informs
http://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 21, No. 4, Fall 2019, pp. 723–741
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Patient Prioritization in Emergency Department Triage Systems: An Empirical Study of the Canadian Triage and Acuity Scale (CTAS)

Yichuan Ding,[a,b] Eric Park,[c] Mahesh Nagarajan,[a] Eric Grafstein[d]

[a] Sauder School of Business, University of British Columbia, Vancouver, British Columbia V6T 1Z2, Canada; [b] School of Information Management and Engineering, Shanghai University of Finance and Economics, 200083 Shanghai, China; [c] Faculty of Business and Economics, University of Hong Kong, Pokfulam, Hong Kong, China; [d] Providence Health Care and Vancouver Coastal Health, Vancouver, British Columbia V6E 3V6, Canada

Contact: daniel.ding@sauder.ubc.ca, http://orcid.org/0000-0003-3014-8973 (YD); ericpark@hku.hk, http://orcid.org/0000-0002-3331-3164 (EP); mahesh.nagarajan@sauder.ubc.ca, http://orcid.org/0000-0001-6942-4627 (MN); eric.grafstein@vch.ca, http://orcid.org/0000-0003-3327-7615 (EG)

Received: September 8, 2016
Revised: August 21, 2017; December 31, 2017
Accepted: January 12, 2018
Published Online in Articles in Advance: April 8, 2019

https://doi.org/10.1287/msom.2018.0719

Abstract. Emergency departments (EDs) typically use a triage system to classify patients into priority levels. However, most triage systems do not specify how exactly to route patients across and within the assigned triage levels. Therefore, decision makers in EDs often have to use their own discretion to route patients. Also, how patient waiting is perceived and accounted for in ED operations is not clearly understood. In this paper, using patient-level ED visit data, we structurally estimate the waiting cost structure of ED patients as perceived by the decision makers who make ED patient routing decisions. We derive policy implications and make suggestions for improving triage systems. We analyze the patient routing behaviors of ED decision makers in four EDs in the metro Vancouver, British Columbia, area. They all use the Canadian Triage and Acuity Scale, which has a wait time–related target service level objective. We propose a general discrete choice framework, consistent with queueing literature, as a tool to analyze prioritization behaviors in multiclass queues under mild assumptions. We find that the decision makers in all four EDs (1) apply a delay-dependent prioritization across different triage levels; (2) have a perceived marginal ED patient waiting cost that is best fit by a piece-wise linear concave function in wait time; (3) generally follow, in the same triage level, the first-come first-served principle, but their adherence to the principle decreases for patients who wait past a certain threshold; and (4) do not use patient complexity as a major criterion in prioritization decisions.

Funding: The research of Y. Ding and M. Nagarajan is partially supported by the National Sciences and Engineering Research Council of Canada (NSERC) [Grants PGPIN 436156-13 and 13R81646, respectively]. E. Park is partially supported by The University of Hong Kong, Seed Funding for Basic Research [Grant 201611159008]. E. Grafstein thanks Vancouver Coastal Health for its generous support.
Supplemental Material: The online appendices are available at https://doi.org/10.1287/msom.2018.0719.

Keywords: empirical research • emergency department • dynamic priority • discrete choice • public policy • generalized $c\mu$ rule

1. Introduction

Emergency department (ED) overcrowding has been a prevalent issue for several decades in hospitals around the world (Graff 1999, Pines et al. 2011). The alarmingly overcrowded status in the U.S. healthcare system has been well documented (Derlet and Richards 2000). Canada, which operates a universal, publicly funded healthcare system, is no exception (Drummond 2002, Ospina et al. 2007). Overcrowding occurs when demand exceeds available capacity. Given the limited capacity of many EDs, some patients may experience excessive wait times, which can be critical—even life threatening (Bernstein et al. 2009). Combined with the fact that many ED patients require immediate service, prioritization of ED patients who are waiting for treatment is critical to the performance of the ED and, consequently, public health. Hence, most EDs around the world use a resource allocation plan known as "triage," which provides guidelines for classifying patients into priority groups (triage levels) based on their acuity, urgency, and resource needs. For EDs in Canada, the Canadian Triage and Acuity Scale (CTAS) (Beveridge 1998, p. 510) proposes a *fractile response objective*. As an example, CTAS guidelines state that "a level-2 patient should be seen within 15 minutes, 95% of the time," in addition to the classification guidelines (Table 1). However, some ED physicians view the fractile response as an operationally unreasonable objective because many EDs across the country are consistently facing high patient volume and operate at high utilization rates. Furthermore, most triage systems, including CTAS, do not provide explicit guidelines on how to route patients within the

Table 1. CTAS Fractile Response Objective

CTAS score (triage level)	Triage acuity	Target wait time	Fractile response objective
1	Resuscitation	Immediately	98%
2	Emergent	15 minutes	95%
3	Urgent	30 minutes	90%
4	Less urgent	60 minutes	85%
5	Nonurgent	120 minutes	80%

assigned triage levels.[1] Hence, in practice, ED decision makers[2] often have to use their own *discretion* in making patient-routing decisions rather than following predetermined rules, such as first come, first served (FCFS) within the same triage level and/or strict priority across different triage levels. In fact, we observe from the study data that, on average, FCFS is violated 39.7% and 52.8% of the time within triage levels 2 and 3, respectively, and triage level 3 patients are chosen over triage level 2 patients 57.1% of the time when at least one of each are present in the waiting area. This is an obvious distinction from some other commonly examined service systems. For example, in call centers, a well-studied service setting in the operations management literature, customer routing in most instances follows a given priority rule dictated by the system's objective. In the context of EDs, understanding the prioritization is especially important because EDs are often gatekeepers to the entire healthcare system, and the impact of prioritization can affect crucial operational measures, such as ED length of stay, which, in turn, has serious implications to patient outcomes, including complication and mortality rates.

The goal of this paper is to empirically infer, from patient-level ED visit data, the waiting cost structure of ED patients waiting for treatment as *perceived by the routing decision makers*. This allows us to understand how decision makers route patients in ED triage systems, a discretionary multiclass queuing system. We explore this decision in two dimensions, that is, first within the same triage level and next across different triage levels. We study how the ED decision makers account for patient waiting in the Canadian health system and relate their routing behavior to the policy design of CTAS. Understanding the ED decision makers' perceived waiting cost can benefit both local ED management and global triage policy designs, including CTAS and other triage systems. Moreover, ED administrators can compare the waiting cost perceived in practice with social, clinical, or ethical (such as fairness) expectations and revise the current operation guidelines if needed. If the current operations meet expectations, policy designers can use them as benchmarks for improving the operations in other EDs. Otherwise, by identifying existing issues in the current practice, the policy designer can better determine the future direction for policy refinement.

We study the ED decision makers' patient-routing behavior in more than 186,000 ED patient admissions from April 2013 to November 2014 in the four largest EDs in the metro Vancouver, British Columbia, area. We model the ED in which patients are waiting to see a physician as a multiclass queueing system and investigate how decision makers choose which patient gets to be seen by the next available physician. We observe a few important properties of this queueing system. With those properties, the ED decision makers' routing behavior follows the generalized $c\mu$ rule proposed by Van Mieghem (1995). We estimate the decision makers' perceived ED patient marginal waiting cost (i.e., the extra cost of waiting for another minute) from the observed routing decisions using a discrete choice framework (McFadden 1973). Our framework also allows us to test whether, to improve ED operations, the decision makers incorporate patients' complexity information into their patient routing decisions, which has been suggested by Saghafian et al. (2014). By estimating the cost functions, we find that the routing decisions have the following features, which are robust for all four EDs we studied:

1. The ED decision makers' patient-routing behavior is best fit by a piece-wise linear concave marginal waiting cost function for each triage level. In particular, we find that the marginal waiting cost has a significantly positive slope below the point at which the slope changes (break point) but is nearly *constant* above the break point for most triage levels. The order of estimated break points across triage levels is identical to the order of CTAS triage-level target wait times, and the estimated break point values are close to the target wait times. This implies that the CTAS fractile response objective may be one of the drivers for the flattening in marginal cost curves. The shape of the marginal waiting cost function implies the next two features.

2. Routing behavior within triage levels: ED decision makers generally follow the FCFS principle within the same triage level, but their adherence to FCFS decreases among patients who wait past a certain threshold (break point).

3. Routing behavior across triage levels: ED decision makers apply a delay-dependent prioritization (also called dynamic prioritization by Jackson 1960) across different triage levels in the respective patients' wait times. Generally, higher triage level (e.g., triage level 2)

patients receive priority over lower triage level (e.g., triage level 3) patients. However, a lower triage level patient who has waited longer can be prioritized over a higher triage level patient who has waited less time.

4. Overall, there is no strong evidence that current ED patient routing is based on the service (treatment) time of a patient anticipated by the decision maker.

These findings have important implications for designing prioritization rules in ED triage systems. First, our work points out an important but debatable consequence of the CTAS objective: among patients who have waited past a certain target, those who have waited longer may not receive extra priority, which, at the least, does not meet the conventional fairness standard of FCFS (Iserson and Moskop 2007). In other words, the CTAS target wait time structure may lead to unjustifiably prolonged waits for some patients. This is an interesting behavioral observation, as the CTAS, despite being well advocated in the Canadian medical community, was not imposed by strict adherence rules nor penalty mechanisms in any of the four study EDs. Second, we find that the decision makers apply a delay-dependent prioritization across triage levels. Evidence of a sophisticated prioritization behavior in practice suggests that it would be worthwhile to explore implementing such prioritization rules into triage system guidelines. We highlight the need to consider not just the assigned patient's triage level, but also the patient's actual wait time in routing decisions. Finally, given that the current prioritization may depend solely on the urgency (wait time) of a patient but not on the complexity of service (treatment time), we believe that ED operations can be improved by incorporating both the complexity and urgency information of the patients into the routing decisions. Doing that may require the involvement of physicians in the prioritization decisions as physicians generally have better knowledge about the complexity of the patient treatment process than nonphysician decision makers.

Although our empirical findings apply immediately to ED operations, our proposed structural estimation framework can analyze prioritization behaviors in other service systems that share the following features with EDs: (1) the service provider's objective is not driven by revenue or other explicit measurements, but by less definable goals, for example, social welfare, and (2) Prioritization guidelines are either absent or not detailed enough because of the complexity of the system so that the service providers have to rely on their own discretion when making prioritization decisions. Extensive examples exist in the public sector: government offices, nonprofit hospitals and public healthcare systems, and nongovernmental organizations. Immigration officers, for example, when facing a large backlog of immigrant or visa applications, have to select certain cases to expedite. Likewise, when managing operating rooms, hospitals have to sequence surgeries based on the physician's judgment of patient urgencies among other factors. Our framework provides a tool to understand how these service systems value the waiting costs of customers. To our knowledge, this is the first attempt to study the waiting cost structure perceived by the service provider but not the customers themselves.[3]

2. Literature Review

There are two streams of literature that are closely related to this paper. The first is the multiclass queueing literature, and the second is the literature on ED operations. We review relevant papers in the following sections.

2.1. Scheduling in Multiclass Queues

Our work is closely related to the extensive literature on queuing and job scheduling that explores optimal scheduling policies under different waiting cost structures. When the cumulative waiting cost is a linear function of the sojourn time $W_k(t)$, that is, $C_k(W_k(t)) = c_k W_k(t)$, the well-known prioritization scheme, $c\mu$ rule, is to prioritize queues with a larger value of $c_k\mu_k$ and to use the FCFS rule within each queue (Smith 1956). When the cumulative holding cost $C_k(W_k(t))$ is a nondecreasing convex function, Van Mieghem's (1995) seminal paper shows that the generalized $c\mu$ rule, in which jobs are prioritized according to the order of $C_k'(W_k(t))\mu_k$, minimizes average waiting costs under the heavy traffic asymptotic regime. Van Mieghem's (1995) result does not require stationarity of the arrival process and is robust when there are a few homogeneous servers and countably many job types. Mandelbaum and Stolyar (2004) and Gurvich and Whitt (2009) have studied the queue-length version of the $Gc\mu$ rule, in which the holding cost $C_k(\cdot)$ is a differentiable function of the queue length $Q_k(t)$ instead of $W_k(t)$.

The $Gc\mu$ rule subsumes several classes of scheduling policies as the waiting costs can take various forms. For example, when the waiting cost is a quadratic function of the queue length, that is, $C_k(Q_k(t)) = \beta_k(Q_k(t))^2$, the $Gc\mu$ rule is reduced to a well-known MaxWeight policy in which a server s always serves queue k with the largest index $\beta_k Q_k(t)\mu_{ks}$ at time t (Tassiulas and Ephremides 1992). The $Gc\mu$ rule also applies to scenarios in which jobs face timing requirements, such as laxities and deadlines (Hong et al. 1989). The former requires a job to start service by a specified time, and the latter imposes a due time for service completion. Let W and τ denote the time that a job remains in the system until the beginning of service and until the end of service, respectively. We use d_k and D_k, respectively, to denote the laxity and deadline of a job in queue k relative to its arrival time. There are four cost structures

that can arise from these measures: (a) the expected tardiness with respect to laxities, $\mathbb{E}(W-d_k)^+$; (b) the expected tardiness with respect to deadlines, $\mathbb{E}(\tau-D_k)^+$; (c) the proportion of jobs that violate the laxity constraints, $\Pr(W>d_k)$; and (d) the proportion of jobs that violate the deadline constraints, $\Pr(\tau>D_k)$. Because cost structures (a) and (b) are both nondecreasing convex, the $Gc\mu$ rule asymptotically minimizes cost functions (a) and (b). When the cost structure is of type (d), Van Mieghem (2003) proved that the generalized longest queue (GLQ) or the generalized largest delay (GLD) policy both asymptotically minimize (d) in heavy traffic among the class of work-conservation policies, and both GLQ and GLD can be regarded as special forms of the $Gc\mu$ rule. Cost structure type (c), which might be close to the CTAS fractile response objective, has not been well-studied in the literature of $Gc\mu$ rules.

Our study contributes to this stream of queueing literature by providing an empirical understanding of the possible objective functions that are used in scheduling multiclass patients in typical Canadian EDs. This may open the door for important theoretical work and subsequent empirical studies on scheduling multiclass jobs.

2.2. ED Operations

The ED as a general application has gained significant attention in the operations management (OM) literature in recent years, for example, Kc (2013) and Batt and Terwiesch (2016). The question of how one should route patients in EDs has been studied under different objectives. Dobson et al. (2013) have looked at ED throughput, and Huang et al. (2015) have examined violation of laxity constraints. Saghafian et al. (2012) probed into the question of whether streaming ED patients based on predictions of whether they would be discharged or admitted to the hospital could improve ED performances. Helm et al. (2011) proposed an "expedited patient care queue," an alternative hospital access gateway to the two conventional gateways, ED and scheduled elective admission, as a solution to mitigate ED crowding and blockage. Our work complements this stream of literature by studying the empirical counterpart of patient routing in EDs.

Several other papers have examined ED management from a capacity design perspective. Hu and Benjaafar (2009) studied the partitioning of ED capacity as an alternative to patient prioritization with a pooled capacity. Song et al. (2015) found that average ED patient wait times and length of stay are longer in a queueing system in which physician capacity is pooled compared with a system in which physicians are dedicated to their own stream of patients.

Empirically, Batt and Terwiesch (2015) studied the patient's side of ED operations on how ED congestion and queueing behavior affect patient abandonment in ED triage systems. To our knowledge, we are among the first to empirically study the control side of routing in multiclass queues regardless of application. We refer to Saghafian et al. (2015) for an overview of ED operations literature.

Several papers specifically discuss ED operations in the Canadian health system. Stanford et al. (2014) studied the wait time distribution in time-dependent priority queues in a single server setting. Sharif et al. (2014) generalized the result of Stanford et al. (2014) to a multiserver setting but with treatment time distributed with the same mean for all classes. Both provide a stepping-stone for better managing EDs subject to the CTAS fractile response objective. Our work complements both studies by providing empirical insights into how practitioners respond to the CTAS objective structure.

3. Study Setting and Data

3.1. CTAS

In the mid-1990s, the Canadian Association of Emergency Physicians (CAEP) recognized that, despite EDs being the interface between emergent care and the community, the Canadian health system had invested little to evaluate how ED case mix or changes to care delivery affected patients seeking emergent care. CAEP determined that it was important to standardize the processes and definitions of care for emergency medicine (Beveridge 1998). As a result, the CTAS was introduced in 1998 as an attempt to define patients' needs for timely care more accurately and to allow an ED "to evaluate its resource needs and performance compared to predefined objectives" (Beveridge 1998, p. 508). The CTAS guidelines state that "the primary operational objective of the triage scale is to define the optimal time to see a physician," and each triage level is given a fractile response objective (Table 1). The guidelines note that "the time responses are ideals (objectives) not established care standards" (CAEP 2014, section 1). The rationale behind this is that "the fractile response is a way of describing how often a system operates within its stated objectives."

Most triage systems, such as the Manchester Triage System (MTS) in the United Kingdom and Germany and the Emergency Severity Index (ESI) in the United States, focus on how to *classify* patients into multiple triage levels but do not provide guidelines on how to *prioritize* patients given their triage levels. In the United States, the general expectation is that the most urgent (or potentially most serious) cases will be treated first, followed by less urgent cases, and that urgent cases will be treated equally on a FCFS basis (Iserson and Moskop 2007). The fractile response objective distinguishes CTAS from other triage systems in that it incorporates specific time-based objectives. Since the CTAS was

initially proposed, it has faced intense criticisms and undergone a number of updates and revisions (Murray et al. 2004; Bullard et al. 2008, 2014). A major criticism is that the fractile response objective specified by the CTAS was set mainly for clinical reasons without considering the operational obstacles. Given the excessive demand and limited capacity in most EDs in Canada, the fractile response objectives are most likely not achievable regardless of the prioritization rules. This brings up the question central to our research; that is, how do ED decision makers prioritize patients in the absence of explicit guidelines?

In all four study EDs, neither a financial incentive nor penalty mechanism to induce ED decision makers to meet the CTAS fractile response objective was implemented. However, the CTAS fractile response objective may have still affected the decision makers' behavior in two aspects. First, CTAS, despite being considered inoperable at most times, has been widely advocated as the general principle for patient prioritization in Canadian EDs. Thus, it could have a psychological impact on the decision makers' prioritization behavior. Second, the fractile response is a mandated reporting data element by the Canadian Institute of Health Information, and the performance of each ED can be obtained through the publicly available National Ambulatory Care Reporting System metadata. Hence, the ED decision makers wary of public perception of ED performance may have the incentive to meet the CTAS targets, but the degree to which the decision makers are incentivized to adhere is not clear. Therefore, our empirical analysis reveals to what extent such a soft and flexible fractile objective (CTAS) influences decision makers' behavior.

3.2. Clinical Setting and Data

We analyze ED patient registration data from the four largest EDs in the metro Vancouver, British Columbia, area, which had a population of 2.4 million in 2011. The four study EDs cover a wide range of demographics and hospital types: the flagship hospital of the Vancouver healthcare system, which also serves as the primary trauma center of the metro Vancouver area; a large teaching hospital located near the city center; and two suburban hospitals located in a mainly residential district and a residential/commercial mixed district. The average daily traffic in these EDs ranges from 142 to 243 patients. All four EDs used the CTAS guidelines during the 20-month study period from April 2013 to November 2014. The data are at the patient visit level at which each observation corresponds to a single patient visit to one of the four EDs. For each patient visit, we have three important time stamps: (1) *enter time*—time of entry to the ED at which time the patient is triaged and registered; (2) *selection time*—time when the patient is first selected to enter the treatment area and see a physician; and finally, (3) *exit time*—when the patient is discharged from the ED after completion of treatment.

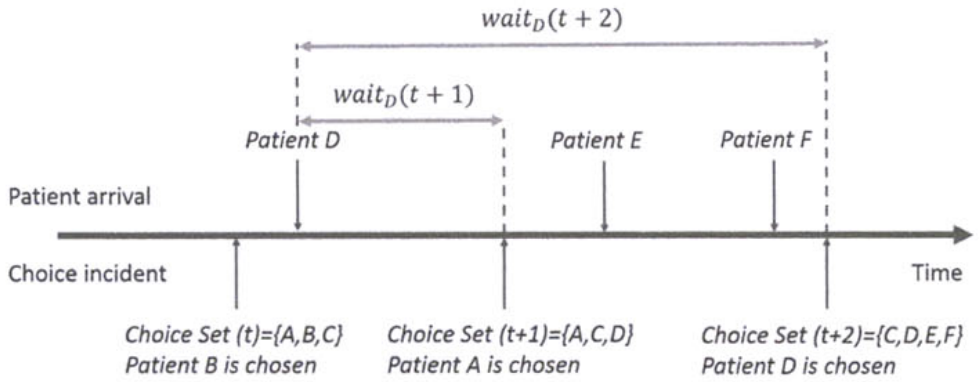
Most patients arrive to the ED either by emergency medical transportation, such as an ambulance, or by their own mode of transport referred to as "walk-in" patients. Regardless of the arrival mode, once a patient arrives, the triage nurse diagnoses the patient as soon as possible and identifies the most appropriate medical code, which has a default triage level. The standardized medical code is described in detail under Control variables. The nurse then inputs the patient information (such as name, age, sex, and personal health number), complaints, medical code, triage level, and other available information into the patient care information system (PCIS). Patients then wait either in the waiting room or on a stretcher-chair if deemed necessary by the triage nurse.

The decision process of when to initiate treatment of a new patient is primarily dictated by the availability of physicians but not beds and seats. The practice in all four study EDs is that physicians are typically involved in multiple tasks and do not have to wait for beds and seats; that is, physicians are the bottleneck resources in the treatment process. And the "call in" is triggered by the time when the physician finds that the physician has enough bandwidth to start accommodating another patient. Then the decision of which patient to treat with that opened bandwidth is made by the decision maker based on the information in the PCIS, which includes the patients' triage level, enter time, and other information.

The dependent variable in our study is whether the patient was chosen at a choice incident when a physician became available to accommodate a new patient. Choice incidents are chronologically ordered according to their selection times. Although the patient may not immediately see a physician and start treatment after being chosen, each choice still reflects different patients' relative priority as perceived by the decision maker. Hence, we use the selected time as a proxy for when prioritization decisions are made. Figure 1 provides a graphic illustration of choice incidents and how the choice sets evolve with time. At choice incident t ($t = 1, 2, \ldots$), only one patient is chosen from those waiting in the ED, which we denote as choice set, $ChoiceSet(t)$. $ChoiceSet(t)$ comprises patients who are currently waiting in the ED, that is, those who entered before choice incident t but were not chosen in any previous choice incidents or were not present in any of them. Patients not chosen remain in choice incident $t + 1$ and comprise the choice set $ChoiceSet(t + 1)$ with the new arrivals, that is, those who arrived at the ED between choice incidents t and $t + 1$.

In all four EDs, triage level 4 and 5 patients have a dedicated fast-track service line that operates

Figure 1. (Color online) Choice Incidents and Evolvement of Choice Sets



separately from the primary service area for triage level 1, 2, and 3 patients. Because wait times for triage level 1, 2, and 3 patients are critical as they need the most urgent care and, therefore, the decision makers are more cautious in their routing, we focus on the decision makers' choices for the primary service patients only. Among those patients, triage level 1 patients are often seen on arrival, and the registration/input to the PCIS happens afterward. For this reason, the *enter time* and *selection time* of triage level 1 patients may not be accurately recorded. Because of the possibility of their arrival triggering a choice incident rather than a physician becoming available to accommodate a new patient, we exclude triage level 1 patients and focus only on triage level 2 and 3 patients in this study. Table 2 briefly summarizes the data for triage level 2 and 3 patients.

Independent variable: patient wait time—our key variable of interest is the patient's wait time in the ED before being selected to be treated. We measure a patient's wait time at choice incident t by the duration of time from registering in the ED system (*enter time*) to the time of being chosen (*selection time*). For example, in Figure 1, patient D has waited $wait_D(t+1)$ until choice incident $t+1$ but was not chosen. At choice incident $t+2$ when patient D was chosen, patient D's wait time was $wait_D(t+2)$. The longest wait time observed in the study EDs is 720 minutes.

Control variables: time-invariant (fixed) patient characteristics—EDs in the province of British Columbia, including the four study EDs, use a standardized hierarchical tree structure to record patients' medical conditions: 19 categories at the clinical department level recorded as "chief complaint system" (CCS) and 474 detailed medical conditions recorded as "chief complaint description" (CCD). The CCD differentiates medical conditions at several levels, which is a strength of our data and allows us to control patient characteristics at a granular level. A few examples of CCD codes include "abdominal pain, moderate pain, episodic vomiting, fever"; "allergic reaction, mild respiratory distress, mild facial/oral edema, extensive rash"; and "acute dizziness/vertigo, + other neurological symptoms > 6 hrs." Each CCD belongs to a single parent CCS. The process of assigning a triage level to a patient who just entered the ED starts by the triage nurse first identifying the most appropriate CCD from a menu of 474 possible conditions, which are

Table 2. Summary Statistics

	ED A		ED B		ED C		ED D	
	Tri 2	Tri 3	Tri 2	Tri 3	Tri 2	Tri 3	Tri 2	Tri 3
Wait time (min)	37.0	77.5	22.3	41.7	32.7	68.0	33.5	68.8
Fractile response	35.7%	26.8%	42.9%	46.9%	22.7%	25.9%	35.6%	28.7%
Service time (min)	418.8	287.8	407.0	250.5	436.1	268.2	444.1	247.8
Age (y)	54.9	52.1	49.3	46.2	55.0	48.7	51.3	46.6
Ambulance arrivals	49.2%	29.1%	46.9%	35.0%	34.8%	22.6%	26.1%	18.2%
Female	45.8%	52.4%	39.1%	43.7%	48.7%	53.3%	51.9%	53.0%
Census of triage 1, 2, 3 in ED when selected	43.1	41.6	29.5	28.7	28.8	27.6	22.8	22.5
Waiting census of triage 1, 2, 3 when selected	7.7	7.2	3.9	3.6	5.4	4.9	4.2	4.1
N	25,098	66,174	15,823	56,728	14,517	47,932	13,356	38,568
N (percentage)*	17.4%	45.8%	12.0%	43.2%	15.4%	51.0%	15.9%	45.8%

Notes. Means are shown except for fractile response (to target wait time), percentage of patients arrived by ambulance, and female patients. A full table of all five triage levels is available upon request. Tri, triage level.

*Percentage among all five triage levels.

shared by all four study EDs. Each CCD code has a default triage level but is subject to adjustment by the triage nurse depending on the patient's specific condition. We use triage level dummy variables to capture prioritization effects across different triage levels and include CCD codes to control heterogeneous medical conditions within each triage level. Because some CCD codes have low frequency and do not appear in all EDs, we use patients with the top 113 common CCDs, which cover 90% of triage level 2 and 3 patients in the four EDs. Other control variables include age group, sex, method of arrival (whether the patient arrived via ambulance—ground or air—or walked in), and discharge decision (whether the patient was discharged home, admitted to ward, or transferred to another facility). All control variables in our data have a finite discrete domain.

4. Model of Patient Choice in ED Triage Systems

4.1. The Conditional Logit-$Gc\mu$ Framework

We discuss four observations of the study ED systems that set the stage for the analytical tool we use in this study—the conditional logit-$Gc\mu$ framework.

Observation 1. When choosing a patient, the ED decision makers' objective is to minimize average cumulated ED patient holding cost.

During the study period in all four EDs, all ED personnel, including physicians, were compensated by a fixed salary for each shift. Hence, the possibility of selecting patients based on their treatment time or medical expenses resulting from personal financial incentives can be ruled out. Furthermore, from our discussion with numerous ED physicians and administrators, we believe that the ED decision makers' incentives are aligned with the general goal of ED operations, that is, to provide prompt medical treatment to the population needing it most urgently or, equivalently, to minimize the total cumulative holding cost of all patients. However, we do observe from the data that a lower triage level (less complicated) patient is more likely to be selected in the last choice incident during a physician's shift.[4] An explanation of that phenomenon is that the ED decision makers may have other objectives during the shift change, such as preventing the physician from working overtime. Studying this end-of-shift effect is not the main interest of this paper. To avoid complicating the main model, from the analyzed data, we excluded the last choice incident during each physician's shift, which takes up about 8% of the available data.

Observation 2. A patient's marginal holding cost is continuous and nondecreasing in wait time before seeing a physician and can be any constant value afterward.

During a patient's visit, we refer to the time interval before and after a patient being seen by a physician for the first time as the waiting period and treatment period, respectively. We define the marginal holding cost during the waiting period as the *marginal waiting cost*. In many healthcare settings, including EDs, it has been discovered that patients' clinical conditions deteriorate faster the longer they wait for treatment (Derlet and Richards 2000, Ostendorf et al. 2004, Diercks et al. 2007). This suggests that the marginal waiting cost is nonnegative and nondecreasing in the patient's wait time. Without loss of generality, we can assume that the marginal waiting cost is continuous because a discontinuous function with finite jumps can be approximated by a continuous function. Once the patient transitions from waiting to treatment, we allow the marginal holding cost to jump to a (small) constant value as immediate measures are taken that put the patient's risk under control.

Note that in the $Gc\mu$ rule proposed by Van Mieghem (1995), c refers to the marginal waiting cost. When making routing decisions by the $Gc\mu$ rule, only patients who are waiting need to be evaluated and compared by their $Gc\mu$ values at the time of choice. Hence, the marginal holding cost during the treatment period is not taken into account when one applies the $Gc\mu$ rule. However, regarding the property of the marginal holding cost, it may not be nondecreasing throughout a patient's entire stay when one considers the treatment period in addition to the waiting period. This merits a discussion that is to validate the asymptotic optimality of the $Gc\mu$ rule in our study setting, which we provide in Online Appendix A.

Observation 3. All servers, the ED physicians, are homogeneous, and there is no skill-based patient routing.

From our observation of the study EDs, physicians can be considered as the bottleneck resource at most times and, hence, regarded as the "servers" in the ED queueing system. In our consultation with several ED physicians from both study and nonstudy EDs, the consensus was that ED physicians are generalists and are supposed to have the capability to treat patients of all types. Matching a physician with an ED patient based on the clinical diagnosis is not the norm. In fact, this is the expectation for physicians in all EDs and not only in Canada (Zink 2006). To further validate the homogeneity of servers, we run conditional independence tests (Agresti 1996) for each ED between the treating physician ID and the treated patient's triage level or CCS conditional on the day-of-week and hour-of-day combination, which controls for patient arrival and physician shift patterns. We find that the association between physician ID and clinical diagnosis

is statistically insignificant at the 5% level for all four EDs, indicating that there is no skill-based routing among the ED physicians (Online Appendix B).

Observation 4. The EDs are critically loaded during the study period.

According to Armony et al. (2015), EDs can be viewed as critically loaded systems between late morning and late evening. From our data, we find that the peak load hours in the four study EDs can be best approximated by the period from 10 am to 2 am the next day. Thus, we keep and analyze choice incidents in 10 am–2 am only, when the EDs can be regarded as critically loaded.

With these four observations, the ED decision makers can be considered to be minimizing the total cumulative holding cost in a multiclass queueing system with multiple homogeneous servers, in which the marginal holding cost exhibits certain properties. Furthermore, by our discussion in Online Appendix A, the $Gc\mu$ rule is asymptotically optimal in such a system. This provides a strong justification and basis for us to model the patient routing decision process using a $Gc\mu$-type choice behavior with which the decision maker assesses patients in the choice set and evaluates the $Gc\mu$ value for each patient. The decision maker then chooses the patient with the highest value to be treated by the next available physician.

Specifically, a decision maker i's own valuation of choosing patient j with characteristics $\boldsymbol{X}_j$ and wait time $wait_j(t)$ at choice incident t has the following expression:

$$V_{ijt}(wait_j(t), \boldsymbol{X}_j, \boldsymbol{Y}_i) = c_j^i(wait_j(t), \boldsymbol{X}_j, \boldsymbol{Y}_i)\mu_j^i. \tag{1}$$

In the above equation, $c_j^i(wait_j(t), \boldsymbol{X}_j, \boldsymbol{Y}_i)$ represents the marginal holding cost conditional on the patient still waiting, hence, the *marginal waiting cost*; $1/\mu_j^i$ represents the service time of patient j expected by the decision maker i before treatment commences, and $\boldsymbol{Y}_i$ represents decision maker i's own attributes.

In most EDs, patient-choice decisions are made by the chief nurse or ED administrator with the occasional input from the physician, and although there are no publicly documented guidelines on how to manage the routing duties, EDs are expected to maintain consistency in their operations. Hence, although the decision makers are shuffled across different shifts, their behavior can be considered to be consistent. For this reason, we assume that a single decision maker chooses patients in each ED and estimate the choice behavior for each ED separately. Hence, we need not consider the decision maker's attributes in the conditional logit model. We perform robustness analysis for potential decision maker heterogeneity in Online Appendix C.

By assuming homogeneity across the decision makers, we can drop the subscript i in Equation (1) and get

$$V_{jt}(wait_j(t), \boldsymbol{X}_j) = c_j(wait_j(t), \boldsymbol{X}_j)\mu_j \tag{2}$$

as the valuation function. In Section 4.2, we discuss the various functional forms of the marginal waiting cost term, $c_j(wait_j(t), \boldsymbol{X}_j)$, in detail.

To account for the randomness in the routing decisions that are not captured in the available data, we combine the $Gc\mu$ rule with a discrete choice structure consistent with the additive random utility theory. Discrete choice models statistically relate the decision maker's choice to the attributes of both the decision maker and the available choice candidates. In ED patient routing, variation in medical conditions across patients is the key driver of the decision maker's choice behavior rather than the individual decision maker's attribute, which renders McFadden's conditional logit model (McFadden 1973) as the most suitable for analysis.

In the conditional logit framework, at each choice incident, the decision maker assesses patients in the choice set and evaluates the utility gained by initiating the treatment of each patient at that moment. The decision maker then chooses the patient with the highest utility. The utility of patient j at choice incident t, U_{jt}, has two components where V_{jt} has the form of Equation (2):

$$U_{jt} = V_{jt}(wait_j(t), \boldsymbol{X}_j) + \epsilon_{jt}. \tag{3}$$

We assume that the idiosyncratic random shock, ϵ_{jt}, which represents external factors that affect the patient's utility perceived by the decision maker, is i.i.d. type-I extreme value distributed.

Given this assumption, the probability of patient j being chosen in choice incident t is given by the logit form of

$$P(j \mid \boldsymbol{\Sigma}(t)) = \frac{\exp(V_{jt}(wait_j(t), \boldsymbol{X}_j))}{\sum_{p \in ChoiceSet(t)} \exp(V_{pt}(wait_p(t), \boldsymbol{X}_p))}, \tag{4}$$

where

$$\boldsymbol{\Sigma}(t) := \{(wait_p(t), \boldsymbol{X}_p) \mid p \in ChoiceSet(t)\} \tag{5}$$

denotes the patient information for choice incident t—wait time and the fixed patient characteristics for all patients waiting to be chosen in that incident.

The log-likelihood of observing the sequence of choices can be expressed as

$$\ln L = \sum_t \ln P(c(t) \mid \boldsymbol{\Sigma}(t)), \tag{6}$$

where $c(t)$ represents the index of the patient chosen at choice incident t.

For each ED, we separately estimate the decision maker's patient valuation term, $V_{jt}(wait_j(t), X_j)$, by maximizing the likelihood of the sequence of observed choices. We account for heteroscedasticity in the random shock term, ϵ_{jt}, using the Huber/White/sandwich variance estimator clustered by the choice incident, t. This allows the term to capture external shocks at choice incident t that are common to all patients waiting to be seen by a physician.

The $Gc\mu$-type choice behavior is not only grounded in classical queueing theory, but also has a meaningful clinical interpretation. The primary objectives of triage systems are to provide detailed instructions for prioritizing patients based on the observed medical conditions and to ensure that patients are treated based on urgency, acuity, and resource needs (CAEP 2014). The marginal waiting cost term, $c_j(wait_j(t), X_j)$, incorporates both urgency and acuity of the patient, the former captured by wait time, $wait_j(t)$, and the latter by the fixed characteristics in X_j. The service rate term μ_j captures the complexity of treatment (a close proxy for resource needs) for certain types of patients. For instance, a heart failure patient is likely to have a lower service rate (longer treatment time) than a patient with a non–life-threatening cut.

Saghafian et al. (2014) showed that a triage system that also incorporates patient complexity information in routing decisions can improve ED patient flow compared with a triage system that uses patient urgency information only. For each of the four study EDs, we test whether the decision maker incorporates patient complexity information into the decision maker's routing decisions, and if so, at what level. To do that, we calibrate three possible models and compare their goodness of fit. First, we fit an *Urgency(only)-based* model in which the decision maker does not use complexity information at all. Hence, we can assume μ_j to be a constant $\mu_{overall}$ for the entire patient population. Such a model can be considered as a Gc-type choice behavior because service rate μ_j is not utilized in the decision-making process. Second, we fit a *Complexity (triage)-based* model in which complexity of treating patient j is assessed at a coarse patient triage level, $Tri(j)(=2,3)$, that is, $\mu_j = \mu_{Tri(j)}$. Finally, we fit a *Complexity(CCD)-based* model in which complexity of patient j is assessed at a more granular clinical condition level, CCD, with 113 distinct codes. Hence, $\mu_j = \mu_{CCD(j)}$ for a patient with CCD code $CCD(j)$. By comparing the model fit of these models, we can examine which model best represents how the complexity information has been used in practice.

4.2. Marginal Waiting Cost Function

Our main interest is to understand how the ED decision maker incorporates each patient's wait time information into the patient prioritization decision. We achieve this by inferring the patient waiting cost structure within the conditional logit-$Gc\mu$ framework. We decompose the marginal waiting cost term $c_j(wait_j(t), X_j)$ in Equation (2) into two parts: $f_w^{Tri(j)}(wait_j(t))$ and $f_c(X_j)$. The first component, $f_w^{Tri(j)}$, is a function of the patient's (cumulative) wait time, $wait_j(t)$, and triage level, $Tri(j)$. The second component, f_c, is a linear function of the patient's fixed characteristics. We, thus, derive the following expression of the decision maker's patient valuation:

$$V_{jt}(wait_j(t), X_j) = (f_w^{Tri(j)}(wait_j(t)) + f_c(X_j))\mu_j. \quad (7)$$

The decomposition of $c_j(wait_j(t), X_j)$ allows us to explore the functional forms of $f_w^{Tri(j)}(wait_j(t))$ and infer how the decision maker's perceived patient waiting cost depends on the two variables mostly to our interest: triage level and wait time. We model $f_w^{Tri(j)}(wait_j(t))$ with various functional forms and compare their fits with the observed data to identify the form that best characterizes the marginal ED patient waiting cost function perceived by the ED decision maker. We assume that both triage levels 2 and 3 have the same functional form of $f_w^{Tri(j)}(wait_j(t))$ but with parameters that may differ by triage level. We consider the following five functional forms for $f_w^{Tri(j)}(wait_j(t))$.

Constant marginal waiting cost function, $f_w^{Tri(j)}(wait_j(t)) = \beta_1^{Tri(j)}$, corresponds to a linear cumulative waiting cost most commonly assumed in the literature (Mendelson and Whang 1990, Aksin et al. 2013, Yu et al. 2016). However, in the ED setting, one may conjecture that the (cumulative) wait time has a nonlinear (usually increasing in margin) effect on patient conditions.

Hence, we also fit a *linear* marginal waiting cost function, $f_w^{Tri(j)}(wait_j(t)) = \beta_1^{Tri(j)} \cdot wait_j(t)$, and higher degree polynomials such as *quadratic*, $f_w^{Tri(j)}(wait_j(t)) = \beta_1^{Tri(j)} \cdot wait_j(t) + \beta_2^{Tri(j)} \cdot wait_j(t)^2$, and *cubic*, $f_w^{Tri(j)}(wait_j(t)) = \beta_1^{Tri(j)} \cdot wait_j(t) + \beta_2^{Tri(j)} \cdot wait_j(t)^2 + \beta_3^{Tri(j)} \cdot wait_j(t)^3$, which have all been studied in the literature as well (Dewan and Mendelson 1990, Parlar and Sharafali 2014).

We also consider a *piece-wise linear* function whose slope may have an abrupt change at certain points. This accounts for the possible impact of the CTAS target wait times on a patient's marginal waiting cost perceived by the ED decision maker. For computational tractability, we propose a piece-wise linear function with a single break point:

$$f_w^{Tri(j)}(wait_j(t)) = \beta_1^{Tri(j)} \cdot wait_j(t) + \beta_2^{Tri(j)} \cdot (wait_j(t) - \gamma_1^{Tri(j)})^+. \quad (8)$$

Table 3. Model Fit by Marginal Waiting Cost Function and Patient Complexity Information Used in Patient-Routing Decisions

			Complexity information					
			No complexity (urgency only)		Complexity (triage)		Complexity (CCD)	
ED	Marginal waiting cost function	df	Log-likelihood	BIC	Log-likelihood	BIC	Log-likelihood	BIC
A	Constant	125	−110,282	222,202	−110,339	222,314	−110,438	222,513
	Linear	127	−107,549	216,761	−107,610	216,882	−107,756	217,175
	Quadratic	129	−106,681	215,052	−106,747	215,182	−107,008	215,704
	Cubic	131	−106,358	214,432	−106,425	214,565	−106,674	215,063
	Piece-wise linear	131	−105,692	213,099	−105,760	213,235	−105,998	213,712
B	Constant	125	−62,943	127,423	−62,943	127,423	−62,978	127,493
	Linear	127	−58,505	118,571	−58,506	118,574	−59,109	119,779
	Quadratic	129	−53,791	109,168	−53,797	109,179	−56,186	113,957
	Cubic	131	−53,045	107,701	−53,053	107,716	−55,696	113,003
	Piece-wise linear	131	−51,884	105,378	−51,892	105,394	−53,237	108,085
C	Constant	125	−63,287	128,123	−63,294	128,138	−63,369	128,287
	Linear	127	−58,985	119,545	−58,995	119,564	−59,502	120,579
	Quadratic	129	−56,676	114,951	−56,697	114,994	−57,933	117,465
	Cubic	131	−56,238	114,099	−56,261	114,146	−57,617	116,857
	Piece-wise linear	131	−54,799	111,223	−54,821	111,267	−55,812	113,247
D	Constant	125	−47,536	96,579	−47,541	96,588	−47,632	96,771
	Linear	127	−44,735	91,001	−44,733	90,997	−45,406	92,343
	Quadratic	129	−43,989	89,533	−43,992	89,539	−45,017	91,589
	Cubic	131	−43,843	89,264	−43,846	89,271	−44,870	91,319
	Piece-wise linear	131	−43,254	88,087	−43,255	88,088	−44,171	89,921

Note. df, degrees of freedom; BIC, Bayesian information criterion.

According to Equation (8), $\beta_1^{Tri(j)}$ and $\beta_1^{Tri(j)} + \beta_2^{Tri(j)}$, respectively, represent the slope of the marginal waiting cost below and above the break point $\gamma_1^{Tri(j)}$. All three parameters depend on the triage level of the focus patient j, $Tri(j)$; hence, the piece-wise linear model estimates a total of six parameters.

However, the existence of a break point is not guaranteed in the data generation process, in which case a standard linear function may fit the data better. The estimation method of the piece-wise linear specification is general enough to capture the nonexistence of break points. We refer the readers to Muggeo (2003) for details on the break point estimation procedure in a piece-wise linear specification. In the model that we use to derive our main findings and the policy implications thereof, we limit the maximum number of break points per triage level to one. In Online Appendix F, we relax this assumption and consider multiple break points per triage level.

5. Results

We use the maximum-likelihood method (Equation (6)) to estimate the parameters in Equation (7). For each of the four EDs, we individually explore which combination of the two dimensions discussed in the previous section best represents the ED decision maker's patient-routing decisions. We compare model fits across (a) different functional forms of $f_w^{Tri(j)}(wait_j(t))\ \forall Tri(j) \in \{2,3\}$ in the marginal waiting cost term $c_j(wait_j(t), \boldsymbol{X}_j)$ and (b) different patient complexity information levels, μ_j. We then visualize the model that best fits the observed data—plotting the decision maker's valuation of treating a patient (Equation (2)) as a function of wait time for each triage level $l = 2, 3$—and derive managerial insights and policy implications.

5.1. Model Fit: Marginal Waiting Cost Function and Patient Complexity Information

We use the Bayesian information criterion (BIC) to compare the model fits. When fitting models, adding parameters can improve the likelihood, but it may result in overfitting. BIC measures this trade-off by rewarding models with the best statistical fit (likelihood) and penalizing for model complexity (degree of freedom) proportional to the size of the data (natural log of number of total observations). Statistically, the model with the lowest BIC score is preferred (Burnham and Anderson 2002). Table 3 reports the model fit across the two dimensions of interest: functional form of $f_w^{Tri(j)}(wait_j(t))$ in the marginal waiting cost and the patient complexity information used.

First, we explore the form of the marginal waiting cost function. In all four EDs, within each complexity information model, we find that the ED decision makers' perceived ED patient waiting costs are best approximated by a piece-wise linear marginal waiting cost function. Compared with the standard linear function, the piece-wise linear function fits the data significantly better in all situations by a large margin. This is despite the penalty for having four more parameters to estimate as captured in the BIC

Table 4. Estimation Results: Piece-Wise Linear Marginal Waiting Cost Function in Urgency-Only Model

	Coefficients				Odds ratio			
	ED A	ED B	ED C	ED D	ED A	ED B	ED C	ED D
Break point (mins)								
Triage 2	8.5***	13.6***	18.4***	12.7***				
	(0.302)	(0.163)	(0.239)	(0.372)				
Triage 3	19.5***	18.4***	18.5***	19.3***				
	(0.320)	(0.131)	(0.195)	(0.369)				
Slopes								
Triage 2	0.145***	0.297***	0.208***	0.161***	1.156***	1.345***	1.231***	1.175***
Below break point	(0.008)	(0.007)	(0.005)	(0.008)	(0.009)	(0.009)	(0.006)	(0.009)
Triage 2	0.002***	0.001	0.001	0.001***	1.002***	1.001	1.001	1.001***
Above break point	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)
Triage 3	0.083***	0.201***	0.190***	0.107***	1.087***	1.223***	1.209***	1.113***
Below break point	(0.002)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.005)	(0.004)
Triage 3	0.005***	0.009***	0.009***	0.009***	1.005***	1.005***	1.009***	1.009***
Above break point	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Intercept								
Triage 2	1.119***	0.704***	1.421***	1.217***	3.061***	2.022***	4.143***	3.376***
	(0.056)	(0.080)	(0.096)	(0.083)	(0.171)	(0.162)	(0.398)	(0.281)
Average CCD effect								
Triage 2	0.223	0.229	0.230	0.219				
	((0.444))	((0.336))	((0.352))	((0.288))				
Triage 3	0.061	0.009	0.134	0.077				
	((0.220))	((0.226))	((0.222))	((0.160))				
N (observations)	485,895	217,922	241,689	171,777				
N (choice incidents)	56,604	43,669	38,331	31,427				
McFadden's R^2	0.078	0.198	0.165	0.122				

Notes. Clustered standard errors in single parentheses. Standard deviation of average CCD effects in double parentheses.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

calculation. The piece-wise linear function also outperforms the cubic function, which has the same degree of freedom. In the family of polynomial functions, increasing the degree of the polynomial significantly improves the model fit, which suggests a strong nonlinearity of the marginal cost function.

Next, fixing $f_w^{Tri(j)}(wait_j(t))$ as the piece-wise linear function, we find that the *Urgency(only)-based* model consistently outperforms the two *Complexity-based* models in all four EDs. Especially, the gap in BIC scores widens as the complexity information becomes more granular in the order of *Urgency(only)-based*, *Complexity(triage)-based*, and *Complexity(CCD)-based*. These results suggest that the ED decision makers likely *do not* incorporate the complexity information into their routing decisions. This finding is consistent with the literature (see Saghafian et al. 2014) and the general belief held by many ED physicians in the metro Vancouver area with whom we conducted interviews.

5.2. Estimation Results: Piece-Wise Linear Marginal Waiting Cost Function

By comparing the fit of the different models, we find that the piece-wise linear marginal waiting cost function and *Urgency(only)-based* model best represents the ED decision makers' patient-routing decisions in all four study EDs. Given this model choice, we interpret the coefficient estimates and infer the decision makers' prioritization behaviors within each triage level and across different triage levels. Table 4 reports estimation results from the maximum likelihood estimation of Equation (6) with the piece-wise linear marginal waiting cost function (Equation (8)) in the patient valuation term (Equation (7)). Columns (1) to (4) list the coefficient estimates and relative statistics of the four EDs, and columns (5) to (8) list the corresponding odds ratios for applicable independent variables.

The first section of rows reports, in minutes, the location of the estimated break point, γ_1 in Equation (8), for each triage level. The piece-wise linear estimation procedure concluded that for both triage levels in all four EDs, a piece-wise linear function is a better fit than a standard linear function. For all four EDs, break points monotonically increase in the order of the triage levels, which is consistent with the order of CTAS target wait times. The orders are strict in all EDs apart from ED C, in which the break points are statistically not different. The exact locations of the break points remain fairly close to the suggested CTAS target wait times—15 minutes for triage level 2 and 30 minutes for triage level 3—suggesting that the CTAS fractile response objective may be a key driver of the break-point phenomena.

The second section of rows reports the estimation of β_1 and $\beta_1 + \beta_2$ in Equation (8), which represent the slopes of the marginal waiting cost function below and above, respectively, the estimated break point, γ_1. These parameters have important implications in regard to the decision makers' routing behaviors *within* the same triage level. For both triage level 2 and 3 patients in all four EDs, marginal waiting costs have a significant positive slope below the break point. This suggests that the routing behavior is close to FCFS for patients within the same triage level and with wait times below the break point. For instance, according to the odds ratio in column (6), for triage level 2 patients in ED B with a wait time less than 13.6 minutes, waiting an extra minute increases the odds of being chosen by a factor of 1.345. However, once patients wait beyond the break point of their triage level, the decision makers' adherence to the FCFS principle significantly decreases. This is suggested by the fact that $\beta_1 + \beta_2$ is substantially smaller than β_1 in all four EDs. For example, for triage level 2 patients in ED B with wait times longer than 13.6 minutes, the slope of the marginal waiting cost, $\beta_1 + \beta_2$, is close to zero, which indicates that the marginal waiting cost is nearly a constant above the break point, and waiting an extra minute increases the odds of being chosen by a factor of only 1.001. According to the $Gc\mu$ rule, these patients will receive almost no extra priority by waiting longer. This result confirms a plausible conjecture about the impact of the CTAS fractile response objective on the patient-routing behavior: the decision maker has less incentive to choose a patient who waited the longest from among those having already waited longer than the target wait time. This implies that the CTAS objective may disincentivize the decision makers to follow the expected practice of first treating patients who have had longer wait times among the patients who have already missed the target wait time.

In Section 5.3, we explain how the magnitude of the slopes in the piece-wise linear marginal waiting cost function reflect the degree of adherence to the FCFS principle using an example of two individual patients.

The third section of rows reports the estimated triage level 2 intercept (dummy variable), which captures the decision makers' prioritization behavior *across* different triage levels. Triage level 3 is excluded as the base category. The results suggested by the positively significant coefficients in all four EDs are consistent with clinical expectations: triage level 2 receives priority over triage level 3. However, this comparison across triage level intercepts is conditional on all patients waiting zero minutes. The effect of increase in wait times may differ by triage level. We demonstrate this effect in Section 5.3.

Finally, in the fourth section of rows, because of the large number of distinct CCD codes, we only report the average and standard deviation of the coefficient values (intercept) of the CCD control (dummy) variables tabulated by the triage level of the patient. The CCD intercepts are identified by excluding the most common code "abdominal pain, moderate pain, episodic vomiting, fever," which has a default triage level of 3, as a base category. In all four EDs, triage level 2 patients have a larger average intercept value contributed by their respective CCD dummies than triage level 3 patients. The distinct average values and non-negligible standard deviations support the validity of CCD intercepts as an effective control of patients' heterogeneous medical conditions.

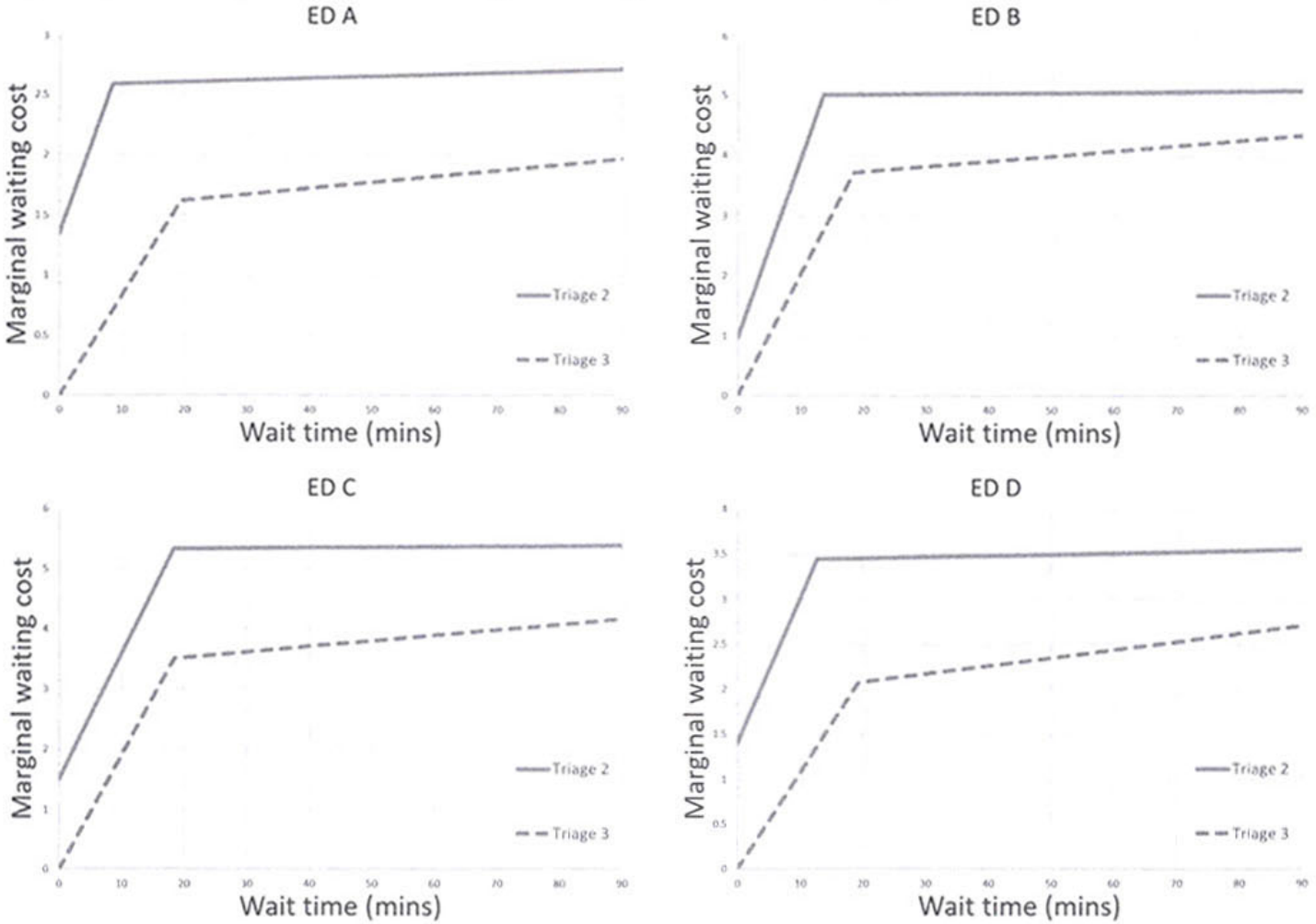
As a reference for how well the conditional logit-$Gc\mu$ framework captures the decision makers' patient-routing behaviors, we use McFadden's pseudo R^2 as a measure of goodness of fit. McFadden suggested that pseudo R^2 values between 0.2 and 0.4 should be considered indicative of *extremely good* model fits (Louviere et al. 2000). Simulations by Domencich and McFadden (1975) equivalenced this range to R^2 values of 0.7 to 0.9 for linear regression. Analysts should not expect to obtain pseudo R^2 values as high as the R^2 values commonly obtained in many ordinary least square models. Apart from ED A, the pseudo R^2s range from 0.122 to 0.198, which can be considered reasonably good fits. We discuss the validity of the conditional logit-$Gc\mu$ framework in representing the decision makers' patient-routing behaviors in more detail in Section 6.

5.3. Delay-Dependent Patient Prioritization

In Figure 2, for each of the four EDs, we use the estimation results from Table 4 to calculate and plot $\mathbb{E}[V_{jt}(wait, \boldsymbol{X}_j) \mid Tri(j) = l]$ as a function of *wait* for $l = 2, 3$. The functional value of $\mathbb{E}[V_{jt}(wait, \boldsymbol{X}_j) \mid Tri(j) = l]$ stands for the decision makers' average valuation of choosing a patient from triage level l with wait time *wait*. According to the *Urgency(only)-based* model,[5] we have $\mathbb{E}[V_{jt}(wait, \boldsymbol{X}_j) \mid Tri(j) = l] = f^l_w(wait) + \mathbb{E}[f_c(\boldsymbol{X}_j) \mid Tri(j) = l]$ by Equations (2) and (7), where $f^l_w(wait)$ is the estimated piece-wise linear function (Equation (8)) and $\mathbb{E}[f_c(\boldsymbol{X}_j) \mid Tri(j) = l]$ is the average contribution of fixed characteristics of patients in triage class l, including the average CCD effect. Note that we have pinned the intercept of the triage level 3 curve to zero as subtracting a constant from the intercept of both curves will not change the choice behavior.

The plotted curves illustrate the main attributes of the ED decision makers' perceived ED patient marginal waiting cost. First, they show clear piece-wise linearity in wait time, especially the flattening of the marginal cost beyond the break points. Second, the plotted functional values quantify the aggregate impact of wait time and triage level on a patient's priority in comparison with the random shock ϵ_{jt}, whose standard deviation has been fixed as one unit. For example, in

Figure 2. (Color online) Piece-Wise Linear Marginal Waiting Cost Function by Triage Level



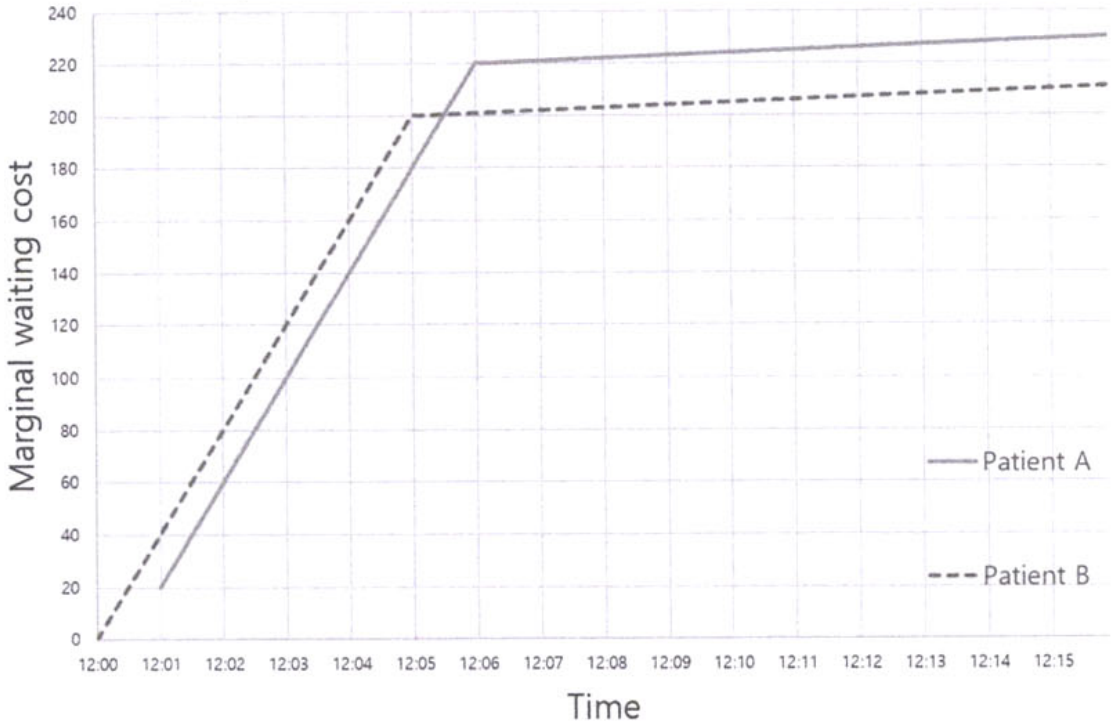
ED C, on average, triage level 2 patients have about 1.5 unit priority over triage level 3 patients when both have just entered the ED, that is, both *wait* = 0 (Figure 2). If a choice incident occurs immediately upon a simultaneous arrival of both an *average* triage level 2 and 3 patient, the odds for the triage level 2 patient being chosen is $exp(1.5) = 4.48$ times that of the triage level 3 patient.

Figure 2 visualizes the delay-dependent aspect of patient prioritization behavior in CTAS EDs. *In general*, higher triage level patients receive priority over lower triage level patients as one would expect based on general medical guidelines. This is supported by the marginal waiting cost curve of triage level 2 being stacked above triage level 3 in all four EDs. However, we observe possible instances in which the triage level-wise prioritization order is reversed; depending on their respective wait times, lower triage level patients can be prioritized over higher triage level patients. In all four EDs, the marginal waiting cost of triage level 2 patients can be smaller than that of triage level 3 patients who have waited for a much longer period of time. For instance, in ED B, a triage level 3 patient who has waited 15 minutes has a marginal waiting cost valued around three whereas a triage level 2 patient who has waited less than eight minutes has a marginal waiting cost smaller than three. This observation suggests that patients are routed not only by triage level (static) priorities but also by their actual (dynamic) wait time, suggesting that all four EDs are using delay-dependent prioritization.

A noticeable observation is the gap between the triage level curves varying with wait time, suggesting that wait time may have a nonhomogeneous effect on patient priorities. The priority between triage levels changes with time rather than being invariant. This is particularly evident when the wait time is less than 20 minutes, at which time the curve begins to "plateau." In all four EDs, more than 75% of the observations within each triage level are in the range above the respective break points. Hence, the flattening of the marginal cost curve past the break point is not driven by the lack of data points in the region.

One should note that the plotted marginal waiting cost values in Figure 2 do not indicate an individual patient's relative priority as the values only reflect the *average* in each triage level without considering the

Figure 3. (Color online) Delay-Dependent Patient Prioritization: An Example of Two Patients in Same Triage Level



individual patient's characteristics. For example, in ED A, even though triage level 2 patients who waited more than five minutes dominate triage level 3 patients who waited less than 90 minutes, certain triage level 2 patients who waited longer than five minutes can have lower marginal waiting costs than triage level 3 patients who waited less than 90 minutes.

To further illustrate the interpretation of the curves and the estimated slope values of the piece-wise linear marginal waiting cost function in relation to adherence to the FCFS principle, we provide an example of comparing the priority between two individual patients in the same triage level in Figure 3.

Suppose two patients A and B have different characteristics but are assigned to the same triage level. The triage level they belong to has an estimated marginal waiting cost slope of 40 below the estimated break point of five minutes and a slope of one above the break point. The wait time–independent fixed attributes term of Patient A is larger than that of Patient B by 20, that is, $f_c(\boldsymbol{X}_A) - f_c(\boldsymbol{X}_B) = 20$, but Patient B has arrived one minute earlier than Patient A. The shapes of the wait time–dependent component of the marginal waiting costs, $f_w^{Tri(j)}(wait_j(t))$, are identical for the two patients; however, the intercepts (or starting values of the curves) are different as the wait time–independent characteristics component $f_c(\boldsymbol{X}_j)$ varies for the two. The marginal waiting cost curve in Figure 3 represents the sum of the two components as a function of the respective patient's actual wait time along the horizontal axis in real time starting at 12:00. Up until 12:05, patient B has a higher total marginal waiting cost and, thus, a larger odds ratio to be selected compared with patient A; hence, the FCFS principle is more likely to be adhered to. This is a result of the fact that patient B receives additional priority (40) by arriving one minute earlier than patient A, and the difference in the characteristics term is not enough to overcome. However, after 12:05, the late-arrived patient A has a higher total marginal waiting cost and, thus, a larger odds ratio; thus, the FCFS principle is likely to be violated. Because the slope after the break point, one, is smaller than the characteristic term gap, 20, wait time matters less once both patients wait past the break point, and the priority is dominated by the characteristic term. In this manner, the magnitude of the slope in the piece-wise linear marginal waiting cost function reflects the degree of adherence to the FCFS principle.

6. Model Validation

This section consists of three parts. The first part lays out the ground for justifying the structural assumptions that have been imposed by the conditional logit-$Gc\mu$ framework. In the second part, we prove consistency of the maximum log-likelihood estimator (MLE) under the conditional logit-$Gc\mu$ framework. This justifies the main insights of the paper, which are derived based on the MLE results. The last part tests the goodness of fit of the selected model (i.e., *Urgency (only)-based* routing and piece-wise linear marginal waiting cost) for out-of-sample data.

6.1. Justification of Framework Assumptions

Our conditional logit-$Gc\mu$ framework falls into the category of structural estimation methods, which are developed to approximate complicated decision-making processes and derive estimations for certain decision parameters (e.g., Cohen et al. 2003 and Olivares et al. 2008). As with other structural estimation methods, our conditional logit-$Gc\mu$ framework has imposed certain underlying structural assumptions. In this subsection, we provide the rationale for imposing these structural assumptions, which justifies the conditional logit-$Gc\mu$ framework and the results we derived therein.

The conditional logit-$Gc\mu$ framework has restricted patient routing decisions to *myopic choices*. Formally, a myopic choice means that a decision maker i always chooses a patient $j^* \in \arg\max\{U_{ijt} \mid j \in ChoiceSet(t)\}$. The value function U_{ijt} can be calculated based on the attributes and wait time of patient j and attributes of the decision maker i. We consulted with administrators and physicians from the four study EDs and received consensus response that given the high uncertainty in ED operations, it is unclear how to make forward-looking choices, and the decisions are mostly myopic in practice. When a choice decision is to be made, the ED decision maker reads information of each patient from the PCIS screen, which shows age, CCS, CCD, triage level, arrival method, and wait time (duration since time of triage). The PCIS, however, does not provide any predictive analytic or sophisticated guidance that can facilitate forward-looking decisions. The only exception is that the decision maker becomes more likely to choose easier cases near the physician's shift change by anticipating that, otherwise, the physician's shift may get prolonged. Because the last choice incidents in the physicians' shift have already been removed from the data, myopic choice appears to be a reasonable assumption for the rest.

We perform robustness tests for other structural assumptions that have been imposed, and the details are discussed in the appendices, including decision maker heterogeneity (Online Appendix C), unobserved patient heterogeneity (Online Appendix D), the independence from irrelevant alternatives (IIA) property of the conditional logit model (Online Appendix E), and number of break points in the piece-wise linear specification of $f_w^{Tri}(\cdot)$ (Online Appendix F). We also perform an out-of-sample test to further justify the validity of our structural estimation framework representing the patient-routing decision process (see Online Appendix H).

6.2. Consistency of the Maximum Log-Likelihood Estimator

Our consistency result is developed for the conditional logit-$Gc\mu$ framework with $f_w^{Tri}(\cdot)$ in a general function class—polynomial regression splines, which cover all five functional forms that we have studied. Although our proof uses standard methods, it cannot be directly implied from the existing results on the consistency of MLE for generalized linear models (e.g., Fahrmeir and Kaufmann 1985 and Newey and McFadden 1994) because the $f_w^{Tri}(\cdot)$ term in our model can be nonlinear and nonsmooth, and the observed choice sequences are not i.i.d. A rigorous statement of the consistency results and the proof containing the technical details are shown in Online Appendix G. We also show that the distribution of MLE for our model is generally not asymptotically normal because of the boundary constraint.

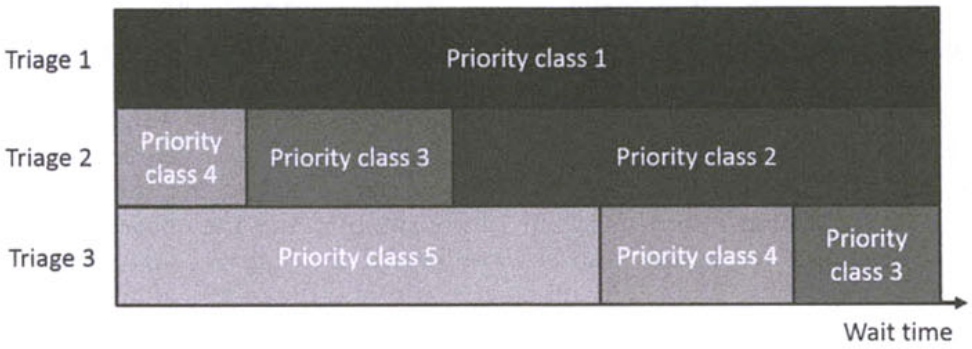
7. Policy Implications

We highlight several important policy implications derived from our estimation results.

First, the CTAS fractile response objective may provide incentives that lead to unintended consequences. The ED decision makers generally follow the FCFS principle within the same triage level, but their adherence to the FCFS principle decreases among patients who have waited past a certain break point, that is, 13.3 minutes for triage level 2 patients and 18.9 minutes for triage level 3 patients, on average. This might be because of the reduced incentive of the decision maker to choose the patient who has waited the longest from among those who have already waited more than the CTAS target wait time. As a result, patients who have waited past that target wait time are likely to wait even longer because they are not given extra priority for having waited longer. This result has implications for improving CTAS. Both from urgency and fairness standpoints, for patients in the same triage level, treating those who have waited longer is the reasonable expectation. Although the target time was developed as an "ideal," the existence of an explicit fractile response objective may have adversely affected patients by making those who have waited a significant amount of time wait even longer. Monitoring patient wait time from other angles, such as looking at the longest wait over a certain period of time, may reduce such outcomes. A simpler but limited alternative would be to implement multiple target wait times within a triage level. This might prevent prolonged waits within the range of the largest target wait time but still be susceptible outside of it. It is possible that the target fractile is, realistically, not achievable in every ED across Canada. Hence, implementing target wait times adjusted to the risk factors and congestion level of each ED may be another way of improving the current CTAS structure.

Second, we find that ED decision makers use a delay-dependent prioritization policy in which the relative priority across different triage levels may depend on

Figure 4. (Color online) Delay-dependent Patient Prioritization in Triage Systems



patient wait times. This suggests that the ED decision makers are making sophisticated routing decisions in the sense of not just following a strict absolute prioritization across triage levels. The added complexity may have a positive impact on patient outcomes: lower triage level patients would not get pushed back too far. This allows even low triage level patients to be treated within a reasonable time frame, which was one of the key motivations of the CTAS fractile response objective and could possibly reduce ED patient abandonments and revisits at a later time when patients are in potentially worse conditions.

The idea of delay-dependent priority was initiated by Jackson (1960) and Kleinrock (1964) to allow decision makers extra freedom in routing decisions so that they could manipulate the relative wait times in each priority level. This is an important aspect of the ED setting as patient risk is highly dependent on the time delay and varies by triage level. However, there is a gap between the current design of the CTAS system and how patient prioritization is executed in practice. Our results show that practitioners are using a delay-dependent priority rule in practice, yet the CTAS design lags in this regard as it does not provide guidelines for prioritizing across triage levels. They only acknowledge the relative risk of patient delay in the form of the target wait times differing by triage level. It is unclear whether the delay-dependent priority rule currently used in practice is clinically appropriate or optimal. To this end, further examination of this rule is needed. If implementing such a delay-dependent priority rule is acceptable to the medical community, then the policy maker should consider providing corresponding guidelines. We give an example of such a guideline in Figure 4, in which patients with equal priority are grouped into the same priority class and represented by the same color. A patient's priority class depends on both the patient's triage level and actual wait time. For example, in Figure 4, triage level 2 patients with short wait times and triage level 3 patients with intermediate wait times both belong to priority class 4. Because this guideline provides a relative priority rule, it can be adjusted to each ED's unique situation and varying congestion level by adjusting the priority grouping cutoff wait times.

Third, we find only minimal evidence of patient complexity information at the granular clinical condition level being incorporated into current prioritization decisions. This finding is confirmed during our discussions with Vancouver ED physicians. The physicians' responses are that there are two possible reasons why ED personnel rarely think of patient complexity in the patient choice decisions: the decision maker is incapable of properly assessing patient complexity at the moment, and the CTAS lacks a structured guideline on how to assess complexity, which could overcome such incapability. The benefit of incorporating complexity information into patient triage, a practice called complexity-based triage, has been discussed by Saghafian et al. (2014). Because of the proven optimality of the $Gc\mu$ rule, from an operational perspective when it comes to routing, if one incorporates patient complexity information into routing decisions in the CTAS setting, it will likely lead to improved patient outcomes.

Finally, the implementation of a delay-dependent prioritization policy and the incorporation of complexity information both call for decision makers to have high levels of expertise, which suggests that it would be preferable to hire physicians for both triage (classification) and routing (prioritization) in contrast to having the ED administrator/chief nurse doing those tasks. Physician triage has been implemented in some hospitals and was found to improve certain operational performance measures such as ED length of stay, number of patients who left without being seen, and total time and number of days on ambulance diversion (Han et al. 2010, Rowe et al. 2011, Imperato et al. 2012). It may be worthwhile exploring whether implementing physician decisions in the entire ED patient-flow process—not only in triage but also in routing—would improve ED operations. Nevertheless, physicians are an expensive resource, and the efficiency of allotting the physician's time to non-patient treatment activities, for example, triage decisions, is questionable (Rowe et al. 2011). It is worth exploring whether the process of assessing patient priority and

complexity can be standardized into a protocol that can be used by nonphysician ED decision makers who do not have the requisite medical knowledge.

8. Conclusions and Future Research

In this paper, we studied the decision makers' patient-routing behaviors in Canadian ED triage systems. We modeled the patient choice behavior in a discrete choice-$Gc\mu$-type framework and found that a decision maker's perceived marginal patient waiting cost is best fit by a piece-wise linear concave function in wait time for each triage level.

The cost of ED patients waiting can be understood from three different perspectives: a clinical perspective purely driven by clinical outcomes; the patient's perspective driven by the patient's own satisfaction and utility; and the routing decision maker's perspective driven by various aspects, including objectives of the care-providing organization and clinical outcomes. The first two perspectives have been examined in the emergency medicine and OM literature. Guttmann et al. (2011) found that patients present in an ED during shifts with longer average wait times were associated with higher mortality rates and a greater chance of being admitted to a hospital within seven days of discharge from the ED. However, the clinical cost of waiting is not yet clearly understood at the individual patient level. Batt and Terwiesch (2015) empirically studied how ED congestion and queueing dynamics affect patient abandonment behavior. We studied the third perspective by identifying how patient waiting is perceived by the ED decision makers at the individual patient level.

One of our main findings is that the ED decision makers' perceived marginal patient waiting costs flatten above certain threshold points. Aligned with the views of the ED physicians to whom we presented our results, we believe these phenomena may be driven by the CTAS fractile response objective, supported by the fact that the threshold points are close to the CTAS target wait times. However, there is the possibility that other incentives in the Canadian ED system unknown to us may also be driving such results. As our framework is general enough to apply to other EDs with similar patient–visit level data, it would be interesting to compare our results to patient-routing behavior in other EDs. In particular, exploring EDs without the fractile response objective would provide a control group for comparison. Even Canadian EDs subject to the same fractile response objective but in different regions of the country are worth exploring because such a study could identify local effects that may affect patient routing in certain ways. Such results would help improve the CTAS design to accommodate varying local patient characteristics, for example, offering different target wait times by region, which the current design does not do. Furthermore, repeating the analysis on U.S. EDs, which use a different triage system (ESI) without the fractile response objective, can add insights into understanding patient routing behavior in EDs in general.

To our knowledge, we are among the first to empirically study the control side of queueing decisions in a multiclass queue regardless of application. A natural extension would be to apply our framework to other applications. In call centers, which is the primary application of studies in multiclass queue controls, the decision maker generally follows a predetermined routing rule. However, in some other applications, the decision maker may have discretion to route the customers and does not need to adhere exactly to the predetermined system routing rule. It may be interesting to explore human factors in routing decisions to understand when the decision maker adheres to the system's predetermined rule and when the decision maker does not and whether it is related to queue length, average wait times in specific priority classes, or other operational performance measures of queues.

Acknowledgments

The authors thank the two anonymous reviewers, the associate editor, and the area editor for their many helpful suggestions. They also thank Keith Head, Garth Hunte, and MacKenzie Winston for their valuable comments.

Endnotes

[1] In this paper, patient routing includes both (1) prioritization across different triage levels and (2) service disciplines within the same triage level, for example, first-come first-served or not.

[2] In the study EDs, this would often be the chief nurse or ED administrator with occasional input from the physician.

[3] Several papers have empirically studied the customer's own perception of waiting, for example, Aksin et al. (2013) and Yu et al. (2016). These papers aim at understanding a customer's waiting experience rather than how the prioritization decisions are made. Another few empirical papers have looked into the server's routing behavior, such as when to deviate from FCFS (Ibanez et al. 2017) and how the routing decision depends on workload, speed, and service quality (Tan and Staats 2017).

[4] http://blogs.ubc.ca/ycding/files/2018/03/PatientChoice-Final-supplementary.pdf.

[5] In the *Urgency(only)-based* model, we can further assume the constant service speed $\mu_{overall} \equiv 1$ as the constant service speed can be absorbed into the coefficients in the function $f_w^{Tri}(\cdot)$ and $f_c(\cdot)$.

References

Agresti A (1996) *An Introduction to Categorical Data Analysis*, Vol. 135 (Wiley, New York).

Aksin Z, Ata B, Emadi SM, Su C-L (2013) Structural estimation of callers' delay sensitivity in call centers. *Management Sci.* 59(12): 2727–2746.

Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Syst.* 5(1):146–194.

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

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

Bernstein SL, Aronsky D, Duseja R, Epstein S, Handel D, Hwang U, McCarthy M, et al. (2009) The effect of emergency department crowding on clinically oriented outcomes. *Acad. Emerg. Medicine* 16(1):1–10.

Beveridge R (1998) CAEP issues. The Canadian Triage and Acuity Scale: A new and critical element in health care reform. Canadian Association of Emergency Physicians. *J. Emerg. Medicine* 16(3):507–511.

Bullard MJ, Unger B, Spence J, Grafstein E (2008) Revisions to the Canadian Emergency Department Triage and Acuity Scale (CTAS) adult guidelines. *Can. J. Emerg. Medicine* 10(2):136–142.

Bullard MJ, Chan T, Brayman C, Warren D, Musgrave E, Unger B; Members of the CTAS National Working Group (2014) Revisions to the Canadian Emergency Department Triage and Acuity Scale (CTAS) guidelines. *Can. J. Emerg. Medicine* 16(6): 485–489.

Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer-Verlag, New York).

CAEP (2014) CTAS implementation guidelines. Retrieved February 25, 2015, http://caep.ca/resources/ctas/implementation-guidelines.

Cohen MA, Ho TH, Ren ZJ, Terwiesch C (2003) Measuring imputed cost in the semiconductor equipment supply chain. *Management Sci.* 49(12):1653–1670.

Derlet RW, Richards JR (2000) Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. *Ann. Emerg. Medicine* 35(1):63–68.

Dewan S, Mendelson H (1990) User delay costs and internal pricing for a service facility. *Management Sci.* 36(12):1502–1517.

Diercks DB, Roe MT, Chen AY, Peacock WF, Kirk JD, Pollack CV, Gibler WB, Smith SC, Ohman M, Peterson ED (2007) Prolonged emergency department stays of non–ST-segment-elevation myocardial infarction patients are associated with worse adherence to the American College of Cardiology/American Heart Association guidelines for management and increased adverse events. *Ann. Emerg. Medicine* 50(5):489–496.

Dobson G, Tezcan T, Tilson V (2013) Optimal workflow decisions for investigators in systems with interruptions. *Management Sci.* 59(5):1125–1141.

Domencich TA, McFadden D (1975) *Urban travel demand: A behavioral analysis* (North-Holland, Amsterdam).

Drummond AJ (2002) No room at the inn: Overcrowding in Ontario's emergency departments. *Can. J. Emerg. Medicine* 4(2):91–97.

Fahrmeir L, Kaufmann H (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* 13(1):342–368.

Graff L (1999) Overcrowding in the ED: An international symptom of health care system failure. *Amer. J. Emerg. Medicine* 17(2): 208–209.

Gurvich I, Whitt W (2009) Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* 11(2):237–253.

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. *Brit. Med. J.* 342:d2983.

Han JH, France DJ, Levin SR, Jones ID, Storrow AB, Aronsky D (2010) The effect of physician triage on emergency department length of stay. *J. Emerg. Medicine* 39(2):227–233.

Helm JE, AhmadBeygi S, Van Oyen MP (2011) Design and analysis of hospital admission control for operational effectiveness. *Production Oper. Management* 20(3):359–374.

Hong J, Tan X, Towsley D (1989) A performance analysis of minimum laxity and earliest deadline scheduling in a real-time system. *IEEE Trans. Comput.* 38(12):1736–1744.

Hu B, Benjaafar S (2009) Partitioning of servers in queueing systems during rush hour. *Manufacturing Service Oper. Management* 11(3):416–428.

Huang J, Carmeli B, Mandelbaum A (2015) Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Oper. Res.* 63(4):892–908.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2017) Discretionary task ordering: Queue management in radiological services. *Management Sci.* 64(9):4389–4407.

Imperato J, Morris DS, Binder D, Fischer C, Patrick J, Sanchez LD, Setnik G (2012) Physician in triage improves emergency department patient throughput. *Intern. Emerg. Medicine* 7(5): 457–462.

Iserson KV, Moskop JC (2007) Triage in medicine, part I: Concept, history, and types. *Ann. Emerg. Medicine* 49(3):275–281.

Jackson JR (1960) Some problems in queueing with dynamic priorities. *Naval Res. Logist. Quart.* 7(3):235–249.

Kc DS (2013) Does multitasking improve performance? Evidence from the emergency department. *Manufacturing Service Oper. Management* 16(2):168–183.

Kleinrock L (1964) A delay dependent queue discipline. *Naval Res. Logist. Quart.* 11(3–4):329–341.

Louviere JJ, Hensher DA (1983) Using discrete choice models with experimental design data to forecast consumer demand for a unique cultural event. *J. Consumer Res.* 10(3):348–361.

Louviere JJ, Hensher DA, Swait JD (2000) *Stated Choice Methods: Analysis and Applications* (Cambridge University Press, Cambridge, UK).

Mandelbaum A, Stolyar AL (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* 52(6):836–855.

McFadden D (1973) Conditional logit analysis of qualitative choice behavior. Zarembka P, ed. *Frontiers in Econometrics* (Academic Press, New York), 105–142.

Mendelson H, Whang S (1990) Optimal incentive-compatible priority pricing for the m/m/1 queue. *Oper. Res.* 38(5):870–883.

Muggeo VM (2003) Estimating regression models with unknown break-points. *Statist. Medicine* 22(19):3055–3071.

Murray M, Bullard M, Grafstein E, the CTAS (2004) Revisions to the Canadian Emergency Department Triage and Acuity Scale implementation guidelines. *Can. J. Emerg. Medicine* 6(6):421–427.

Newey WK, McFadden D (1994) Large sample estimation and hypothesis testing. Engle RF, McFadden DL, eds. *Handbook of Econometrics* (Elsevier, Amsterdam), 2111–2245.

Olivares M, Terwiesch C, Cassorla L (2008) Structural estimation of the newsvendor model: An application to reserving operating room time. *Management Sci.* 54(1):41–55.

Ospina MB, Bond K, Schull M, Innes G, Blitz S, Rowe BH (2007) Key indicators of overcrowding in Canadian emergency departments: A Delphi study. *Can. J. Emerg. Medicine* 9(5):339–346.

Ostendorf M, Buskens E, van Stel H, Schrijvers A, Marting L, Dhert W, Verbout A (2004) Waiting for total hip arthroplasty: Avoidable loss in quality time and preventable deterioration. *J. Arthroplasty* 19(3):302–309.

Parlar M, Sharafali M (2014) Optimal design of multi-server Markovian queues with polynomial waiting and service costs. *Appl. Stochastic Models Bus. Indust.* 30(4):429–443.

Pines JM, Hilton JA, Weber EJ, Alkemade AJ, Al Shabanah H, Anderson PD, Bernhard M, et al. (2011) International perspectives on emergency department crowding. *Acad. Emerg. Medicine* 18(12):1358–1370.

Rowe BH, Guo X, Villa-Roel C, Schull M, Holroyd B, Bullard M, Vandermeer B, Ospina M, Innes G (2011) The role of triage liaison

physicians on mitigating overcrowding in emergency departments: A systematic review. *Acad. Emerg. Medicine* 18(2):111–120.

Saghafian S, Austin G, Traub SJ (2015) Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Trans. Healthcare Syst. Engrg.* 5(2):101–123.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Oper. Res.* 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 Oper. Management* 16(3):329–345.

Sharif AB, Stanford DA, Taylor P, Ziedins I (2014) A multi-class multi-server accumulating priority queue with application to health care. *Oper. Res. Health Care* 3(2):73–79.

Smith WE (1956) Various optimizers for single-stage production. *Naval Res. Logist. Quart.* 3(1–2):59–66.

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

Stanford DA, Taylor P, Ziedins I (2014) Waiting time distributions in the accumulating priority queue. *Queueing Syst.* 77(3):297–330.

Tan T, Staats BR (2017) Behavioral drivers of routing decisions: Evidence from restaurant table assignment. Working paper, Southern Methodist University, Dallas.

Tassiulas L, Ephremides A (1992) Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Autom. Control* 37(12):1936–1948.

Train K (2000) Halton Sequences for Mixed Logit. Unpublished manuscript, Department of Economics, University of California, Berkeley.

Van Mieghem JA (1995) Dynamic scheduling with convex delay costs: The generalized $c|\mu$ rule. *Ann. Appl. Probab.* 5(3):809–833.

Van Mieghem JA (2003) Due-date scheduling: Asymptotic optimality of generalized longest queue and generalized largest delay rules. *Oper. Res.* 51(1):113–122.

Yu Q, Allon G, Bassamboo A (2016) How do delay announcements shape customer behavior? An empirical study. *Management Sci.* 63(1):1–20.

Zink BJ (2006) *Anyone, Anything, Anytime: A History of Emergency Medicine* (Elsevier/Mosby, Philadelphia).