noting that other congestion measures (i.e., *EDCongestion* and *MainCongestion*) exhibit statistically significant correlations with certain observable patient severity measures, indicating the need for constructing the relative congestion measure.

Finally, it is important to note that the IV approach measures the local average treatment effect (LATE), which essentially quantifies the average treatment effect for individuals whose treatments are influenced by the instrument. In the context of our study, this implies that when employing the IV approach, we establish a causal relationship between the FT routing decisions and the outcomes for patients whose FT routing decision is affected by our IV (i.e., the relative congestion level between the main area and the entire ED).

4.3. Estimation

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

4.3.1. Continuous Outcome Variable We start with the outcome *LOS*. Given that LOS_i is a continuous variable while the endogenous variable FT_i is binary, employing a standard two-stage least squares (2SLS) approach would treat the first stage as a linear probability model. This approach, although unbiased, may result in potentially inefficient estimation. For more efficient estimation, we adopt a non-linear parametric modeling approach to simultaneously estimate both the FT routing decision model and the patient outcome model, as outlined in Chan et al. (2018). We first formulate the FT routing decision using a latent variable approach as follows:

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

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

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

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

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

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

4.3.2. Binary Outcome Variables We next consider two binary outcome variables: $Revisit_i$ and $LWBS_i$. In these cases, both the endogenous variable (i.e., the FT routing decision) and the outcome variables are binary. We follow Kim et al. (2015) and Chan et al. (2018) and use a nonlinear parametric model approach to jointly estimate $Outcome_i$ (i.e. $Revisit_i$ and $LWBS_i$) and FT_i. More specifically, we employ the recursive bivariate probit model (see Maddala 1986 and Greene 2018), which contains two probit models with correlated error terms as follows:

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

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

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

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

where $Outcome_i$ represents $Revisit_i$ or $LWBS_i$; and FT_i^* and $Outcome_i^*$ are the latent variables associated with FT_i and $Outcome_i$, respectively. The error terms ε_i and ξ_i of the FT routing decision and patient outcome models are jointly distributed following a bivariate normal distribution $\Phi_2(\varepsilon_i, \xi_i; \rho)$ with correlation coefficient ρ. The vector $\mathbf{X_i}$ is the same as in Equation (4). Additionally, we control for the patient waiting time ($WaitTime_i$). It is worth noting that $WaitTime_i$ is not included in the *LOS* regression model in Equation (4) because LOS_i is the sum of $WaitTime_i$ and patient i's diagnosis and treatment time (see Figure 1). If we control $WaitTime_i$ (equivalent to conditional on patient waiting time), Equation (4) examines the variation in the diagnosis and treatment time only.

Note that LWBS patients did not receive treatment at EDs. Hence, when estimating $LWBS_i$, we exclude $AvgOccTreated_i$ and $Workload_i$ from the control variables as they are measured during the treatment period. To control the ED congestion level, we introduce a new census variable denoted as $Census_i$. Following the

approach outlined by Batt and Terwiesch (2015), we define $Census_i$ as the number of patients in the waiting area where patient i is assigned at the moment when patient i is assigned to this area. Furthermore, the *WaitTime* for a LWBS patient is defined as the time interval from the end of triage to the time when the patient is identified as having left the ED, which is analogous to the offered wait variable defined in Batt and Terwiesch (2015). Finally, we cluster standard errors at the physician level and estimate the model through FMLE.

4.4. Patient Classification

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

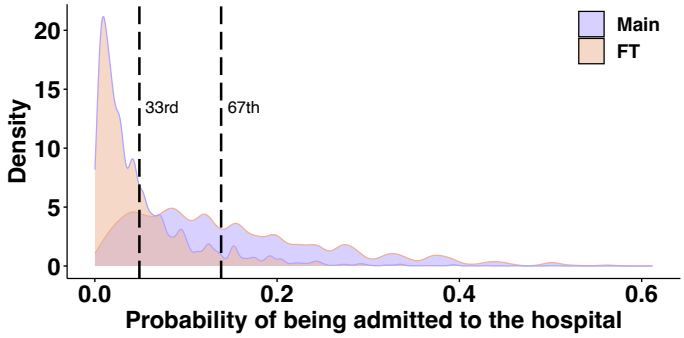
In this regard, a patient streaming strategy based on predicted disposition (i.e., admitted to the hospital vs. discharged from the ED) has been found to be successful by ED practitioners (O'Brien et al. 2006, Kelly et al. 2007). Moreover, in the OM literature, Saghafian et al. (2012, 2014) demonstrate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

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

where $\mathbf{X_i^p}$ denotes a vector of readily available patient characteristics during the triage process, which include triage score, age group, gender, and chief complaint. Additionally, ϕ_i represents the unobservable component following a standard normal distribution. We then create patient complexity classes by partitioning the fitted probability of admission (denoted as $\hat{M}_i$) based on patient clinical characteristics collected during triage. The fitted probability $\hat{M}_i$ here is computed as $\hat{M}_i = \Phi(\hat{\beta}_p \mathbf{X_i^p})$, where $\hat{\beta}_p$ is the estimated β_p and $\Phi(\cdot)$ is

the cumulative distribution function of the standard normal distribution. Intuitively, the higher the fitted probability, the more likely the patient would be admitted to the hospital, and hence, this patient is more likely to be classified as of a higher complexity level. Figure 2 depicts the fitted probability distribution for patients routed to the main and FT area, respectively. We observe that most patients with a high value of $\hat{M}_i$ were routed to the main area, whereas most patients with a low value of $\hat{M}_i$ were routed to the FT area. This observation is consistent with our intuition that patients with a higher probability of being admitted to the hospital are likely to be higher-complexity patients who should be treated in the main area. Nevertheless, we still observe a few patients with a high value of $\hat{M}_i$ who were routed to the FT area and vice versa. Therefore, we are interested in understanding whether any hidden consequence exists for patients treated in the FT area but would have been routed to the main area in a less congested ED.

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



Next, we consider the following complexity classification approach: a patient belongs to (i) the high-complexity class if $\hat{M}_i > t_2$, (ii) the low-complexity class if $\hat{M}_i < t_1$, and (iii) the medium-complexity class if $t_1 \leq \hat{M}_i \leq t_2$, where the two thresholds t_1 and t_2 are determined based on the density function of the fitted probability $\hat{M}_i$. Following a similar logic as the thresholds choice in Chan et al. (2018), a larger t_2 increases the percentage of patients with $\hat{M}_i > t_2$ being routed to the main area; similarly, a smaller t_1 increases the percentage of patients with $\hat{M}_i < t_1$ being routed to the FT area. The goal of our selected thresholds (t_1, t_2) is then twofold: (i) to balance the increasing percentage of patients in the high- (low-) complexity group being routed to the main (FT) area while maintaining a large enough patient sample in each group for meaningful statistical analyses and (ii) to facilitate generalizability to other hospital settings. Consequently, we set t_1 at the 33rd percentile and t_2 at the 67th percentile of $\hat{M}_i$, thereby dividing the patient population into three groups of equal size. This division not only guarantees the generalizability of our results but also creates well-balanced and representative groups across a range of patient complexities. Tables 7 and 8 in Appendix A present the summary statistics of patient characteristics and outcomes (i.e., revisits, LWBS, and LOS) for the three complexity classes. We can then estimate the impact of FT routing decisions on patient outcomes based on patient complexity subgroups. Later, we also conduct robustness checks with alternative choices of t_1 and t_2 and show our empirical results remain consistent.

5. Estimation Results

In this section, we present our estimation results.

5.1. Correlation Between Operational Status and FT Routing Decisions

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

$$FT_i^* = \beta \mathbf{X_i} + \alpha MEBusyRatio_i + \omega_h + \tau_m + \theta_t + \varepsilon_i \tag{10}$$

$$FT_i = \mathbb{1}\{FT_i^* > 0\} \tag{11}$$

where vector $\mathbf{X_i}$ again includes the age group, gender, chief complaint, triage score, and triage time of patient i. The variables ω_h, τ_m, and θ_t represent the hospital, month-year, and weekday fixed effects. The error term ε_i follows a standard normal distribution. The variable of interest $MEBusyRatio_i$ here measures the relative congestion level between the main area and the entire ED. Table 2 below presents the estimation results for all patients as well as patients in each complexity group; see Table 9 in Appendix A for the full estimation results. In addition, we also include the average marginal effect (AME) of $MEBusyRatio_i$ on the FT routing decision FT_i computed based on the estimated coefficients.[3]

Based on results in Table 2, we find that the coefficient of $MEBusyRatio_i$ is positive and significant (p-value < 0.01) for all the analyses (i.e., full patient sample, high-, medium-, and low-complexity groups), indicating a positive correlation between the relative congestion level of the main area to the entire ED and the likelihood of being routed to the FT area. Specifically, based on the AME in Table 2, we find that one standard deviation increase in $MEBusyRatio_i$ is associated with a 1.2% increase in the likelihood of being routed to the FT area based on the full patient sample. Moreover, this positive correlation varies across different patient complexity groups. In particular, one standard deviation increase in $MEBusyRatio_i$ corresponds to a 0.6%, 1.4%, and 1.8% higher likelihood of being routed to the FT area for the high-, medium-, and low-complexity groups, respectively. These results suggest that FT routing decisions are not purely clinical-driven, ED operational status related to congestion is also a critical factor in the FT routing decision-making process. Additionally, given that $MEBusyRatio_i$ also serves as the IV to analyze the effects of being routed to the FT, the heterogeneous effects observed across complexity groups may imply that the treatment effect of FT routing could vary across different complexity groups. This suggests that our proposed IV is not merely a random or weak proxy for treatment assignment; instead, it appears to have a meaningful relationship with the treatment and outcomes within these specific patient subgroups. This observation aligns with the previously mentioned Cragg-Donald Wald F statistics, which suggest that our identification

[3] To calculate the average marginal effect (AME) for continuous variable in this case, we first derive the marginal effect for each observation in our dataset, which is the derivative of the probability of being routed to the FT area with respect to *MEBusyRatio*. The AME is then the average of these marginal effects across all observations.

approach is not weak, providing further support for the robustness of our identification approach. Besides, based on results in Table 9 in Appendix A, we find that clinical factors, such as age group, triage score, gender, and triage time, are also associated with FT routing decisions.

Table 2 Results on the correlation between operational status and the fast-track routing decisions.

	All patients	High-complexity	Medium-complexity	Low-complexity
MEBusyRatio	0.083***	0.084***	0.093***	0.079***
	(0.005)	(0.013)	(0.009)	(0.007)
AME	0.012***	0.006***	0.014***	0.018***
	(0.001)	(0.001)	(0.001)	(0.002)
N	142,683	46,600	48,772	46,789
Pseudo R^2	0.541	0.292	0.441	0.405

Standard errors in parentheses. Some observations are dropped due to the perfect separation.
See Table 9 in Appendix A for the full estimation results. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.2. Impact of FT Routing Decisions on Patient Outcomes

This section discusses our main results on the impact of FT routing decisions on patient outcomes. Table 3 presents results both with and without IV to illustrate the potential estimation bias without IV. It is important to note that the IV approach estimates the LATE, which essentially quantifies the average treatment effect for individuals whose treatments are affected by the instrument. Additionally, instead of the estimated coefficient of the FT routing variable FT_i (γ in Equations (4) and (7)), we present the AME of FT_i on each patient outcome variable for the interpretation purpose. The estimated coefficient γ can be found in Table 10 in Appendix A, which presents the full estimation results.

To start with, we consider analyses with all patients. From panel A of Table 3, we find that being routed to the FT area reduces the average *LOS* (i.e., a negative AME of -0.388 with p-value < 0.01). To understand the *LOS* reduction in hours, we compute the predicted values of *LOS* when patients were routed to the main versus FT area using our estimation results, which gives us an average reduction of 0.89 hospital hours in *LOS* (i.e., $4.02 - 3.13 = 0.89$). We remark that here we cannot directly interpret the *LOS* reduction in hours using AME values in Table 3 because the dependent variable *LOS* is log-transformed and the AME measures the marginal effect of log(*LOS*). As a result, we interpret our results using predicted values. Although earlier medical literature has shown the effectiveness of FT on reducing patient *LOS* (see Sanchez et al. 2006, Devkaran et al. 2009, Chrusciel et al. 2019, and Grant et al. 2020), our work further validates this result with a more comprehensive approach and a new hospital setting (data) in Canada. Furthermore, our analysis reveals that being routed to the FT reduces the occurrence of patients leaving without being seen by 1.5% when considering the entire patient cohort. Additionally, when we estimate the effects of FT routing on patient revisits using all patient data without accounting for variations in care needs across patient complexity groups, we do not find statistically significant effects on $Revisit_{48h}$. However, as discussed earlier,

the impact of FT routing decisions may differ among patients with varying complexity levels. Therefore, we proceed to investigate the effects based on patient complexity groups.

Table 3 The AME of being routed to fast-track on patient outcomes.

	With IV				Without IV			
Outcome variables	AME (SE)		ρ (SE)		Test $\rho = 0$	AME (SE)	*Pseudo* R^2	*N*
Panel A: All patients								
$Revisit_{48h}$	-0.004	(0.007)	0.009	(0.033)	0.778	-0.002 (0.004)	0.404	123,655
LWBS	-0.015***	(0.004)	0.000	(0.047)	0.996	-0.015*** (0.001)	0.463	146,481
log(*LOS*)	-0.388***	(0.117)	0.035	(0.075)	0.640	-0.339*** (0.019)	0.247	142,683
Panel B: High-complexity patients								
$Revisit_{48h}$	0.082**	(0.035)	-0.181**	(0.075)	0.018	0.017** (0.007)	0.187	41,171
LWBS	-0.012	(0.008)	0.275	(0.238)	0.272	0.001 (0.004)	0.235	48,404
log(*LOS*)	-0.746***	(0.121)	0.123*	(0.071)	0.086	-0.551*** (0.023)	0.105	47,122
Panel C: Medium-complexity patients								
$Revisit_{48h}$	0.023*	(0.013)	-0.093*	(0.049)	0.058	0.002 (0.005)	0.335	41,701
LWBS	-0.020***	(0.007)	0.200**	(0.095)	0.041	-0.006* (0.003)	0.365	49,858
log(*LOS*)	-0.652***	(0.072)	0.199***	(0.045)	0.000	-0.371*** (0.020)	0.194	48,772
Panel D: Low-complexity patients								
$Revisit_{48h}$	0.003	(0.009)	-0.085	(0.060)	0.163	-0.009*** (0.004)	0.332	40,783
LWBS	-0.007	(0.008)	-0.131*	(0.075)	0.086	-0.020*** (0.003)	0.346	48,219
log(*LOS*)	-0.695***	(0.069)	0.308***	(0.045)	0.000	-0.270*** (0.018)	0.212	46,789

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

5.2.1. Effects on Different Complexity Groups Panels B, C, and D in Table 3 present the impacts of FT routing on high-, medium, and low-complexity patient groups, respectively. First, we observe a consistent reduction in LOS across all complexity groups. This is evident from the negative coefficients of -0.746, -0.652, and -0.695 for high-, medium-, and low-complexity groups, respectively, all statistically significant with p-value < 0.01. More specifically, high-complexity patients experience an average *LOS* reduction of 1.92 hours, medium-complexity patients 1.39 hours, and low-complexity patients 1.29 hours. These reductions are calculated by comparing the predicted *LOS* values when patients are routed to the main versus FT area using our estimation results. Moreover, for medium-complexity patients, being routed to the FT reduces the likelihood of leaving without being seen by 2%.

However, the impact on the quality of care varies. Notably, we find that being routed to the FT area could negatively affect the quality of care, leading to an 8.2% increase in the likelihood of 48-hour revisits for high-complexity patients and a 2.3% increase for medium-complexity patients. Given that the IV approach estimates the LATE, which primarily pertains to patients whose routing decisions are affected by the relative ED congestion, it is likely that this subset of patients is more substantial within the lower complexity groups and less so within the higher complexity groups. The results presented in Section 5.1 provide some

evidence of this conjecture, showing a stronger correlation between relative ED congestion and FT routing decisions for the low-complexity group, followed by the medium- and high-complexity groups. In other words, patients in the lower complexity groups are more likely to receive the treatment due to the instrument (i.e., relative ED congestion), and the reverse is true for higher complexity groups. However, we notice significant negative effects on higher complexity groups, who are expected to be less likely to receive the treatment. This observation suggests that for patients within the high- or medium-complexity groups who are induced to receive the treatment due to the instrument, the treatment has a more substantial negative impact on their outcomes compared to individuals in the low-complexity group. In fact, for low-complexity patients, we do not observe statistically significant effects on outcomes except for *LOS*. These findings support the rationale behind introducing the FT area: to expedite the treatment of low-complexity patients and improve operational efficiency without compromising the quality of care.

To sum up, these findings emphasize the need for hospital and ED managers to carefully balance the tradeoff between care efficiency and quality, particularly for high- and medium-complexity group patients. The full estimation results can be found in Table 10 in Appendix A, which also reveal a positive correlation between waiting time and 48-hour revisits. However, an extra hour of waiting is associated with only a 0.4% increase in 48-hour revisits for high-complexity patients and a 1% increase for medium-complexity patients.

5.2.2. Discussion on the Likelihood Ratio Tests The third column of Table 3 shows the estimated correlation $\rho(SE)$ between the error terms of the FT routing decision equation and the outcome equation. The fourth column of Table 3 presents the p-values of the likelihood ratio test results "Test $\rho = 0$" that compares the log-likelihood of our full model with the sum of log-likelihood of two separate models. Similar to the Hausman test, the likelihood ratio test checks the exogeneity of a dummy independent variable with a dummy dependent variable (Knapp and Seaks 1998). We can see from panel B of Table 3 that for high-complexity patients, the p-values of the likelihood ratio test is less than 0.05 for the 48-hour revisits and less than 0.1 for the *LOS*, indicating a strong endogeneity issue, which also explains the difference between the results with IV and without IV. Similarly, for medium-complexity patients, the likelihood ratio test indicates the existence of a strong endogeneity issue across all the outcome variables; see panel C of Table 3. Finally, for low-complexity patients, the likelihood ratio test suggests the existence of endogeneity issues in the estimation of the *LOS* and *LWBS* outcome variables. These results indicate the importance of adopting an IV approach to get consistent estimates of the impact of FT routing decisions on patient outcomes.

5.3. Discussion on the Mechanism

Our main results, as presented in Section 5.2, show that being routed to the FT area significantly reduces the *LOS*. However, there are hidden consequences for high- and medium-complexity group patients, as being routed to FT increases their revisit rates. To investigate the underlying factors contributing to these results,

we first analyze the components of LOS, namely waiting time and treatment time. Subsequently, we delve into the impact on ordered tests.

Our analysis begins with an assessment of the impact of FT routing decisions on waiting and treatment times, applying the models from Equations (2)–(4). The treatment time here is measured as the actual interaction time a physician spent on this patient, derived from the physician's activity logs.[4] This approach excludes non-interactive periods such as the time duration when patients are going through tests and waiting for results while the attending physician is caring for other patients, and thus provides a more accurate measure of direct patient care time. Additionally, waiting time is measured as the period from the end of triage to the moment when this patient is picked up by a physician, as explained in Section 3.3.2. Columns (1) and (2) in Table 4 present the estimation results for different patient groups. The full estimation results are presented in Tables 16 and 17 in the supplementary document. Our analysis reveals that being routed to the FT area leads to a decrease in wait time across all patient groups, as well as a reduction in treatment time (direct patient care time) for the medium-complexity group.

The observed reduction in treatment time in the FT area offers a plausible explanation for the decreased quality of care, especially for medium-complexity patients. To further understand underlying factors contributing to adverse effects on the quality of care, we explore the following measures related to ordered tests: the number of lab tests, prescribed medications, and diagnostic images (including CT scans and X-ray tests). These factors reflect the complexity of a patient's diagnosis and treatment process following their routing to a specific treatment area. Since all these factors are count variables, we employ a negative binomial model to estimate the impact of the FT routing decision while controlling for variables such as age group, gender, chief complaint, triage score, triage time, hospital fixed effect, month-year fixed effect, and weekday fixed effect. Columns (3)–(6) in Table 4 present the estimation results.

To begin with, we find that being routed to the FT area reduces the number of lab tests, prescribed medications, and CT scans for patients across all complexity groups. These findings suggest that FT physicians tend to simplify the diagnosis and treatment process for patients in the FT area. However, it is worth noting that FT routing leads to an increase in the number of X-ray tests for medium- and low-complexity patients. One plausible explanation is that physicians working in the FT area may tend to order fewer CT scans, which produce detailed and high-quality images, and instead opt for more X-ray tests, which are faster and less expensive but provide less detailed information compared to CT scans. Physicians in the FT area understand that the primary goal of the FT area line is for the fast delivery of care for less complex and less urgent patients. Hence, they may prioritize expediting patient flow by simplifying patient diagnostic procedures. Consequently, the quality of care may be compromised for high- and medium-complexity patients routed

[4] Our data records the start time of each physician-patient interaction, including the initial assessment and all follow-ups, but does not have the end time. Hence, we use the start time of the next activity as the end time of the current activity of the same physician. The difference between the start time and end time serves as an estimate of the interaction time.

to the FT area. On the other hand, low-complexity patients are often easier to diagnose compared to high- and medium-complexity patients. Hence, the simplified patient diagnostic process in the FT area may have a minimal impact on their quality of care.

Table 4 Estimation results for the impact of the FT routing decisions on the numbers of lab tests, the number of medication orders, the number of CT (computerized tomography) scans, the number of X-ray tests, treatment time, and wait time across patients of different complexity levels for Dataset III.

	(1)	(2)	(3)	(4)	(5)	(6)
	TreatTime	*WaitTime*	*LabTests*	*Medications*	*CTScans*	*X-rayTests*
High-complexity	-0.133	-1.847***	-5.193***	-1.382***	-0.132***	-0.006
	(0.082)	(0.046)	(0.121)	(0.078)	(0.016)	(0.017)
Medium-complexity	-0.360***	-1.230***	-3.436***	-0.887***	-0.086***	0.085***
	(0.102)	(0.069)	(0.075)	(0.039)	(0.007)	(0.009)
Low-complexity	0.024	-0.518***	-2.101***	-0.459***	-0.047***	0.162***
	(0.144)	(0.092)	(0.062)	(0.018)	(0.003)	(0.010)

5.4. Robustness Checks

This section briefly summarizes our robustness checks. To start with, we explore several alternative IVs, including adjusting for area workload by dividing the number of physicians on duty in a particular area and directly using the measure *FTCongestion* as an IV. We also consider alternative IVs calculated from different time intervals (i.e., 0.5, 1, and 2 hours) prior to patient triage start time. Furthermore, we investigate additional or alternative control variables. We conduct a robustness check by controlling for patient comorbidity information and explore an alternative method for quantifying physician workload based on the number of patients receiving care from physician p_i during the treatment process of patient i. We also conduct an additional analysis that controls for the ED congestion at the time when the attending physician makes the disposition decision. Besides, we consider alternative patient classification cutoffs, alternative outcomes with 72-hour revisits, and alternative samples, including the removal of extreme observations with triage times longer than 17 minutes and the inclusion of patients with triage level 1. We also employ a matching approach to create comparable patient samples between the main and FT areas. Finally, we consider alternative model specifications by including physician and time-of-day fixed effects, as well as another robustness check to incorporate the ED congestion when the attending physician makes disposition decisions and physician workload into the patient classification model. Across all these robustness checks, our results remain consistent. See Appendix C for details.

6. Comparison of Alternative Fast-Track Routing Policies

Based on the empirical results, we find that hidden consequences on patient outcomes might occur when routed to the FT area. In this section, we build a discrete event simulation model to simulate the ED patient flow process. The objective of the simulation is to compare different routing policies and derive managerial

implications to help guide FT routing decisions. To supplement this section, we present the queueing model framework details in Appendix D, MDP formulation in Appendix E, constrained MDP in Appendix F, additional simulation results and model validation in Appendix G.

6.1. Simulation Design, Input Modeling, and Validation

In this section, we describe the simulation design, input modeling, and validation in detail.

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

Patient Routing and Abandonment. Based on predefined routing policies (see more details in Section 6.2), patients will be routed to the main treatment area or the FT area, waiting in the corresponding queue. The probability of patients left without being seen during waiting is calculated by our estimation model outlined in Section 4.3.2. The FT area in our study hospitals operates from 10 am to midnight; hence, no patients will be routed to FT outside this period. When the FT area closes at midnight, we assume that an exhaustive service discipline is applied (Ingolfsson et al. 2007), i.e., the FT physician completes the treatment of the patient whose diagnosis is in process before they leave work. Other patients waiting in the FT area are moved to the main area instantaneously.

Service Process. Physicians go to work according to a shift-based schedule. In the simulation, we use the actual schedule from our study hospital. As a result, the number of physicians on duty is time-varying, determined by the shift schedule (the FT area always has one working physician). We assume that physicians do not idle if there are patients waiting to be seen. Physicians select the next patient to treat based on a discrete choice model, in which a patient's priority of being seen mainly depends on the triage score and the current waiting time (each triage level is associated with a quadratic marginal waiting cost function, see, e.g., Ding et al. 2019). We acknowledge that existing studies have established the dependence of service times on system states such as the physician workload; see, e.g., KC and Terwiesch (2009, 2012), Batt and Terwiesch (2016), Berry Jaeker and Tucker (2016). Motivated by this literature, we consider a shift-hour-dependent service rate, which is the number of new patients seen by a physician at the corresponding shift hour observed from the data, as shift hours are crucial in determining service rates (Zaerpour et al. 2022). This level of abstraction has been shown to be sufficient to generate dynamics that match the data from the ED process (Ouyang et al. 2021).

Implementation and Validation. The simulation model is implemented in Python using SimPy 4.0. For the purpose of model validation, we start the simulation with an empty ED and run 30 replications with a replication length of 500 weeks (the first 100 weeks are identified as the warm-up period and thus

removed). The routing policy used in the simulation for validation is based on the estimated current routing policy (Policy CP in Section 6.2). All parameters related to the inter-arrival time generation, service time generation, and the current routing policy are estimated using data from one of our study hospitals from January 2015 to July 2015, as the shift schedule was fixed during this period. We validate our simulation model using data and see the detailed results in Appendix G.

6.2. Alternative Fast-Track Routing Policies

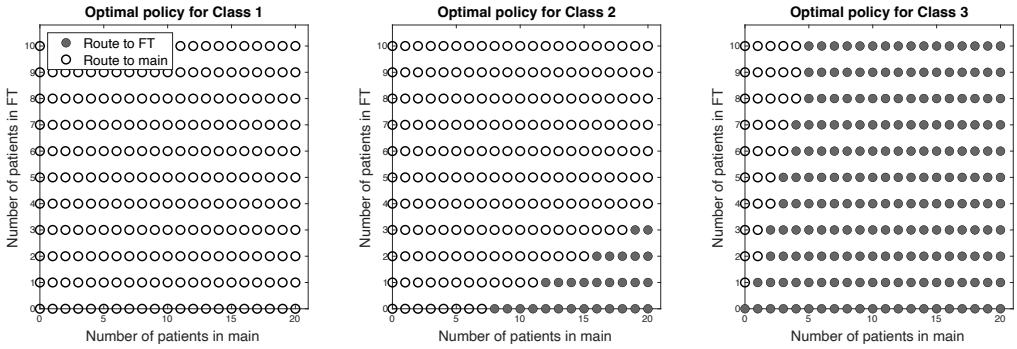
In this section, we compare the performance of different fast-track routing policies by simulation. These policies differ in their state dependency or the classification of patients that are routed to FT. It is worth noting that the purpose here is not to propose routing policies that are readily implementable in EDs; rather, our goal is to provide managerial implications for hospital FT routing decisions.

We first estimate the current routing policy implemented in our study hospitals, referred to as *Policy CP* hereafter, which serves as a benchmark for other routing policies under consideration. The estimation adheres to the probit model in Equation (10). Note that Policy CP depends on both the patient's characteristics and ED system state (see estimation results in Section 5.1). We are interested in if better state-dependent policies can be devised by leveraging our estimation results. Hence, we follow the procedure described in Appendix E to model the FT routing decisions by a Markov decision process (MDP) formulation and solve for the optimal routing policy numerically. We refer to this policy as *Policy OP* hereafter. Note that the MDP formulation assumes aggregated patient arrivals and services at the daily level. Hence, the optimal policy for the MDP model is *not* necessarily the optimal policy for our simulation setup which assumes time-of-day dependent arrival and service processes. Nevertheless, we believe the structure of the optimal policy can provide insights into the FT routing decisions.

Figure 3 illustrates the optimal policy used in the simulation study. From Figure 3, we observe that Class 1 (i.e., high-complexity) patients should almost always be routed to the main area, whereas it is optimal to route Class 3 (i.e., low-complexity) patients to the FT area under most circumstances. The dynamic routing mainly applies to Class 2 (i.e., medium-complexity) patients. Specifically, when the main area is significantly more crowded than the FT area, it is optimal to route more patients of Class 2 to the FT area. For example, it can reduce the waiting time of Class 2 patients by routing them to FT when there are 20 patients in the main area and less than 4 patients in the FT area, which also eases the congestion level in the main area.

FT area aims toward fast care delivery and rapid discharge for patients with less urgent conditions. Hence, it is imperative that the waiting time (or LOS) at FT and the severity levels of patients routed to FT remain similar to that under the current practice (i.e., Policy CP). To align with the primary objective of FT, we consider a constrained MDP formulation by adding a constraint on the queue length in the FT area to the above-mentioned MDP model (see more details in Appendix F). Interestingly, the policy from the constrained MDP formulation shares the same structure as the MDP except that less patients are routed to

Figure 3 An illustration of the optimal routing policy (Policy OP) used in our simulation study.



the FT area (see Figure 4 in Appendix F) so as to retain the fast delivery of service at FT. We refer to the optimal policy of the constrained MDP as *Policy COP* hereafter.

Motivated by the insights from structure of the optimal routing policy that it is better to route patients of lower complexity to FT, we propose a static routing policy, referred to as *Policy SP*, as follows. Specifically, we route a patient to FT if their predicted admission probability (defined in Equation (9)) is lower than a given percentile; otherwise, the patient is routed to the main area. We adjust the percentile so that the proportion of patients treated in FT is comparable to that under Policy CP. It is worth noting that the admission prediction model only uses information collected during triage (such as age, gender, triage level, and chief complaints). In fact, existing studies have shown that a patient's disposition can be fairly accurately predicted using triage information (see, e.g., Holdgate et al. 2007, Vaghasiya et al. 2014).

Finally, we also consider the policy that routes patients of triage levels 4 and 5 to the FT area, and patients of triage levels 1–3 to the main area. We refer to this triage-score-based routing policy as *Policy TP*. Plausibly due to its simplicity, the triage-score-based routing policy has been implemented in some Canadian EDs (Ding et al. 2019, Al Darrab et al. 2006) despite the lack of understanding of its effectiveness. The percentage of patients routed to FT under TP will be about 1.5% higher than that under Policy CP, which lead to higher workload and congestion level, as shown by our simulation results in Table 5. To ensure a fair comparison, we consider a modified policy (referred to as *Policy TP-C* hereafter) that routes the additional 1.5% patients to the main area so that the workload under CP and TP-C are similar.

6.3. Discussion of Results

In this section, we compare the performance of alternative policies by simulation. We run the simulation under each routing policy for 30 replications, where each replication has a length of 500 weeks with the first 100 weeks as the warm-up period. We calculate the 48-hour revisits, the average patient LOS, and the LWBS rate for each routing policy and include them in Table 5. The percentages of patients routed to FT by

their complexity or triage levels are also included in Table 5, which we believe can provide further insights into the performance of each routing policy. We obtain the following observations.[5]

Table 5 The 95% confidence interval for the 48-hour revisits, LOS, LWBS, the percentage of patients routed to FT under each routing policy, and the percentage reduction in the 48-hour revisits by using alternative policies over CP.

	CP	TP	TP-C	SP	COP	OP
The 48-hour patient revisits	$5,189 \pm 6$	$5,363 \pm 5$	$5,335 \pm 5$	$5,095 \pm 4$	$5,021 \pm 5$	$4,994 \pm 5$
% reduction in revisits		-3.35 ± 0.17	-2.82 ± 0.12	1.82 ± 0.16	3.24 ± 0.14	3.75 ± 0.13
Average LOS (hours)						
All patients	2.06 ± 0.01	2.02 ± 0.01	2.05 ± 0.01	2.10 ± 0.01	1.90 ± 0.01	1.83 ± 0.01
Patients in main area	2.10 ± 0.01	1.91 ± 0.01	2.03 ± 0.01	2.16 ± 0.01	1.92 ± 0.01	1.73 ± 0.01
Patients in FT area	1.84 ± 0.01	2.45 ± 0.01	2.11 ± 0.01	1.86 ± 0.01	1.85 ± 0.01	2.25 ± 0.01
LWBS (%)	1.84 ± 0.03	2.16 ± 0.03	2.77 ± 0.04	2.73 ± 0.04	1.22 ± 0.03	0.95 ± 0.02
% of patients routed to FT	24.01 ± 0.02	25.49 ± 0.03	23.92 ± 0.03	23.46 ± 0.02	24.44 ± 0.03	25.72 ± 0.03
% of patients routed to FT by complexity						
High-complexity	2.33 ± 0.02	4.64 ± 0.02	4.29 ± 0.02	0.49 ± 0.01	0.49 ± 0.01	0.49 ± 0.01
Medium-complexity	18.91 ± 0.04	20.24 ± 0.04	18.93 ± 0.04	1.26 ± 0.01	5.64 ± 0.05	5.65 ± 0.05
Low-complexity	59.76 ± 0.05	60.67 ± 0.05	57.41 ± 0.06	78.36 ± 0.05	77.18 ± 0.06	81.79 ± 0.06
% of patients routed to FT by triage scores						
CTAS 2	7.30 ± 0.03	0.59 ± 0.01	0.59 ± 0.01	4.75 ± 0.02	5.42 ± 0.03	5.63 ± 0.03
CTAS 3	22.00 ± 0.05	1.25 ± 0.01	1.26 ± 0.01	15.63 ± 0.03	19.96 ± 0.03	20.91 ± 0.04
CTAS 4	45.38 ± 0.07	90.74 ± 0.06	82.47 ± 0.05	55.92 ± 0.07	52.51 ± 0.07	55.44 ± 0.07
CTAS 5	55.00 ± 0.13	92.46 ± 0.07	92.82 ± 0.07	66.11 ± 0.10	60.89 ± 0.14	64.42 ± 0.11

Notes: The calculation of the 48-hour patient revisits is based on the total number of discharged patients during FT open hours for the two EDs in 24 months (i.e., a total of 123,655 observations). There is a small percentage of patients of CTAS 2 and 3 routed to FT under TP and TP-C because admitted patients are routed based on the actual routing decisions in the data.

Observation 1. *(i) The forward-looking policy OP performs the best among all the routing policies in terms of reducing the 48-hour patient revisits, the average LOS, and the LWBS rate; (ii) Policy COP outperforms CP and all static routing policies (SP, TP, TP-C) while maintaining fast delivery of care at FT.*

Our simulation results reveal that the fine-tuned state-dependent policy OP (devised based on the MDP formulation) reduces the 48-hour patient revisits over the current routing policy used in our study EDs (Policy CP) by 3.75%. The LWBS rate decreases from 1.84% under CP to 0.95% under OP. At the same time, Policy OP reduces the average LOS of all patients by 11.2%, compared to CP. However, our results also show that 25.72% patients are routed to the FT area under OP, whereas the FT area treats 24.01% patients under CP. As a result, the average LOS for FT patients increases from 1.84 hours to 2.25 hours, i.e., FT patients stay in the ED 24.6 minutes longer under OP, which undermines the primary objective of setting up an FT area and also motivates Policy COP.

[5] The average queue length of FT and the average patient waiting time are also recorded. Interested readers can refer to Table 11 in Appendix G.

By adjusting the constraint in the constrained MDP formulation, we make sure that the percentage of patients routed to FT and the average LOS of FT patients under Policy COP are similar to that under Policy CP (i.e., 24.44% vs. 24.01% and 1.85 hours vs. 1.84 hours, respectively). Although COP performs slightly worse than OP (in terms of revisits, LOS, and LWBS), it surpasses policy CP and all static routing policies (SP, TP, TP-C) in performance by a significant margin. A closer look at Policy COP further finds that compared to Policy CP, patients of higher complexity level or higher urgency level (i.e., smaller triage score) are routed to FT less frequently, which may explain the reduction in revisits under Policy COP.

Hence, we conclude that a forward-looking state-dependent policy can improve over the current routing policy in practice while maintaining the primary objective of FT—delivering fast care to patients with minor conditions. Note that our empirical results have shown that Policy CP also takes account the system state information, particularly the relative congestion levels between the main area and FT. It is plausible that triage nurses are not forward-looking and only considers the current state information under Policy CP.

Observation 2. *(i) The static policy SP reduces the 48-hour patient revisits over CP; (ii) The triage-score-based static policies TP and TP-C perform worse than all other policies under consideration.*

The results on SP and TP-C—both are static routing policies—are of particular interest because the percentage of patients routed to FT under them is similar to that under policy CP. The simulation results show that Policy SP can reduce the 48-hour patient revisits over CP by 1.84%. On the other hand, the triage-score-based policy TP-C increases patient revisits (compared to CP) by 2.8%, which implies that patient classification based on their level of complexity can pick more suitable patients to be routed to FT to reduce patient revisits and improve patient outcomes. It is interesting to note that the average LOS of FT patients under SP and TP-C is longer than that under CP despite that the FT area treats slightly fewer patients under SP and TP-C (the results on patient LWBS rates are similar as LWBS depends heavily on patient waiting time). A plausible explanation is that SP and TP-C are the only policies that do not consider ED congestion levels in the routing decisions. Note that Policy TP actually routes more patients to FT than CP, which does not facilitate a fair comparison. We include the results on TP in Table 5 just for reference.

Finally, we discuss the insights from the comparison results and provide managerial implications to potentially help guide ED FT routing decisions. First of all, due to the hidden consequences of FT routing decisions, triage nurses should be more cautious and choose the "right" patients to route to FT. Our complexity-based classification method provides an option for hospital management to consider. Secondly, forward-looking state-dependent policies perform the best. The intuition is that the dynamic routing policy benefits from the server pooling effect, which, to a certain extent, makes up the "anti-pooling" deficit from setting up the FT area by placing physicians (also nurses and beds) into separate areas with dedicated queues. Lastly, despite being a popular policy in practice, routing patients purely based on their triage scores is not

recommended, and one potential drawback is the lack of accounting for the dynamics of ED congestion (compared to COP).[6]

7. Conclusion and Future Research

This paper studies the role of operational factors related to ED congestion in FT routing decisions and the subsequent impact of being routed to FT on patient outcomes using data from two Canadian EDs. The purpose of introducing an FT area is to reduce the waiting time for less urgent and less complex patients. However, the FT area forms a separate queue with a fixed allocation of medical resources, which may create the "anti-pooling" effect, as Saghafian et al. (2012) cautioned in their study. Triage nurses, the decision makers of FT routing, are aware of the congestion levels at both the main and the FT areas. Hence, it seems to be an intuitive and sensible decision to route patients who would be sent to the main area when the ED is less congested into the FT area when the main area is significantly more crowded, so as to reduce their waiting times. In fact, we find a positive correlation between the ED congestion level and the likelihood of being routed to the FT area. To a certain extent, routing decisions based on congestion levels achieve resource pooling between the main area and FT. Indeed, our results show that the congestion-dependent routing practice in our study EDs improves efficiency by reducing patient *LOS* and *LWBS*, 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 that the 48-hour revisits increase by 8.2% for the high-complexity group and by 2.3% for the medium-complexity group. Therefore, we advise caution since it has unintended consequences on the quality of care, especially for patients with more complex care conditions. Being aware of this important trade-off between care efficiency and quality, we evaluate the performance of different routing policies through simulation studies. Our results show that a better-informed routing policy can improve both care efficiency and quality of care compared to the current routing policy in our study hospitals. Interestingly, the triage-score-based policy, which routes all (and only) patients with triage scores 4 and 5 to the FT area, performs the worst among all the policies under consideration, despite its prevalent use as a guideline for making FT routing decisions in many hospitals. Our work, therefore, calls for attention from healthcare decision makers to carefully balance the trade-off between efficiency and the quality of care when making FT routing decisions.

As more hospitals have implemented FT areas in their EDs, it becomes increasingly important to establish consistent and evidence-based guidelines for FT routing decisions. Our study serves as an important step towards this goal. In what follows, we discuss some limitations of our study and point out opportunities

[6] It is, however, important to note that this finding is based on data from hospitals where the Canadian triage system (i.e., CTAS) is used and one should be cautious when extending the results to other triage systems such as the ESI. For example, CTAS focuses on the urgency of patients' medical conditions and does not consider the anticipated resource needs (unlike ESI, see, e.g., Gilboy et al. 2012). As a result, the efficacy and relevance of the triage-score-based routing policy in hospitals using the ESI system might vary.

for future research. First, our study focuses on two Canadian EDs where the main area and FT area share the same pool of physicians. While we believe many EDs have similar settings to ours (Ding et al. 2019, Al Darrab et al. 2006), we note that the staffing of FT areas in some hospitals can be different. For example, the ED studied by Sanchez et al. (2006) staffed physician assistants and nurse practitioners to provide care for patients routed to FT. Therefore, our results may not be directly applied to those hospitals, and it would be valuable to conduct analyses using data from more hospitals based on our framework. Second, we stratify patients into three complexity classes based on their predicted dispositions. It would be of interest for future studies to examine other classification methods that reflect patients' heterogeneous care needs from alternative perspectives. Finally, in our simulation study, we consider a shift-hour-dependent service rate. We recognize that a more precise system-state-dependent service rate could incorporate a physician's workload state, provided that the instantaneous workload could be accurately tracked. However, this modeling approach would require more granular data on physician-patient interactions, which is beyond the scope of our current dataset. Future research with more granular data might consider extending our simulation framework with such as a workload-dependent service rate.

References

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

Al Darrab A, Fan J, Fernandes CM, Zimmerman R, Smith R, Worster A, Smith T, O'Connor K (2006) How does fast track affect quality of care in the emergency department? *European Journal of Emergency Medicine* 13(1):32–35.

Altman E (2021) *Constrained Markov decision processes* (Routledge).

Anand KS, Paç MF, Veeraraghavan S (2011) Quality–speed conundrum: Trade-offs in customer-intensive services. *Management Science* 57(1):40–56.

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

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

Batt RJ, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* 61(1):39–59.

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

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

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

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

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

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

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

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

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

Feizi A, Carson A, Jaeker JB, Baker WE (2023) To batch or not to batch? impact of admission batching on emergency department boarding time and physician productivity. *Operations Research* 71(3):939–957.

Freeman M, Robinson S, Scholtes S (2021) Gatekeeping, fast and slow: An empirical study of referral errors in the emergency department. *Management Science* 67(7):4209–4232.

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

Gilboy N, Tanabe P, Travers D, Rosenau AM, et al. (2012) Emergency severity index (esi): a triage tool for emergency department care, version 4. *Implementation handbook* 2012:12–0014.

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

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

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

Green LV, Savin S, Savva N (2013) "nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science* 59(10):2237–2256.

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

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

Hampers LC, Cha S, Gutglass DJ, Binns HJ, Krug SE (1999) Fast track and the pediatric emergency department: resource utilization and patient outcomes. *Academic emergency medicine* 6(11):1153–1159.

Holdgate A, Morris J, Fry M, Zecevic M (2007) Accuracy of triage nurses in predicting patient disposition. *Emergency Medicine Australasia* 19(4):341–345.

Hopp WJ, Iravani SM, Yuen GY (2007) Operations systems with discretionary task completion. *Management Science* 53(1):61–77.

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

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

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

KC DS, Staats BR, Kouchaki M, Gino F (2020b) Task selection and workload: A focus on completing easy tasks hurts performance. *Management Science* 66(10):4397–4416.

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

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

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

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

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

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

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

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

Li W, Sun Z, Hong LJ (2023) Who is next: Patient prioritization under emergency department blocking. *Operations Research* 71(3):821–842.

Li X, Guo P, Lian Z (2016) Quality-speed competition in customer-intensive services with boundedly rational customers. *Production and Operations Management* 25(11):1885–1901.

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

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

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

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

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

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

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

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

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

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

Rubin DB (2006) *Matched sampling for causal effects* (Cambridge University Press).

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

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

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

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

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

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

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

Sun S, Lu SF, Rui H (2020) Does telemedicine reduce emergency room congestion? evidence from new york state. *Information Systems Research* 31(3):972–986.

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

Vaghasiya MR, Murphy M, O'Flynn D, Shetty A (2014) The emergency department prediction of disposition (epod) study. *Australasian Emergency Nursing Journal* 17(4):161–166.

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

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

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

Appendices

Appendix A. Tables

Table 6 Results for exogeneity of ED utilization as an IV. The regression models also control for the chief complaint, month-year, and weekday fixed effects.

	MEBusyRatio	*EDCongestion*	*MainCongestion*	*FTCongestion*
Age group in years (Base=0–25)				
25–40	0.007	−0.0003	−0.00005	−0.003
	(0.008)	(0.001)	(0.001)	(0.002)
40–55	0.007	0.002**	0.003***	−0.002
	(0.008)	(0.001)	(0.001)	(0.002)
55–70	0.004	0.004***	0.004***	0.001
	(0.009)	(0.001)	(0.001)	(0.002)
> 70	0.001	0.006***	0.006***	0.002
	(0.010)	(0.001)	(0.001)	(0.003)
Triage score (Base=CTAS 2)				
CTAS 3	−0.006	−0.002***	−0.002***	−0.0004
	(0.006)	(0.001)	(0.001)	(0.002)
CTAS 4	−0.004	−0.005***	−0.005***	−0.003
	(0.008)	(0.001)	(0.001)	(0.002)
CTAS 5	−0.015	−0.006***	−0.007***	−0.003
	(0.012)	(0.001)	(0.001)	(0.003)
Gender (Base=Female)				
Male	0.002	−0.001	−0.001	−0.001
	(0.005)	(0.001)	(0.001)	(0.001)
N	142,683	142,683	142,683	142,683
R^2	0.042	0.294	0.299	0.071

*Notes: *p<0.1; **p<0.05; ***p<0.01*

Table 7 Summary statistics for patients of different complexity classes (Dataset II).

	High-complexity				Medium-complexity				Low-complexity			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Age (years)	60.03	17.56	0.00	106.50	41.34	17.57	0.00	104.20	33.90	15.63	0.00	100.30
Gender (Male %)	54.83	49.77	0.00	100.00	38.89	48.75	0.00	100.00	44.94	49.74	0.00	100.00
Triage score (%)												
CTAS 2	57.21	49.48	0.00	100.00	26.78	44.28	0.00	100.00	11.35	31.72	0.00	100.00
CTAS 3	37.62	48.44	0.00	100.00	56.35	49.60	0.00	100.00	32.33	46.77	0.00	100.00
CTAS 4	4.20	20.05	0.00	100.00	14.06	34.76	0.00	100.00	38.88	48.75	0.00	100.00
CTAS 5	0.97	9.81	0.00	100.00	2.81	16.52	0.00	100.00	17.44	37.95	0.00	100.00

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

Table 8 Summary statistics for patient outcomes of different complexity classes.

		High-complexity		Medium-complexity		Low-complexity	
		Main Area	Fast-Track	Main Area	Fast-Track	Main Area	Fast-Track
	Dataset	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
LWBS (%)	I	1.74 (13.08)	0.60 (7.70)	3.37 (18.05)	0.71 (8.40)	6.30 (24.29)	1.09 (10.38)
LOS (hours)	II	5.74 (3.43)	3.49 (2.25)	4.58 (2.96)	3.24 (2.04)	3.70 (2.50)	2.71 (1.65)
$Revisit_{48h}$ (%)	III	6.65 (24.91)	6.80 (25.17)	7.38 (26.14)	5.18 (22.16)	5.09 (21.97)	3.27 (17.79)

Notes. SD = standard deviation; *LOS* = length of stay; *LWBS* = left without being seen.

Table 9 Full results on the correlation between operational status and FT routing decisions.

	All patients	High-complexity	Medium-complexity	Low-complexity
MEBusyRatio	0.083***	0.084***	0.093***	0.079***
	(0.005)	(0.013)	(0.009)	(0.007)
Age group in years (Base=0–25)				
25–40	-0.082***	-0.037	-0.039	-0.076***
	(0.015)	(0.122)	(0.036)	(0.018)
40–55	-0.123***	-0.336***	-0.082*	-0.079***
	(0.016)	(0.122)	(0.042)	(0.023)
55–70	-0.192***	-0.478***	-0.128**	-0.021
	(0.017)	(0.124)	(0.051)	(0.031)
> 70	-0.346***	-0.564***	-0.254***	-0.164***
	(0.019)	(0.127)	(0.065)	(0.054)
Triage score (Base=CTAS 2)				
CTAS 3	0.524***	0.471***	0.612***	0.345***
	(0.014)	(0.031)	(0.029)	(0.029)
CTAS 4	0.884***	0.821***	0.972***	0.653***
	(0.016)	(0.055)	(0.045)	(0.030)
CTAS 5	0.954***	1.187***	1.264***	0.653***
	(0.020)	(0.083)	(0.067)	(0.034)
Gender (Base=Female)				
Male	0.201***	0.051**	0.200***	0.296***
	(0.010)	(0.025)	(0.020)	(0.015)
Hospital (Base=ED A)				
ED B	-0.077***	-0.054**	-0.117***	-0.060***
	(0.010)	(0.024)	(0.018)	(0.015)
TriageTime	-0.188***	-0.152***	-0.169***	-0.212***
	(0.006)	(0.013)	(0.010)	(0.008)
N	142,683	46,600	48,772	46,789
Pseudo R^2	0.541	0.292	0.441	0.405

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

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

Table 10 Complete estimation results of the effects of FT routing on patient outcomes.

	All patients			High-complexity			Medium-complexity			Low-complexity		
	$Revisit_{48h}$	*LWBS*	log(*LOS*)	$Revisit_{48h}$	*LWBS*	log(*LOS*)	$Revisit_{48h}$	*LWBS*	log(*LOS*)	$Revisit_{48h}$	*LWBS*	log(*LOS*)
FT	-0.037	-0.341***	-0.388***	0.475***	-0.475**	-0.746***	0.181**	-0.443***	-0.652***	0.033	-0.117	-0.695***
	(0.068)	(0.112)	(0.117)	(0.160)	(0.212)	(0.121)	(0.091)	(0.095)	(0.072)	(0.113)	(0.129)	(0.069)
Age group in years (Base=0–25)												
25–40	0.108***	-0.053***	0.080***	-0.012	-0.059***	0.088***	0.068**	-0.068***	0.053***	0.178***	-0.034***	0.062***
	(0.020)	(0.008)	(0.008)	(0.040)	(0.021)	(0.028)	(0.030)	(0.012)	(0.014)	(0.030)	(0.008)	(0.010)
40–55	0.047*	-0.192***	0.172***	-0.110**	-0.205***	0.169***	-0.000	-0.230***	0.136***	0.195***	-0.137***	0.126***
	(0.024)	(0.025)	(0.008)	(0.043)	(0.036)	(0.029)	(0.046)	(0.034)	(0.017)	(0.038)	(0.016)	(0.012)
55–70	0.048*	-0.359***	0.249***	-0.137***	-0.367***	0.248***	-0.006	-0.359***	0.232***	0.243***	-0.348***	0.153***
	(0.026)	(0.046)	(0.009)	(0.043)	(0.053)	(0.029)	(0.056)	(0.050)	(0.022)	(0.048)	(0.035)	(0.014)
> 70	0.118***	-0.560***	0.373***	-0.058	-0.588***	0.391***	0.043	-0.501***	0.336***	0.322***	-0.532***	0.169***
	(0.028)	(0.072)	(0.014)	(0.049)	(0.076)	(0.032)	(0.078)	(0.068)	(0.029)	(0.112)	(0.065)	(0.028)
Triage score (Base=CTAS 2)												
CTAS 3	-0.064***	0.196***	-0.082***	-0.126***	0.240***	-0.086***	0.007	0.158***	-0.021	-0.072	0.159***	0.008
	(0.016)	(0.030)	(0.009)	(0.024)	(0.034)	(0.010)	(0.035)	(0.027)	(0.013)	(0.064)	(0.024)	(0.015)
CTAS 4	-0.190***	0.369***	-0.189***	-0.212***	0.506***	-0.286***	-0.110*	0.354***	-0.124***	-0.226***	0.322***	-0.022
	(0.019)	(0.057)	(0.020)	(0.052)	(0.072)	(0.024)	(0.057)	(0.055)	(0.023)	(0.064)	(0.048)	(0.018)
CTAS 5	-0.250***	0.566***	-0.239***	-0.353***	0.527***	-0.316***	-0.120	0.606***	-0.220***	-0.269***	0.515***	-0.039**
	(0.028)	(0.081)	(0.028)	(0.103)	(0.083)	(0.046)	(0.077)	(0.089)	(0.038)	(0.076)	(0.069)	(0.018)
Gender (Base=Female)												
Male	-0.020	0.030***	-0.015**	0.000	-0.009	-0.009	-0.082***	0.063***	-0.021**	0.033	0.030**	0.000
	(0.013)	(0.006)	(0.006)	(0.020)	(0.007)	(0.008)	(0.027)	(0.012)	(0.009)	(0.029)	(0.012)	(0.009)
Hospital (Base=ED A)												
ED B	0.115***	-0.302***	-0.144***	0.101***	-0.333***	-0.158***	0.179***	-0.330***	-0.146***	0.069**	-0.256***	-0.129***
	(0.019)	(0.082)	(0.029)	(0.027)	(0.082)	(0.030)	(0.024)	(0.088)	(0.030)	(0.029)	(0.083)	(0.035)
TriageTime	0.005	-0.034***	0.063***	0.010	-0.034***	0.065***	-0.004	-0.034***	0.056***	0.037**	-0.039***	0.039***
	(0.007)	(0.006)	(0.004)	(0.011)	(0.007)	(0.004)	(0.010)	(0.006)	(0.005)	(0.015)	(0.009)	(0.006)
Workload	0.012*		0.024***	0.024**		-0.001	0.020*		0.031***	-0.029***		0.044***
	(0.007)		(0.009)	(0.010)		(0.007)	(0.012)		(0.010)	(0.011)		(0.013)
AvgOccTreated	0.032***		0.385***	-0.017		0.361***	0.023*		0.414***	0.088***		0.374***
	(0.008)		(0.008)	(0.015)		(0.012)	(0.012)		(0.009)	(0.011)		(0.011)
WaitTime	0.056***	0.079***		0.031***	0.094***		0.088***	0.083***		0.040***	0.054**	
	(0.007)	(0.020)		(0.009)	(0.018)		(0.011)	(0.018)		(0.014)	(0.024)	
Census		0.338***			0.268***			0.333***			0.415***	
		(0.046)			(0.037)			(0.049)			(0.056)	
N	123,655	146,481	142,683	41,171	48,404	47,122	41,701	49,858	48,772	40,783	48,219	46,789
Pseudo R^2	0.404	0.463	0.247	0.187	0.235	0.105	0.335	0.365	0.194	0.332	0.346	0.212

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

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

Appendix B. Hypotheses

The FT area is a separate ED area designed to offer dedicated pathways to streamline patient flow and alleviate congestion. Nonetheless, if routing decisions are solely based on clinical conditions, it can result in a mismatch between demand and supply in both the FT and main ED areas, leading to operational inefficiencies. This potential imbalance in workload between the two treatment areas can become especially pronounced in congested systems due to significant demand fluctuations. Consequently, triage nurses may opt to take into account the operational conditions of the ED when making FT routing decisions, aiming to align healthcare resources more effectively with patient demands. Therefore, we propose the following hypothesis:

HYPOTHESIS 1. *FT routing decisions may not be purely clinical-driven, and operational factors related to ED congestion could also be crucial to the decisions made.*

Previous medical studies have reported the implementation of the FT area in enhancing the ED operational efficiency, as demonstrated in works such as Sanchez et al. (2006), Ieraci et al. (2008), Devkaran et al. (2009), Chrusciel et al. (2019), and Grant et al. (2020). Consequently, we anticipate direct improvements in key ED efficiency metrics, such as LOS and LWBS. We formally present our hypotheses below:

HYPOTHESIS 2. *Being routed to the FT area improves patient care efficiency.*

Given the efficiency improvement, one might naturally expect that the adoption of the FT would also lead to improvements in healthcare quality, such as the reduction in patient revisits, assuming that the FT routing decision consistently selects the "right" patients for treatment in the FT area. However, as per Hypothesis 1, patients with similar clinical conditions may receive treatment in different areas of the ED based on varying ED congestion levels. As a result, it remains uncertain whether being routed to the FT will have adverse effects on patients. Given the diverse complexity levels of patients routed into the FT area, we anticipate that the impact of FT routing decisions may vary depending on the complexity of the patient's condition.

We detail our method for categorizing patients into low-, medium-, and high-complexity groups in Section 4.4. Patients categorized under high- and medium-complexity groups typically have intricate medical conditions that often require thorough diagnostic assessments and may lead to hospital admission for extensive treatment. Routing such patients to the FT area, which might occur due to factors like ED congestion and operational considerations, could potentially result in inadequate care and increased 48-hour revisits. As noted in Hampers et al. (1999), it was observed that the FT area in the pediatric ED ordered fewer tests compared to the main area of the pediatric ED, a difference that could not be attributed to patient characteristics. This raises concerns about the adequacy of diagnosis and care for complex patients in the FT area, which could potentially result in adverse outcomes. On the other hand, low-complexity patients, who are the primary target for treatment in the FT area, typically present with minimal medical complexity

and are likely to be discharged. Therefore, we anticipate that receiving treatment in the FT area will not lead to adverse outcomes for these patients. To articulate our expectations more formally, we present our hypotheses below:

HYPOTHESIS 3 **(High and Medium-Complexity Patients)**. *High and medium-complexity patients will have worse quality outcomes (i.e., increased 48-hour revisit rates) if routed to the FT area rather than the main ED area.*

HYPOTHESIS 4 **(Low-Complexity Patients)**. *Low-complexity patients will have no worse quality outcomes if routed to the FT area rather than the main ED area.*

Appendix C. Details on Robustness Checks

In this section, we describe our robustness checks in detail.

Appendix C.1. Alternative IVs We consider several alternative IVs in this robustness check. First, in our main model, the IV measures the relative congestion level between the main area and the entire ED. As explained in Section 4.2, the congestion level in a particular area is computed as the area workload divided by the area capacity, where the area workload is the number of patients in that area, observed directly by triage nurses. However, due to varying physician shift schedules, the number of physicians working in an area may differ throughout the day. To obtain an area workload that is independent of the scale of the supply side, we adjust for the number of physicians on duty and consider an alternative IV in this robustness check that takes into account the number of physicians. Following this idea, the new area workload is computed as the total number of patients waiting and receiving treatment in a specific area divided by the number of physicians on duty at that time. Using the new area workload measure, we can compute the revised area congestion measures. Subsequently, we utilize these revised congestion measures to compute an alternative IV, which again assesses the relative congestion level between the main area and the entire ED. Table 18 in the supplementary document shows the estimation results using this alternative IV, which are consistent with our main results.

Next, in this robustness check, we consider an alternative IV by directly using *FTCongestion*, which may provide a more precise representation of the congestion effect. The results, as shown in Table 19 in the supplementary document, remain consistent. It is worth noting that, as evidenced in Table 6 in Appendix A, *FTCongestion* is the only congestion measure that does not exhibit a statistically significant correlation with key patient severity measures, potentially making it a valid IV.

Finally, the IV used in our main analyses is computed at the triage start time of the focal patient. However, triage nurses may use past congestion information to inform current routing decisions. Therefore, we also consider alternative IVs computed using information from 0.5, 1, and 2 hours before patient i's triage start time, respectively. Tables 20–22 in the supplementary document show the estimation results with these alternative IVs. We find all the results are consistent with our main findings.

Appendix C.2. Additional or Alternative Control Variables First, it is natural to expect that patient comorbidities may affect patient outcomes and be associated with complexity classes. Hence, we conduct a robustness check by controlling patient comorbidity information. However, our dataset does not contain a numerical measure of patient comorbidity that can be directly incorporated into our econometric model. Therefore, we use the textual medical history data collected by triage nurses to construct the Charlson comorbidity index (Charlson et al. 1987) as a control variable for patient comorbidity. Since the medical history data only covers 11 months of our study period (from September 2014 to July 2015), we decided to use this information only in the robustness check instead of including in our main analysis. Specifically, we first include the Charlson comorbidity index ($Charlson_i$) in the patient classification model discussed in Section 4.4 and then incorporate it into Equations (2), (4), (5), and (7) of our main empirical analyses. The results shown in Table 23 of the supplementary document are consistent with our main findings.

Next, we consider an alternative way to calculate physician workload. Earlier work (e.g., KC and Terwiesch 2009) has shown that increased physician workload leads to physician behavioral changes that might negatively affect patient outcomes. As a result, we control physician workload in our estimation. One important physician behavior or decision that might affect the entire treatment period is the treatment plan, such as diagnostic tests. These plans are usually determined during the initial patient encounter, which occurs immediately after the patient assignment. Therefore, in our main analyses, we calculate physician workload based on the number of patients assigned to physician p_i who were still present in the ED at the time of patient i's assignment. We now consider an alternative way to reflect the number of patients receiving care from physician p_i during the treatment process of patient i. This idea is similar to the way we calculate the area occupancy ($AvgOccTreated_i$). Table 24 in the supplementary document presents the estimation results, which are consistent with our main findings.

Lastly, previous literature indicates that ED congestion may impact hospital admission decisions (Gorski et al. 2017, Chen et al. 2020, Freeman et al. 2021). To address this concern and ensure the robustness of our main findings, we have conducted an additional analysis that controls for the ED congestion at the time when the attending physician made the disposition decision for patient i, denoted as $EDCongestion_i^d$. Table 25 in the supplementary document shows consistent results as our main findings, further supporting the validity of our main results.

Appendix C.3. Alternative Patient Classification Cutoffs We then explore alternative cutoff values to classify patients into complexity classes. In our main analyses, the cutoffs t_1 and t_2 are set at the 33rd and 67th percentiles, respectively, of the hospital admission likelihood distribution to divide the patient groups into three equal sized categories. To show the robustness of our findings, we consider alternative pairs of thresholds for (t_1, t_2), specifically the (28th, 67th), (38th, 67th), (33rd, 62nd), and (33rd, 72nd) percentiles. The estimation results presented in Tables 26–29 in the supplementary document generally align with our

main findings, with minor differences in the significance levels. Specifically, when using the 28th percentile as the threshold for t_1, we observe a consistent impact on 48-hour revisits for the medium complexity group with a positive sign, although this effect does not reach statistical significance.

Appendix C.4. Alternative Outcome Measure In addition to the 48-hour revisit used in our main analyses, the 72-hour revisit has also been used in prior studies to measure the patient outcome (Batt et al. 2019). Hence, we use the 72-hour revisit as an alternative outcome variable and run our analyses again; see Table 30 in the supplementary document. We find that being routed to the FT increases 72-hour revisits for both high- and medium-complexity patients, which aligns with our main findings. This result further supports the potential hidden consequences on the quality of care associated with FT routing decisions.

Appendix C.5. Alternative Samples We consider three alternative data samples to assess the robustness of our findings. First, we construct an alternative sample by removing extreme observations with triage times longer than 17 minutes (i.e., outside the 99.9th percentile), as shown in Table 31 of the supplementary document. The estimation results are consistent with our main findings.

Second, in our main analyses, we exclude patients with triage level 1 due to their typically urgent conditions requiring immediate attention (Ding et al. 2019). To further test the robustness of our results, we run our analysis again by including these patients, and the results are presented in Table 32 of the supplementary document. Once again, we find the estimation results to be consistent with our main findings.

Third, we adopt a matching approach to create more comparable samples of patients treated in the main area as opposed to the FT area, which emulates a randomized experimental study design (see Rubin 2006). Specifically, we match patients from the main and FT areas based on their propensity to be routed to the FT area, thereby reducing biases from observable characteristics. The estimation results in Table 33 of the supplementary document are generally consistent with our main findings, despite slight differences in significance levels. Specifically, we observe a consistent impact on 48-hour revisits for the medium complexity group with a positive sign, although this effect does not attain statistical significance.

Appendix C.6. Alternative Model Specifications Given the complex dynamics involved in patient treatment, we revise our empirical estimation to include physician fixed effects. This addition aims to capture the variability in physician practices and their potential impact on patient outcomes. Additionally, we include time-of-day fixed effects, taking into account the fluctuations in patient demand and service capacity throughout the day. The results are presented in Table 34 of the supplementary document. We again find consistent results.

Next, since previous work has shown that ED congestion (Gorski et al. 2017, Chen et al. 2020, Freeman et al. 2021) and physician workload (Gorski et al. 2017) may affect hospital admission decisions. We conduct another robustness check to incorporate the ED congestion at the time when the attending physician makes the disposition decision for patient i ($EDCongestion_i^d$) as well as physician workload ($Workload_i$) into the

patient classification model (Equation (9)). The results shown in Table 35 in the supplementary document demonstrate consistency with our main results.

Appendix D. 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 at the ED according to a time-homogeneous Poisson process with arrival rate λ_i, where $i = 1, 2, 3$, representing patients of high-, medium-, and low-complexity classes as defined in Section 4.4, respectively. We are aware that a nonstationary Poisson process with time-dependent arrival rates is a better model for the patient arrival process (Kim and Whitt 2014). We make the stationary assumption to simplify the model and have relaxed it in our simulation. Each station has a single server (which is relaxed to multiple shift-based servers in our simulation model) and a queue with infinite capacity. At station j, the service times (*diagnosis and treatment time*) are i.i.d. exponentially distributed with rate $\mu_j,\ j = 1, 2$, (again, this assumption is relaxed in the simulation). Patients are served on a first-come-first-served (FCFS) basis at each station. We are aware that ED decision-makers do not always adhere to the FCFS rule in real settings (Ding et al. 2019). Hence, in our simulation, the process of selecting the next available patient to treat is formulated by a discrete choice model whose parameters are estimated from our data.

Upon arrival, patients are routed to one of the two queues by the decision maker (i.e., triage nurses), waiting to be seen. If a patient of class i is routed to queue j, a cost $r_{ij}(t)$ is incurred upon the completion of service at station j, where $i = 1, 2, 3$, $j = 1, 2$, given that the patient waited t units of time in the queue before being seen by a physician. The cost reflects the inconvenience and quality of care if a patient needs to revisit the ED within a short period of time (e.g., 48 hours) after being discharged from the ED. 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 3). The dependence on the patient's waiting time, as shown by our empirical results (see Table 10 in Appendix A), aligns with the literature (Guttmann et al. 2011). Note that the dependence of the cost term on a patient's characteristics (e.g., age, gender) is reflected by the patient's class. In our simulation study, we explicitly account for patient characteristic information when estimating the cost term $r_{ij}(t)$. The decision maker's objective is to find a routing policy to minimize the expected long-run average cost over an infinite time horizon. Later, we will formulate the decision problem for FT routing as an MDP or constrained MDP and solve them numerically. See Appendix E and Appendix F for details.

Appendix E. The MDP Formulations

We formulate the decision problem for FT routing using an MDP formulation. The decision epochs correspond to patient arrival times to the ED. Denote the system state at time t by $\boldsymbol{x} = (x_1, x_2)$, where x_1 and x_2 represent the number of patients in the main and the FT area, respectively. Hence, the state space is

$\mathcal{S} \equiv \{\boldsymbol{x} = (x_1, x_2) : x_i \in \mathbb{N}, i = 1, 2\}$. Upon the arrival of a new patient, the triage nurse needs to decide which area to route this patient to after triage. Hence, the action space is $\mathcal{A} \equiv \{0, 1\}$, where 0 and 1 represent routing the patient to the main and the FT area, respectively.

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

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

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

The relatively low dimension of the MDP allows us to solve the MDP by the value iteration algorithm. The arrival rates and service times are estimated from data under the stationary assumption. It is however challenging to estimate the cost terms $r_{ij}(t)$, $i = 1, 2, 3,\ j = 1, 2$. We leverage the results of our econometric model for the binary patient outcome variable to estimate the 48-hour revisit cost for a class i patient with characteristics $\mathbf{X}$ who joins queue j and waits t units of time before being seen by physicians as $r_{ij}(t) = \mathrm{E}\left(Revisit_i | FT_i = j, \mathbf{X}\right) = \mathrm{P}\left(\xi_i \geq -\beta_i \mathbf{X} - \mathbb{1}_{\{j=2\}}\gamma_i - h_i t\right)$, where γ_i is the coefficient of FT_i estimated from Equation (7), h_i is the cost per unit time a class i patient waits in the system, and ξ_i is the error term that follows a standard normal distribution based on the observed information from data. Note that for each class i patient, we compute costs associated with both $FT_i = 1$ and $FT_i = 2$ for the optimization problem.

Appendix F. The Constrained MDP Formulation

We consider a constrained MDP formulation by adding a constraint on the FT congestion based on an occupancy measure (Altman 2021). Specifically, let $\psi(\boldsymbol{x}, \boldsymbol{a})$ denote the stationary probability on each state-action pair $(\boldsymbol{x}, \boldsymbol{a})$, where $\boldsymbol{x}$ is defined in Appendix E and $\boldsymbol{a}$ is a tuple of binary variables, each determining the routing decision of patients from one of the three complexity classes. The action space is thus defined as $\tilde{\mathcal{A}} \equiv \{\boldsymbol{a} = (a_1, a_2, a_3) : a_i \in \mathcal{A}, i = 1, 2, 3\}$. The constraint optimal policy can be formulated as a linear program (LP) as follows.

$$\underset{\psi}{\text{Minimize}} \quad \sum_{\boldsymbol{x}\in\mathcal{S}} \sum_{\boldsymbol{a}\in\tilde{\mathcal{A}}} r(\boldsymbol{x}, \boldsymbol{a})\psi(\boldsymbol{x}, \boldsymbol{a})$$

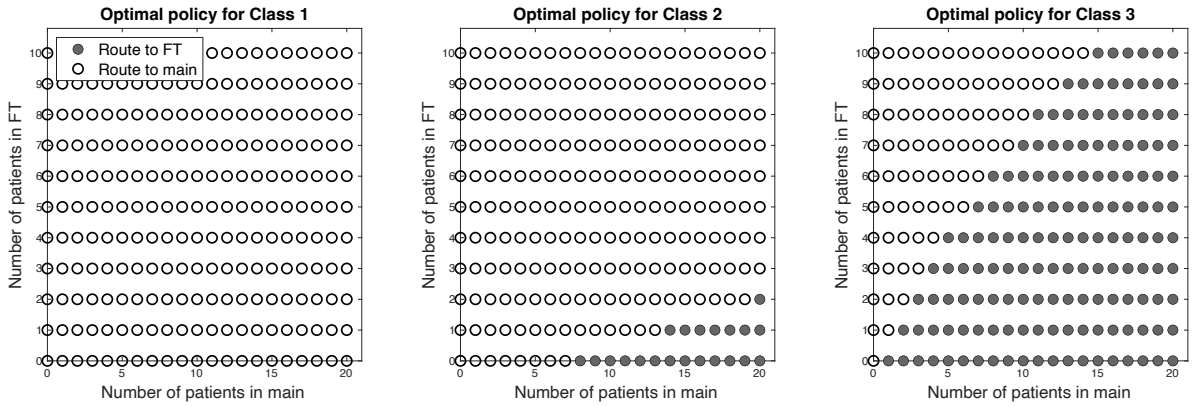
$$
\begin{aligned}
\text{subject to} \quad & \sum_{\boldsymbol{a}\in\tilde{\mathcal{A}}} \psi(\boldsymbol{y},\boldsymbol{a}) - \sum_{\boldsymbol{x}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\tilde{\mathcal{A}}} p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{a})\psi(\boldsymbol{x},\boldsymbol{a}) = 0, \quad \boldsymbol{y}\in\mathcal{S} \\
& \sum_{\boldsymbol{x}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\tilde{\mathcal{A}}} c(\boldsymbol{x},\boldsymbol{a})\psi(\boldsymbol{x},\boldsymbol{a}) \le \omega, \\
& \sum_{\boldsymbol{x}\in\mathcal{S}}\sum_{\boldsymbol{a}\in\tilde{\mathcal{A}}} \psi(\boldsymbol{x},\boldsymbol{a}) = 1, \\
\text{and} \quad & \psi(\boldsymbol{x},\boldsymbol{a}) \ge 0 \quad \text{for all } \boldsymbol{a}\in\tilde{\mathcal{A}}, \boldsymbol{x}\in\mathcal{S},
\end{aligned}
$$

where for a given state-action pair $(\boldsymbol{x},\boldsymbol{a})$, $c(\boldsymbol{x},\boldsymbol{a})$ is the stationary queue length in the FT area, ω is the upper bound we choose, and $r(\boldsymbol{x},\boldsymbol{a}) = \sum_{i=1}^{3}\lambda_i r_{i,a_i+1}(x_{a_i+1}/\mu_{a_i+1})$ is the immediate expected revisit cost. Furthermore, $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{a})$ specifies the transition probability from state $\boldsymbol{x}$ to state $\boldsymbol{y}$ when action $\boldsymbol{a}$ is taken. Specifically, for any $\boldsymbol{x},\boldsymbol{y}\in\mathcal{S}$ and $\boldsymbol{a}\in\mathcal{A}$, we have

$$
p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{a}) = \begin{cases}
\sum_{i=1}^{3}\lambda_i(1-a_i), & \text{when } \boldsymbol{y} = \boldsymbol{x}+\boldsymbol{e_1} \\
\sum_{i=1}^{3}\lambda_i a_i, & \text{when } \boldsymbol{y} = \boldsymbol{x}+\boldsymbol{e_2} \\
\mu_k, & \text{when } \boldsymbol{y} = \boldsymbol{x}-\boldsymbol{e_k},\ k\in\{1,2\} \\
1-\sum_{\boldsymbol{z}\in\mathcal{S},\boldsymbol{z}\neq\boldsymbol{x}} p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{a}), & \text{when } \boldsymbol{y} = \boldsymbol{x} \\
0, & \text{otherwise}
\end{cases}
$$

Finally, we solve the LP to obtain the optimal solutions $\psi^*(\boldsymbol{x},\boldsymbol{a})$. Then, the action $\boldsymbol{a}$ at state $\boldsymbol{x}$ is optimal if $\psi(\boldsymbol{x},\boldsymbol{a}) > 0$. Figure 4 illustrates the optimal policy for the constrained MDP problem, which shares the same structure as that of the unconstrained MDP except that patients are less likely to be routed to FT.

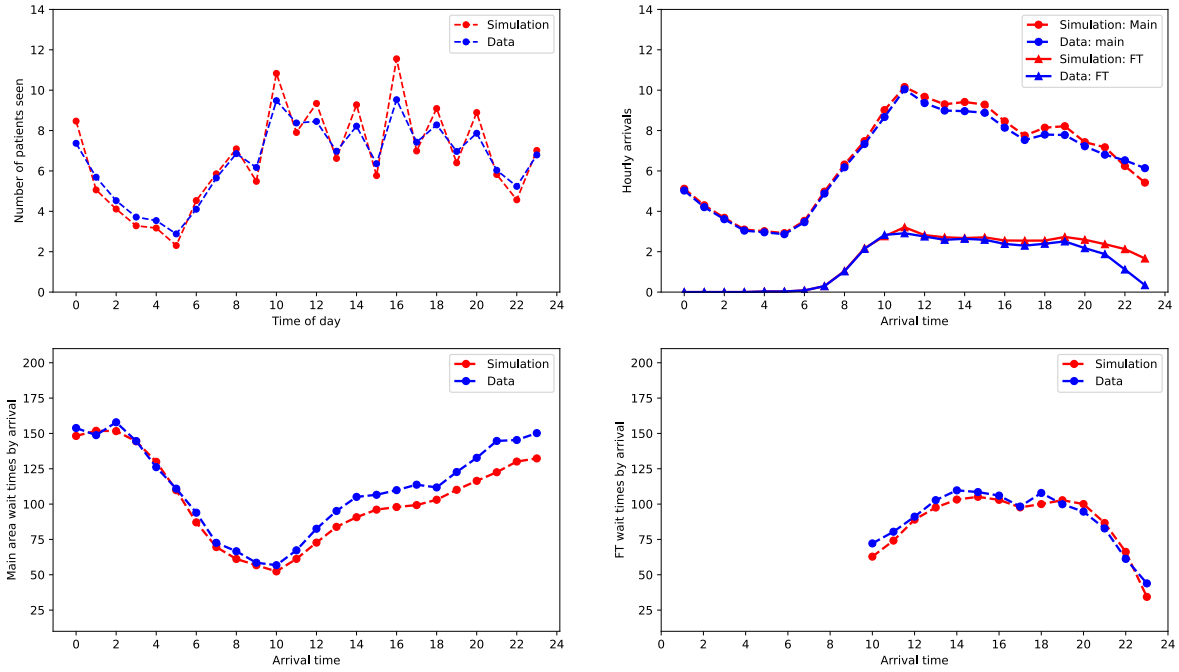
Figure 4 An illustration of the optimal policy for the constrained MDP (Policy COP) used in our simulation study.



Appendix G. Further Simulation Results

The average patient waiting times from the simulation and the data are shown in the bottom two panels of Figure 5 for the main and the FT areas, respectively, which provide evidence that our simulation model captures the trend of the average waiting time from the data reasonably well. We also compare the simulated

Figure 5 **Comparison of the number of patients seen per hour, the hourly number of patients arrived, and the average patient waiting times by the time of day between the simulated and the real data.**



number of patients seen by all physicians on duty per hour and the hourly arrivals with that from the data (shown in the top two panels of Figure 5), which further shows the validity of our simulation model.

Table 11 presents further simulation results on the comparison of FT routing policies. Specifically, we include the average queue length in the FT area, the average patient waiting time in hours by the treatment areas or by the triage levels.

Table 11 **Further simulation results on the average queue length at FT and the average patient waiting times.**

Routing policy	CP	TP	TP-C	SP	COP	OP
Average FT queue length	3.94 ± 0.02	5.68 ± 0.03	4.53 ± 0.02	3.83 ± 0.03	4.01 ± 0.03	5.22 ± 0.04
Average waiting time (hours)						
All patients	1.60 ± 0.01	1.55 ± 0.01	1.59 ± 0.01	1.65 ± 0.01	1.45 ± 0.01	1.37 ± 0.01
Patients in main area	1.62 ± 0.01	1.42 ± 0.01	1.54 ± 0.01	1.67 ± 0.01	1.43 ± 0.01	1.24 ± 0.01
Patients in FT area	1.51 ± 0.01	2.12 ± 0.01	1.78 ± 0.01	1.53 ± 0.01	1.51 ± 0.01	1.93 ± 0.01
Average waiting time (hours) by triage levels						
CTAS 2	1.40 ± 0.01	1.21 ± 0.01	1.34 ± 0.01	1.48 ± 0.01	1.23 ± 0.01	1.06 ± 0.01
CTAS 3	1.72 ± 0.01	1.55 ± 0.01	1.67 ± 0.01	1.78 ± 0.01	1.58 ± 0.01	1.49 ± 0.01
CTAS 4	1.67 ± 0.01	2.04 ± 0.01	1.78 ± 0.01	1.67 ± 0.01	1.53 ± 0.01	1.61 ± 0.01
CTAS 5	1.68 ± 0.01	2.05 ± 0.01	1.83 ± 0.01	1.69 ± 0.01	1.54 ± 0.01	1.64 ± 0.01