Submitted to *Manufacturing & Service Operations Management*
manuscript MSOM-16-311

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

Yichuan Ding
Sauder School of Business, University of British Columbia, Vancouver, British Columbia V6T 1Z2, daniel.ding@sauder.ubc.ca

Eric Park
Faculty of Business and Economics, University of Hong Kong, Pokfulam, Hong Kong, ericpark@hku.hk

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

Eric Grafstein
Regional Head, Department of Emergency Medicine, Providence Health Care and Vancouver Coastal Health, Chair, Regional Emergency Services Program, Providence Health Care and Vancouver Coastal Health, Eric.Grafstein@vch.ca

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 (CTAS), 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 multi-class 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 (FCFS) 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.

Key words: Triage, Emergency Department, Multi-class Queue, Dynamic Priority, Discrete Choice, Structural Estimation, Public Policy, generalized $c\mu$ rule

History: Received: September 7, 2016; Revised: August 21, 2017 and December 31, 3017; Accepted: January 12. 2018

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 United States health care system has been well documented (Derlet and Richards 2000). Canada, which operates a universal publicly funded health care 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) proposes a *fractile response objective*. As an example, CTAS guidelines state that "95% of triage level-2 patients need to be seen within 15 minutes upon arrival" 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 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 pre-determined rules such as 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 level-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 health care 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 multi-class queuing system. We explore this decision in two dimensions, i.e., first within the same

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	Non-urgent	120 minutes	80%

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 over 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 multi-class 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, in order 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 where 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.

The above 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 her 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 patient treatment process than non-physician decision makers.

While our empirical findings apply immediately to ED operations, our proposed structural estimation framework can analyze prioritization behaviors in other service systems which 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; 2) Prioritization

guidelines are either absent or not detailed enough due to 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, non-profit hospitals and public health care systems, and NGOs. 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 judgement of patient urgencies among other factors. Our framework provides a tool to understand how the above 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 multi-class queueing literature and the second is the literature on ED operations. We review relevant papers below.

2.1. Scheduling in Multi-Class Queues

Our work is closely related to the extensive literature on queuing and job scheduling which 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 non-decreasing 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 when 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, $\Pr(\tau>D_k)$. Since 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 minimizes (d) in heavy traffic among the class of work-conservation policies, while both GLQ and GLD can be regarded as special forms of the $Gc\mu$-rule. Cost structure of 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 the above stream of queueing literature by providing an empirical understanding of the possible objective functions that are used in scheduling multi-class patients in typical Canadian EDs. This may open the door for important theoretical work and subsequent empirical studies on scheduling multi-class jobs.

2.2. ED Operations

ED as a general application has gained significant attention in the OM literature in recent years, e.g., (Kc 2013, 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 the above 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 where physician capacity is pooled compared to 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 multi-class

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 Stanford et al.'s (2014) result to a multi server 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. Canadian Triage and Acuity Scale (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 EDs to evaluate their acuity level, resource needs, and performance against certain operating objectives." 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." The rational behind this is that "the fractile response is a way of describing how often a system operates within its stated objectives" (CAEP 2014).

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 (US) 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 US, 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, i.e., 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 as inoperable at most of 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 (CIHI) and the performance of each ED can be obtained through the publicly available National Ambulatory Care Reporting System (NACRS) Metadata. Hence, the ED decision makers wary of public perception on 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 each 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 is at the patient visit level where 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 lastly 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

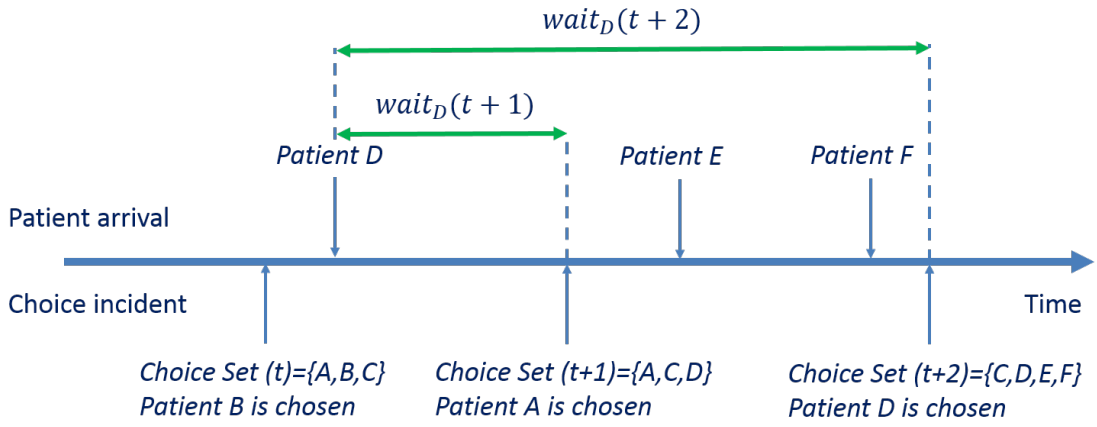


Figure 1 **Choice Incidents and Evolvement of Choice Sets**

as name, age, sex, and personal health number (PHN)), 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, i.e., physicians are the bottleneck resources in the treatment process. And the "call in" is triggered by the time when the physician finds that she 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 graphical 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 $t+1$.

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 (mins)	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 (mins)	418.8	287.8	407.0	250.5	436.1	268.2	444.1	247.8
Age	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%

Note: 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.
* percentage among all five triage levels.

In all four EDs, triage level-4 and 5 patients have a dedicated fast-track service line that operates 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 afterwards. For this reason, the *enter time* and *selection time* of triage level-1 patients may not be accurately recorded. Due to 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 $wait_D(t+1)$ till choice incident $t+1$, but was not chosen. At choice incident $t+2$ when she was chosen, her wait time was $wait_D(t+2)$. The longest wait time observed in the study EDs is 720 mins.

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 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. Since 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 about 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 due to 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 to that phenomenon is that the ED decision makers may have other objectives during the shift change, such as preventing the physician from working over time. 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 non-decreasing in wait time before seeing a physician, and can be any constant value afterwards.*

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 waiting period and treatment period, respectively. We define the marginal holding cost during the waiting period as the *marginal waiting cost*. In many health care 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 non-negative and non-decreasing in the patient's wait time. Without loss of generality, we can assume that the marginal waiting cost is continuous, since, 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 which puts 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 non-decreasing 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 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 of the times, hence, regarded as the "servers" in the ED queueing system. In our consultation with several ED physicians both from study and non-study 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 (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 $10am$ to $2am$ in the next day. Thus, we keep and analyze choice incidents in 10am–2am only, when the EDs can be regarded as critically loaded.

With the above four observations, the ED decision makers can be considered to be minimizing the total cumulative holding cost in a multi-class queueing system with multiple homogeneous servers, where the marginal holding cost exhibits certain properties. Furthermore, by our discussion in 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 where 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,

$$V_{ijt}(wait_j(t), \boldsymbol{X_j}, \boldsymbol{Y_i}) = c_j^i(wait_j(t), \boldsymbol{X_j}, \boldsymbol{Y_i})\mu_j^i. \tag{1}$$

$c_j^i(wait_j(t), \boldsymbol{X_j}, \boldsymbol{Y_i})$ 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 while there are no publicly documented guidelines on how to manage the routing duties, EDs are expected to maintain consistency in their operations. Hence, while 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 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 herself 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

$$P(j|\boldsymbol{\Sigma}(t)) = \frac{\exp(V_{jt}(wait_j(t), \boldsymbol{X_j}))}{\sum_{p \in ChoiceSet(t)} \exp(V_{pt}(wait_p(t), \boldsymbol{X_p}))}, \tag{4}$$

where

$$\boldsymbol{\Sigma}(t) := \{(wait_p(t), \boldsymbol{X_p}) | p \in ChoiceSet(t)\} \tag{5}$$

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

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

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, $V_{jt}(wait_j(t), \boldsymbol{X_j})$, 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 is 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, $c_j(wait_j(t), \boldsymbol{X_j})$, incorporates both urgency and acuity of the patient, the former captured by wait time, $wait_j(t)$, and latter by the fixed characteristics in $\boldsymbol{X_j}$. 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 to 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 her 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 where 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 since service rate μ_j is not utilized in the decision-making process. Second, we fit a *Complexity(triage)-based* model where complexity of treating patient j is assessed at a coarse patient triage level, $Tri(j)(=2,3)$, that is, $\mu_j = \mu_{Tri(j)}$. Lastly, we fit a *Complexity(CCD)-based* model where 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 the above 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 $c_j(wait_j(t), \boldsymbol{X_j})$ in Equation (2) into two parts, $f_w^{Tri(j)}(wait_j(t))$ and $f_c(\boldsymbol{X_j})$. 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:

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

The decomposition of $c_j(wait_j(t), \boldsymbol{X_j})$ 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, $f_w^{Tri(j)}(wait_j(t)) = \beta_1^{Tri(j)}$, which 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 non-linear (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:

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

$\beta_1^{Tri(j)}$ 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 non-existence 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 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.

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

ED	Marginal waiting cost function	df	Complexity information					
			No complexity (Urgency only)		Complexity (Triage)		Complexity (CCD)	
			Log-likelihood	BIC	Log-likelihood	BIC	Log-likelihood	BIC
A	Constant	125	-110282	222202	-110339	222314	-110438	222513
	Linear	127	-107549	216761	-107610	216882	-107756	217175
	Quadratic	129	-106681	215052	-106747	215182	-107008	215704
	Cubic	131	-106358	214432	-106425	214565	-106674	215063
	Piece-wise linear	131	-105692	213099	-105760	213235	-105998	213712
B	Constant	125	-62943	127423	-62943	127423	-62978	127493
	Linear	127	-58505	118571	-58506	118574	-59109	119779
	Quadratic	129	-53791	109168	-53797	109179	-56186	113957
	Cubic	131	-53045	107701	-53053	107716	-55696	113003
	Piece-wise linear	131	-51884	105378	-51892	105394	-53237	108085
C	Constant	125	-63287	128123	-63294	128138	-63369	128287
	Linear	127	-58985	119545	-58995	119564	-59502	120579
	Quadratic	129	-56676	114951	-56697	114994	-57933	117465
	Cubic	131	-56238	114099	-56261	114146	-57617	116857
	Piece-wise linear	131	-54799	111223	-54821	111267	-55812	113247
D	Constant	125	-47536	96579	-47541	96588	-47632	96771
	Linear	127	-44735	91001	-44733	90997	-45406	92343
	Quadratic	129	-43989	89533	-43992	89539	-45017	91589
	Cubic	131	-43843	89264	-43846	89271	-44870	91319
	Piece-wise linear	131	-43254	88087	-43255	88088	-44171	89921

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 tradeoff 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 to 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 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 non-linearity 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.

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				

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

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 the estimated break-point, γ_1, respectively. 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.

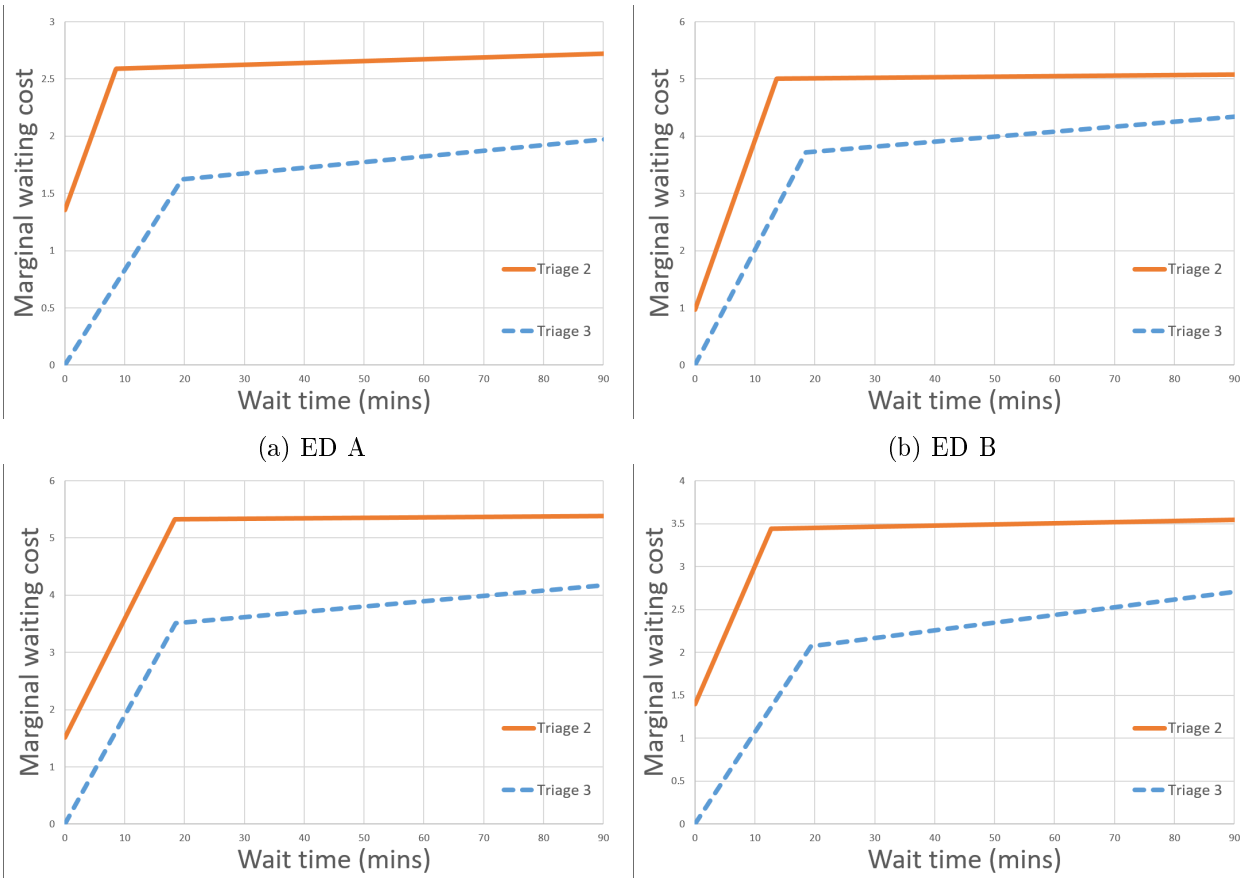
For example, for triage level-2 patients in ED B with wait times longer than 13.6 minutes, the slope of the marginal waiting cost, $\beta_1 + \beta_2$, is close to zero, which indicates that the marginal waiting cost is nearly a constant above the break-point and waiting an extra minute increases the 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 (Figure 3).

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, due to 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



(a) ED A (b) ED B

(c) ED C (d) ED D

Figure 2 Piece-wise Linear Marginal Waiting Cost Function by Triage Level

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 $\mathbb{E}[V_{jt}(wait, \boldsymbol{X_j})|Tri(j) = l]$ as a function of $wait$ for $l = 2, 3$. The functional value of $\mathbb{E}[V_{jt}(wait, \boldsymbol{X_j})|Tri(j) = l]$ 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 $\mathbb{E}[V_{jt}(wait, \boldsymbol{X_j})|Tri(j) = l] = f_w^l(wait) + \mathbb{E}[f_c(\boldsymbol{X_j})|Tri(j) = l]$ by Equation (2) and (7), where $f_w^l(wait)$ is the estimated piece-wise linear function (Equation (8)); and $\mathbb{E}[f_c(\boldsymbol{X_j})|Tri(j) = l]$ 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 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

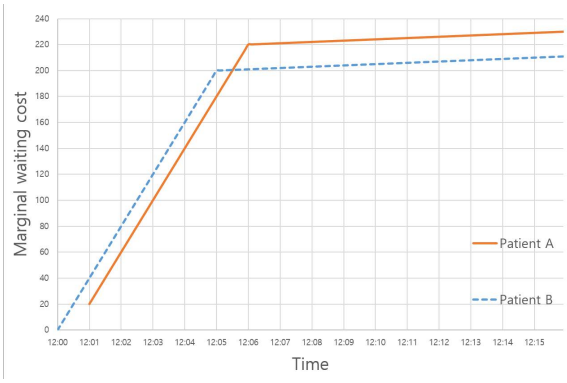


Figure 3 Delay-dependent Patient Prioritization: An Example of Two Patients in Same Triage Level

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 to the random shock ϵ_{jt}, whose standard deviation has been fixed as one unit. For example, in 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, i.e., 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 where 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 3, whereas a triage level-2 patient who has waited less than 8 minutes has a marginal waiting cost smaller than 3. 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 non-homogeneous 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 individual patient's characteristics. For example, in ED A, even though triage level-2 patients who waited more than 5 minutes dominate triage level-3 patients who waited less than 90 minutes, certain triage level-2 patients that waited longer than 5 minutes can have lower marginal waiting costs than triage level-3 patients that waited less than 90 minutes.

In order 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 5 minutes and a slope of 1 above the break-point. The wait time-independent fixed attributes term of Patient A is larger than that of Patient B by 20, i.e., $f_c(\boldsymbol{X_A}) - f_c(\boldsymbol{X_B}) = 20$, but Patient B has arrived 1 min 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 to patient A, hence the FCFS principle is more likely to be adhered. This is due to the fact that patient B receives additional priority (40) by arriving 1 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, 1, 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), Olivares et al. (2008). Like 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 $j^* \in \arg\max\{U_{ijt}|j \in ChoiceSet(t)\}$. 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 which 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. Since 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 Appendices, including decision maker heterogeneity (Appendix C), unobserved patient heterogeneity (Appendix D), the Independence from Irrelevant Alternatives (IIA) property of the conditional logit model (Appendix E), and number of break-points in the piecewise linear specification of $f_w^{Tri}(\cdot)$ (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 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 exiting results on the consistency of MLE for generalized linear models (e.g. Fahrmeir and Kaufmann (1985), Newey and McFadden (1994)), because the $f_w^{Tri}(\cdot)$ term in our model can be nonlinear and non-smooth, 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 attached in Appendix G. We also show that the distribution of MLE for our model is generally not asymptotically normal due to 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 breakpoint, that is, 13.3 minutes for triage level-2 patients and 18.9 minutes for triage level-3 patients, on average. This might be due to the reduced incentive of the decision maker to choose the patient who has waited the longest from among those that 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 an urgency and fairness standpoint, for patients in the same triage level, treating those who have waited longer is the reasonable expectation. While 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 patient wait times. This suggests that the ED decision makers are making sophisticated routing decisions in the sense of not just

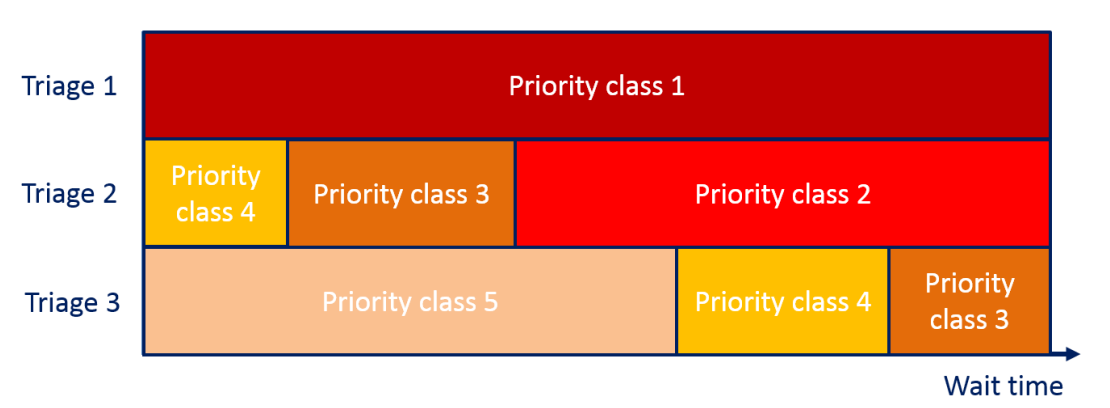


Figure 4 **Delay-dependent Patient Prioritization in Triage Systems**

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 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, where 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 her 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 can 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.

Lastly, 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 non-physician 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 her 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 we presented our results to, 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 US 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 multi-class 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 multi-class 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 she 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

We thank two anonymous reviewers, the associate editor, and the area editor for their many helpful suggestions. We also thank Keith Head, Garth Hunte, and MacKenzie Winston for their valuable comments. The research of the first and the third author is partially supported by National Sciences and Engineering Research Council of Canada (NSERC) PGPIN 436156-13 and 13R81646, respectively. The fourth author thanks the Vancouver Coastal Health for its generous support.

Endnotes

1. In this paper, patient routing includes both 1) prioritization across different triage levels and 2) service disciplines within the same triage level, e.g., first-come first-served (FCFS) 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, e.g., Aksin et al. (2013), Yu et al. (2016). These papers aim at understanding 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 2016).

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

A. Agresti. *An introduction to categorical data analysis*, volume 135. Wiley New York, 1996.

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

A. Antoniadis, I. Gijbels, and M. Nikolova. Penalized likelihood regression for generalized linear models with non-quadratic penalties. *Annals of the Institute of Statistical Mathematics*, 63(3):585–615, 2011.

M. Armony, S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, G. B. Yom-Tov, et al. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, 5(1):146–194, 2015.

R. Batt and C. Terwiesch. Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Sci. Forthcoming*, 2016.

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

S. L. Bernstein, D. Aronsky, R. Duseja, S. Epstein, D. Handel, U. Hwang, M. McCarthy, K. John McConnell, J. M. Pines, N. Rathlev, et al. The effect of emergency department crowding on clinically oriented outcomes. *Academic Emergency Medicine*, 16(1):1–10, 2009.

R. Beveridge. Caep issues. the canadian triage and acuity scale: A new and critical element in health care reform. canadian association of emergency physicians. *Journal of Emergency Medicine*, 16(3):507–11, 1998.

M. J. Bullard, B. Unger, J. Spence, and E. Grafstein. Revisions to the canadian emergency department triage and acuity scale (ctas) adult guidelines. *Cjem*, 10(02):136–142, 2008.

M. J. Bullard, T. Chan, C. Brayman, D. Warren, B. Unger, C. N. W. Group, et al. Revisions to the canadian emergency department triage and acuity scale (ctas) guidelines. *CJEM*, 16(06):485–489, 2014.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2002.

CAEP. CTAS implementation guidelines. Retrieved Feb 25, 2015, from `http://caep.ca/resources/ctas/implementation-guidelines`, 2014.

M. A. Cohen, T. H. Ho, Z. J. Ren, and C. Terwiesch. Measuring imputed cost in the semiconductor equipment supply chain. *Management Science*, 49(12):1653–1670, 2003.

R. W. Derlet and J. R. Richards. Overcrowding in the nations emergency departments: complex causes and disturbing effects. *Annals of emergency medicine*, 35(1):63–68, 2000.

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

D. B. Diercks, M. T. Roe, A. Y. Chen, W. F. Peacock, J. D. Kirk, C. V. Pollack, W. B. Gibler, S. C. Smith, M. Ohman, and E. D. Peterson. 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. *Annals of emergency medicine*, 50(5):489–496, 2007.

P. Dierckx. *Curve and surface fitting with splines*. Oxford University Press, 1995.

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

T. A. Domencich and D. McFadden. Urban travel demand-a behavioral analysis. Technical report, 1975.

A. J. Drummond. No room at the inn: overcrowding in ontarios emergency departments. *CJEM*, 4(02): 91–97, 2002.

L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, pages 342–368, 1985.

T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

L. Graff. Overcrowding in the ed: an international symptom of health care system failure. *The American journal of emergency medicine*, 17(2):208–209, 1999.

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

A. Guttmann, M. J. Schull, M. J. Vermeulen, T. A. Stukel, et al. Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from ontario, canada. *Bmj*, 342:d2983, 2011.

J. H. Han, D. J. France, S. R. Levin, I. D. Jones, A. B. Storrow, and D. Aronsky. The effect of physician triage on emergency department length of stay. *The Journal of emergency medicine*, 39(2):227–233, 2010.

J. E. Helm, S. AhmadBeygi, and M. P. Van Oyen. Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management*, 20(3):359–374, 2011.

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

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

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

M. R. Ibanez, J. R. Clark, R. S. Huckman, and B. R. Staats. Discretionary task ordering: Queue management in radiological services. *Management Science*, 2017.

J. Imperato, D. S. Morris, D. Binder, C. Fischer, J. Patrick, L. D. Sanchez, and G. Setnik. Physician in triage improves emergency department patient throughput. *Internal and emergency medicine*, 7(5): 457–462, 2012.

K. V. Iserson and J. C. Moskop. Triage in medicine, part i: concept, history, and types. *Annals of emergency medicine*, 49(3):275–281, 2007.

J. R. Jackson. Some problems in queueing with dynamic priorities. *Naval Research Logistics Quarterly*, 7 (3):235–249, 1960.

D. S. Kc. Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management*, 16(2):168–183, 2013.

L. Kleinrock. A delay dependent queue discipline. *Naval Research Logistics Quarterly*, 11(3-4):329–341, 1964.

T. Li. Econometric analysis of cross section and panel data (book). *Journal of economic literature*, 40(4): 1239–1241, 2002.

J. J. Louviere and D. A. Hensher. Using discrete choice models with experimental design data to forecast consumer demand for a unique cultural event. *Journal of Consumer research*, 10(3):348–361, 1983.

J. J. Louviere, D. A. Hensher, and J. D. Swait. *Stated choice methods: analysis and applications*. Cambridge University Press, 2000.

D. Lowsky, Y. Ding, D. Lee, C. McCulloch, L. Ross, J. Thistlethwaite, and S. Zenios. Ak-nearest neighbors survival probability prediction method. *Statistics in medicine*, 32(12):2062–2069, 2013.

A. Mandelbaum and A. L. Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research*, 52(6):836–855, 2004.

J. H. McDonald. *Handbook of biological statistics*, volume 2. Sparky House Publishing Baltimore, MD, 2009.

D. McFadden. Conditional logit analysis of qualitative choice behavior. 1973.

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

V. M. Muggeo. Estimating regression models with unknown break-points. *Statistics in medicine*, 22(19): 3055–3071, 2003.

M. Murray, M. Bullard, E. Grafstein, et al. Revisions to the canadian emergency department triage and acuity scale implementation guidelines. *Can J Emerg Med*, 6(6):421–27, 2004.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

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

M. B. Ospina, K. Bond, M. Schull, G. Innes, S. Blitz, and B. H. Rowe. Key indicators of overcrowding in canadian emergency departments: a delphi study. *CJEM*, 9(05):339–346, 2007.

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

I. Pardoe and D. K. Simonton. Applying discrete choice models to predict academy award winners. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2):375–394, 2008.

M. Parlar and M. Sharafali. Optimal design of multi-server markovian queues with polynomial waiting and service costs. *Applied Stochastic Models in Business and Industry*, 30(4):429–443, 2014.

J. M. Pines, J. A. Hilton, E. J. Weber, A. J. Alkemade, H. Al Shabanah, P. D. Anderson, M. Bernhard, A. Bertini, A. Gries, S. Ferrandiz, et al. International perspectives on emergency department crowding. *Academic Emergency Medicine*, 18(12):1358–1370, 2011.

J. M. Pines, C. R. Carpenter, A. S. Raja, and J. D. Schuur. The epidemiology and statistics of diagnostic testing. *Evidence-Based Emergency Care: Diagnostic Testing and Clinical Decision Rules, Second Edition*, pages 19–35, 2012.

B. H. Rowe, X. Guo, C. Villa-Roel, M. Schull, B. Holroyd, M. Bullard, B. Vandermeer, M. Ospina, and G. Innes. The role of triage liaison physicians on mitigating overcrowding in emergency departments: a systematic review. *Academic Emergency Medicine*, 18(2):111–120, 2011.

D. Ruppert and R. J. Carroll. Penalized regression splines. Technical report, Cornell University Operations Research and Industrial Engineering, 1999.

S. Saghafian, W. J. Hopp, M. P. Van Oyen, J. S. Desmond, and S. L. Kronick. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research*, 60(5):1080–1097, 2012.

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

S. Saghafian, G. Austin, and S. J. Traub. Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering*, 5(2):101–123, 2015.

A. B. Sharif, D. A. Stanford, P. Taylor, and I. Ziedins. A multi-class multi-server accumulating priority queue with application to health care. *Operations Research for Health Care*, 3(2):73–79, 2014.

W. E. Smith. Various optimizers for single-stage production. *Naval Research Logistics Quarterly*, 3(1-2): 59–66, 1956.

H. Song, A. L. Tucker, and K. L. Murrell. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science*, 2015.

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

M.-C. Sung, D. C. McDonald, and J. E. Johnson. Probabilistic forecasting with discrete choice models: Evaluating predictions with pseudo-coefficients of determination. *European Journal of Operational Research*, 248(3):1021–1030, 2016.

T. Tan and B. R. Staats. Behavioral drivers of routing decisions: Evidence from restaurant table assignment. *Working Paper*, 2016.

T. G. Tape. *The Area Under an AUC Curve*. University of Nebraska Medical Center. URL `http://gim.unmc.edu/dxtests/Default.htm`.

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

K. Train. Halton sequences for mixed logit. *Department of Economics, UCB*, 2000.

K. E. Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

J. A. Van Mieghem. Dynamic scheduling with convex delay costs: The generalized c— mu rule. *The Annals of Applied Probability*, pages 809–833, 1995.

J. A. Van Mieghem. Due-date scheduling: Asymptotic optimality of generalized longest queue and generalized largest delay rules. *Operations Research*, 51(1):113–122, 2003.

Q. Yu, G. Allon, and A. Bassamboo. How do delay announcements shape customer behavior? an empirical study. *Management Science*, 2016.

B. J. Zink. *Anyone, anything, anytime: a history of emergency medicine*. Elsevier Health Sciences, 2006.

Appendix A: Asymptotic Optimality of the $Gc\mu$-Rule

The $Gc\mu$ rule has been proved asymptotically optimal (Mandelbaum and Stolyar 2004, Gurvich and Whitt 2009) in a multi-class queueing system with non-decreasing marginal holding cost. However, according to Observation 2, the marginal holding cost can drop to a constant (e.g., zero) when the treatment starts. This violates the non-decreasing assumption of the $Gc\mu$-rule. To reconcile this, we consider a new system where the marginal holding cost during the treatment period has been shifted up by a large constant $\bar{c}$. According to our observations, $c_j(wait_j(t), \boldsymbol{X_j})$ is continuous, $\boldsymbol{X_j}$ has finite support, and $wait_j(t)$ is bounded, hence, we can choose a sufficiently large $\bar{c}$ such that $\bar{c} > c_j(wait_j(t), \boldsymbol{X_j})$ for all patient and wait times. As a result, the marginal holding cost in the new system satisfies the non-decreasing assumption and the asymptotic optimality of the $Gc\mu$-rule can be proved. Since the total holding costs in the new system and original system always differs by a constant ($\bar{c} *$ Total treatment time for all patients) when the same routing policy is used, the cost minimization problems in the two systems are equivalent. That means, if the $Gc\mu$ is asymptotically optimal in the new system, it must be asymptotically optimal in the original system as well. Therefore, allowing the marginal holding cost to drop to a small constant will not undermine the asymptotic optimality of the $Gc\mu$-rule.

Appendix B: No Skill-Based Patient Routing: Conditional Independence Test

Table 5 Independence Test Between Physician IDs and Triage Levels Conditional on Hour-of-Day and Weekday/Weekend

ED	Likelihood Ratio Statistic	p-value	Pearson Statistic	p-value	No. of Obs.	df
A	5384.11	0.799	5394.41	0.770	87,158	5,472
B	4676.08	1.000	5307.55	0.392	69,703	5,280
C	3017.18	1.000	3390.68	1.000	57,302	3,744
D	2668.71	0.144	2672.68	0.132	48,687	2,592

We excluded physician IDs that appear only occasionally in the 20-month study period (less than 10 days). Since those IDs have only treated a few patients, the independence test between those IDs and the triage levels would not be statistically meaningful. We perform independence tests between the remaining physician IDs and the patient triage levels treated by those physician IDs to explore whether more acute and difficult patients are likely assigned to certain physicians. Due to the fact that both physician shifts and patient triage levels have a pattern by hour-of-day and weekday/weekend, we test the independence conditional on hour-of-day and whether it is a weekday (Mon–Fri) or weekend (Sat, Sun). Table 5 reports the independence test results conditional on the hour-of-day and weekday/weekend combination, which contains 48 cells within a week (2 types of day * 24 hours). In our data, the expected counts in each cell is greater than 5 observations in all four EDs. Thus, according to the conventional rule of thumb (McDonald 2009), both the G-test (likelihood ratio statistic) and Chi-square test (Pearson statistic) are acceptable for independence test. For both tests, the null hypothesis of independence between physician IDs and triage levels cannot be rejected at the 5% level of significance. This is consistent with what we have learned from the ED physicians and administrators

that the assigning more acute or more difficult patients to certain physicians is not the discipline in the study EDs and in general. Using the same test methods, we also find independence between physician IDs and patient Chief Complaint System (CCS) codes which classify patients at the clinical department level (the minimum p-value is 0.617 for all four EDs).

Appendix C: Decision Maker Heterogeneity

As alluded to in Section 4.1, we expect every ED to have consistency in the patient routing decisions, and assume a single decision maker in each ED. We test whether our findings are robust when it comes to potential decision maker heterogeneity. Our approach is to estimate a random coefficients model also known as mixed logit, where the coefficients of interest are allowed to vary by a parametric structure (normal distribution) across individual choice makers. We estimate the normally distributed random triage-level intercepts and slopes of the piece-wise linear marginal waiting cost function (Equation (6)) with the break-points fixed at the locations from the non-random model reported in Table 5. The mixed logit model also relaxes the IIA property of the conditional logit model and allows correlation across valuation of patients in the same choice incident. However, the information necessary to identify the choice maker, such as work shift schedules of ED personnel, is not available.

We take two different approaches to isolate the decision maker's identity. First, we approximate identity by work shift combinations of day-of-the-week and day-night groupings. For instance, we treat Monday-day, Monday-night, and Tuesday-day as different shifts. Hence there are a total of 14 shifts per week. Second, we use the masked physician ID information for each patient visit. We use this as the identifier of possible decision maker heterogeneity. Both estimation results show that decision maker heterogeneity is statistically insignificant at the 5% level.

Appendix D: Unobserved Patient Heterogeneity

Our data contains rich information for each individual patient which includes CCD, age, sex, method of arrival, and discharge decision. This allows us to successfully control patient heterogeneity. Yet, there still may be patient characteristics that affect the decision makers' patient choice but are not observed by the researcher. An example may include extreme medical conditions requiring special resources that are not captured by the control variables. If so, omitted variable bias may be a concern in Equation (3), as it violates the iid assumption of the error term ϵ in the conditional logit model.

Our approach in this regard is to model the unobserved heterogeneity as a random intercept,

$$\pi_j \sim \mathcal{N}(0, \sigma_\pi^2), \tag{9}$$

which is associated with patient j and is consistent across choice incident t. The valuation of choosing patient j at choice incident t then has the following expression

$$V_{jt}(\pi_j, wait_j(t), \boldsymbol{X_j}) = (\pi_j + f_w^{Trj(j)}(wait_j(t)) + f_c(\boldsymbol{X_j}))\mu_j. \tag{10}$$

The consistency in choice incidents addresses possible serial correlation in patient valuation across different choice incidents. The likelihood of observing the sequence of choices is given by

$$L = \int \prod_t P(c(t)|\boldsymbol{\Sigma}(t)) f(\pi \mid \sigma_\pi) d\pi, \tag{11}$$

Table 6 Robust Analysis: Estimation Results of Unobserved Patient Heterogeneity Term

	ED A	ED B	ED C	ED D
σ_π	0.0085	0.0058	0.0099	0.0123
	(0.1181)	(0.0821)	(0.1262)	(0.1351)

Standard errors in parentheses.

where choice probability, $P(c(t)|\mathbf{\Sigma}(t))$ is equivalent to Equation (4) with Equation (10) as the valuation term. Unfortunately, the integral in Equation (11) does not have a closed form. Hence, we cannot compute the likelihood function exactly. Instead, we approximate the choice probabilities through simulation and maximize the simulated log-likelihood function. We take R number of draws from $f(\pi \mid \sigma_\pi)$ for each patient and let $\pi_j^{r|\sigma_\pi}$ denote the $r-$th draw of patient j. The simulated log-likelihood function of the observed choice sequence is constructed as

$$\ln SL = \ln \frac{1}{R} \sum_{r=1}^{R} \prod_t P_t(c(t)|\pi_j^{r|\sigma_\pi}, wait_j(t), X_j \ \forall j \in ChoiceSet(t)). \tag{12}$$

The estimation of Equation (12) is computationally difficult as we cannot take advantage of the log-transformation in log-likelihood functions. The dimension of t, the number of choice incidents, is large in all EDs we studied, ranging from 31,427 to 56,604. Hence, the simulated probability of observing the choice sequence, $\prod_t P(c(t)|\pi_{c(t)}^{r|\sigma_\pi}, wait_j(t), X_j \ \forall j \in ChoiceSet(t))$, is very small and brings in computational challenge.

In order to circumvent this problem, we propose an alternative model where we group the choice incidents by each calendar day and assume that the random error term of unobserved patient heterogeneity is drawn from a distribution each day instead of the entire sample path. Driven by Observation 4, we already excluded choice incidents between 2AM and 10AM in each day, so there is no overlap of patients across different days. With this structure, the patient valuation function is:

$$V_{jdt}(\pi_{jd}, wait_j(t), \boldsymbol{X_j}) = (\pi_{jd} + f_w^{Trj(j)}(wait_j(t)) + f_c(\boldsymbol{X_j}))\mu_j, \tag{13}$$

where, $\pi_{jd} \sim \mathcal{N}_d(0, \sigma_\pi^2)$. For the grouped data, there are a total of D days and each day, d, has T_d choice incidents. Then the likelihood function can be expressed as following:

$$L = \prod_{d=1}^{D} \int \prod_{t=1}^{T_d} P(c(t)|\pi_{jd}, wait_j(t), \boldsymbol{X_j} \ \forall j \in ChoiceSet(t)) f(\pi \mid \sigma_\pi) d\pi. \tag{14}$$

And the simulated log-likelihood function for the observed choice sequence is:

$$\ln SL = \sum_{d=1}^{D} \ln \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T_d} P(c(t)|\pi_{pd}^{r|\sigma_\pi}, wait_j(t), \boldsymbol{X_j} \ \forall j \in ChoiceSet(t)). \tag{15}$$

Estimation of Equation (15) is more manageable than Equation (12) as the number of choice incidents in a day, T_d, ranges from 68 to 110. We estimate Equation (15) with a piece-wise linear marginal waiting cost function (Equation (8)) by taking 50 halton draws (Train 2000) from $\mathcal{N}_d(0, \sigma_\pi^2)$. Coefficient estimates of the piece-wise linear marginal waiting cost function are robust to our main findings in Section 5.2. The estimate of σ_π is statistically insignificant for all four EDs suggesting there is not enough evidence to support unobserved patient heterogeneity in our model (Table 6).

Appendix E: Independence from Irrelevant Alternatives (IIA) Property of The Conditional Logit Model

The conditional logit model exhibits a certain substitution pattern across alternatives which is known as the property of independence from irrelevant alternatives (IIA). Specifically, the ratio of the probabilities of patients i and k being chosen in $ChoiceSet(t)$ can be expressed as

$$\frac{P(i|\boldsymbol{\Sigma}(t))}{P(k|\boldsymbol{\Sigma}(t))} = \frac{\frac{\exp(V_{it}(wait_i(t),\boldsymbol{X_i}))}{\sum_{j\in ChoiceSet(t)}\exp(V_{jt}(wait_j(t),\boldsymbol{X_j}))}}{\frac{\exp(V_{kt}(wait_k(t),\boldsymbol{X_k}))}{\sum_{j\in ChoiceSet(t)}\exp(V_{jt}(wait_j(t),\boldsymbol{X_j}))}} = \exp(V_{it}(wait_i(t),\boldsymbol{X_i}) - V_{kt}(wait_k(t),\boldsymbol{X_k})). \quad (16)$$

The relative odds of patient i being chosen over patient k depend only on the characteristics of patients i and k, and are independent of what other patients are present in the ED at $ChoiceSet(t)$ and what characteristics the other patients have. Hence, the substitution pattern is known to be IIA. In the context of ED patient routing, the IIA property of the conditional logit model can be viewed as a restriction on the substitution pattern between two patients.

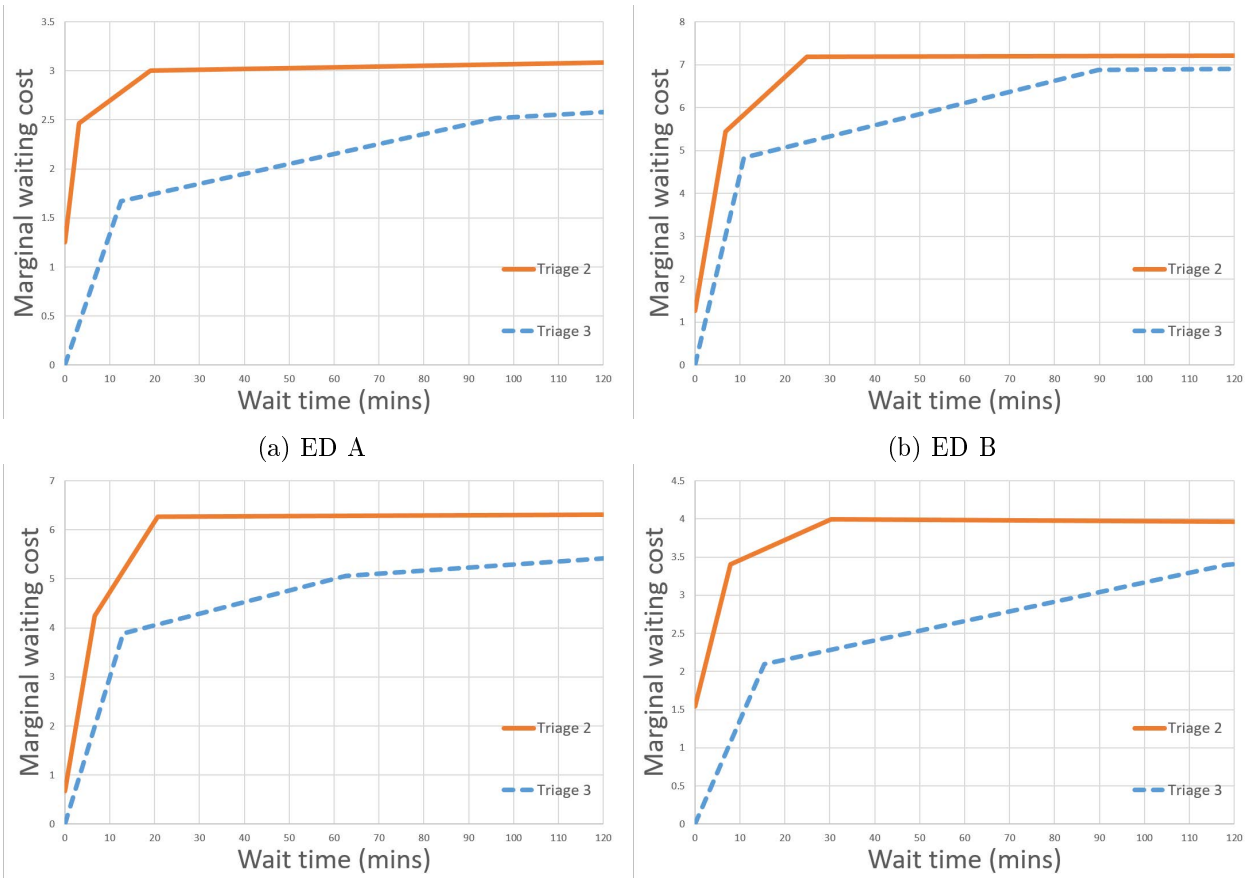
To test whether the IIA property is a reasonable assumption for the observed data, we investigate the mixed logit model, which has been discussed in Appendix C. The conditional logit model used in this paper is a special case of the mixed logit model when the random slopes and intercepts of the piece-wise linear marginal waiting cost specification have zero variance (Train 2009). After fitting the mixed logit model, the statistical insignificance of the variances at the 5% level suggests that the observed data exhibits the IIA pattern.

Appendix F: Number of Break-points in Piece-wise Linear Specification

Our conditional logit-$Gc\mu$ framework has assumed that the piece-wise linear marginal waiting cost functions, $f_w^{Tri(j)}(wait_j(t))\,\forall\, Tri(j)\in\{2,3\}$, have at most one break-point per triage level. To justify this assumption, we fit a marginal cost function with two and three break-points using the estimation method introduced in Muggeo (2003), which can identify multiple break-points. Estimation results from the two-break-point piece-wise linear marginal waiting cost functions are plotted in Figure 5. We find that the marginal waiting cost slope plateaus after the largest break-point for each triage level in a manner similar to the one-break-point model (Figure 2). Hence, the phenomena of the piece-wise linear marginal waiting cost function flattening after a threshold are robust to the number of break-points in the piece-wise specification.

Appendix G: Asymptotic Property of the MLE for the Conditional Logit-$Gc\mu$ Framework

We next prove consistency of the MLE under the conditional logit-$Gc\mu$ framework. We first provide a formal description of the general conditional logit-$Gc\mu$ framework. We strive to define the setting with sufficient generality so that it covers all models we have compared earlier (e.g., Urgency(only)-based or Complexity-based model, different functional forms of $f_w^{Tri}(\cdot)$).



(a) ED A (b) ED B

(c) ED C (d) ED D

Figure 5 Robust Analysis: Two Break-points in Piece-wise Linear Marginal Waiting Cost

Observed Data: The researcher observes a sequence of n choice incidents. In each choice set t, she observes the data for each patient's fixed attributes and waiting time, $\mathbf{\Sigma}(t)$ (defined in (5)), as well as the index of the chosen patient, $c(t)$. Therefore, the observed data for each choice incident can be summarized as $(c, \mathbf{\Sigma})$. To simplify the notation, we may drop the index (t) in the subsequent analysis when there is no ambiguity. Let Ω denote the domain of $(wait_j(t), \boldsymbol{X_j})$. Since the choice set can contain $r(=1,2,\ldots)$ patients, the domain of $\mathbf{\Sigma}$ can be expressed as $\overline{\Omega} := \cup_{r=1}^{+\infty} \Omega^r$.

Model Parameters: The choice probability is predicted using formula (4), with the deterministic value function $V_{jt}(wait_j(t), \boldsymbol{X_j})$ defined in (7). In the expression (7), we assume that the function $f_c(\boldsymbol{X_j})$ is linear and have the following form

$$f_c(\boldsymbol{X_j}) = \alpha_0 + \sum_{m=1}^{M} \alpha_k X_{jm}, \tag{17}$$

where X_{jm} denotes the value of the m^{th} attribute of patient j $(m = 1, \ldots, M)$. We assume the univariate functions, $f_w^{Tri(j)}(wait_j(t)) \, \forall \, Tri(j) \in \{2, 3\}$ are polynomial regression splines with the highest degree D and B break-points, i.e.,

$$f_w^{Tri(j)}(wait_j(t)) = \sum_{d=1}^{D} \beta_d^{Tri(j)} (wait_j(t))^d + \sum_{b=1}^{B} \beta_{D+b}^{Tri(j)} \cdot ((wait_j(t) - \gamma_b^{Tri(j)})^+)^D \tag{18}$$

Note that we assume the polynomial splines do not have a degree-0 term so $f_w^{Tri}(0)=0$ for all triage levels, as the constant intercept α_0 has already been included in the $f_c(\boldsymbol{X_j})$ function.

The polynomial regression spline is a standard tool for fitting continuous but possibly nonlinear and non-smooth functions with unknown parametric forms (Dierckx 1995, Ruppert and Carroll 1999, Antoniadis et al. 2011), and thus well serves for our purpose. It also covers all functional forms that we have discussed earlier in Section 4 and 5. For example, $D=1$, $B=1$ leads to a piece-wise linear function, and $D=3$, $B=0$ corresponds to the cubic model with no break point. This framework also covers the three patient complexity models by plugging different values of μ_j into the expression of V_{jt}.

The parameters in our model thus includes coefficients for the fixed attributes $\boldsymbol{\alpha}=\{\alpha_k|k=0,\ldots,K\}$, coefficients in the piecewise polynomials $\boldsymbol{\beta}:=\{\beta_k^{Tri}|k=1,\ldots,D+B,\ Tri=2,3\}$, and locations of the break-points $\boldsymbol{\gamma}:=\{\gamma_b^{Tri}|b=1,\ldots,B,\ Tri=2,3\}$. We use a vector $\boldsymbol{\theta}:=(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta})$ to record all the parameters. Note that the integers D and B are also parameters that we have to choose. We will first discuss the asymptomatic properties for $\boldsymbol{\theta}$, and then discuss the identification issue for D and B in the end of this section.

The MLE: Let Θ denote the candidate set of $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}^{\boldsymbol{n}}$ denote the MLE $\boldsymbol{\theta}$ for a sequence of n choice incidences, that is,

$$\hat{\boldsymbol{\theta}}^{\boldsymbol{n}}:=\arg\max\{\ln L^n(\hat{\boldsymbol{\theta}})\mid\hat{\boldsymbol{\theta}}\in\Theta\} \tag{19}$$

where

$$\ln L^n(\hat{\boldsymbol{\theta}})=\ln\prod_{t=1}^{n}P(c(t)|\boldsymbol{\Sigma}(t),\hat{\boldsymbol{\theta}})=\sum_{t=1}^{n}\ln P(c(t)|\boldsymbol{\Sigma}(t),\hat{\boldsymbol{\theta}}) \tag{20}$$

with $P(c(t)|\boldsymbol{\Sigma}(t),\hat{\boldsymbol{\theta}})$ given by (4) conditional on parameters $\hat{\boldsymbol{\theta}}$.

We prove that under some regularity conditions, $\hat{\boldsymbol{\theta}}^n$, the MLE for a sequence of n choice incidents, converges to $\boldsymbol{\theta}$, the MLE for the true log-likelihood function, when $n\to\infty$ (The proof and some discussions about the assumptions can be found at http://blogs.ubc.ca/ycding/files/2018/03/PatientChoice-Final-supplementary.pdf).

Theorem G.1 *(Consistency of MLE) Given fixed integers D and B, assume:*

- *(a1) $\{\boldsymbol{\Sigma}(t)|t=1,\ldots,n\}$ is a positive recurrent and periodically stationary Markovian process. Therefore, it is ergodic, which means there exists a limiting probability measure π, such that*

$$\frac{1}{n}\sum_{t=1}^{n}1(\boldsymbol{\Sigma}(t)\in A)\xrightarrow{p}\pi(A)\ \text{for all}\ A\subseteq\overline{\Omega}. \tag{21}$$

- *(a2) Θ is compact.*
- *(a3) The data of fixed patient attributes are not multicollinear, that is, the matrix $((1,\boldsymbol{X}_j)^T|$ all patient j observed in the data$)$ has full column rank.*
- *(a4) $\boldsymbol{X_j}$ has a finite domain; $wait_j$ has a finite upper bound $\overline{W}^{Tri(j)}$ for $Tri(j)=2,3$. (We do not observe any wait times larger than 702 mins and 720 mins for triage level-2 and -3, respectively. Thus, a reasonable upper bound can be $\overline{W}^2=702$ mins and $\overline{W}^3=720$ mins.)*
- *(a5) Any two patient attribute vectors $\boldsymbol{X_j}$ can appear in the same choice set with a positive probability.*
- *(a6) Conditional on any fixed patient attributes $\boldsymbol{X_j}$, $wait_j$ has a positive density over $[0,\overline{W}^{Tri(j)}]$.*

Then when $n\to\infty$,

$$\hat{\boldsymbol{\theta}}^{\boldsymbol{n}}\xrightarrow{p}\boldsymbol{\theta}. \tag{22}$$

Asymptotic Normality: Although $\hat{\boldsymbol{\theta}}^{\boldsymbol{n}} \xrightarrow{p} \boldsymbol{\theta}$, the asymptotic distribution of $\hat{\boldsymbol{\theta}}^{\boldsymbol{n}}$ is generally not normal. This is because the MLE can sometimes be achieved at the boundary of Θ, in which case the asymptotic distribution of the MLE has to be asymmetric and therefore not normal (see the example in p.2144 of (Newey and McFadden 1994)). To see that the MLE can be achieved at the boundary of Θ, recall the previous example in which $V_{jt}(wait_j(t), \boldsymbol{X_j}) = \beta_1 wait_j(t)$ and FCFS holds in all choice incidents. Then the MLE of β_1 is achieved at the boundary of Θ.

Appendix H: Out-of-Sample Test

To perform the out-of-sample test, we create an out-of-sample (test) data that collects all patient visits to the four study EDs during 10am-2am the next day from December 2014 to February 2015, excluding the last choice incident in each physician shift. We estimate the model coefficients (See Table 4) using the in-sample (training) data from April 2013 to November 2014, and predict the choice probability for each patient in the out-of-sample data. These predictions allow us to evaluate the prediction power of the structural estimation framework and further justify the validity of our framework replicating the ED decision makers patient routing decisions. To obtain a robust assessment, we use three different goodness-of-fit metrics.

The first metric is the McFadden's pseudo R^2 (McFadden 1973). For the same data set, a larger pseudo R^2 suggests a better fit in terms of log-likelihood. However, the pseudo R^2 heavily depends on the nature of the data set and thus is not often used as performance measure for out-of-sample test (Train 2009, Sung et al. 2016).

The second metric is the fitted probability (Louviere and Hensher 1983, Pardoe and Simonton 2008). The model marks the patient with the highest predicted probability in each choice set as the predicted choice, and calculate the percentage of correctly predicted choice sets as the fitted probability (Li 2002). The fitted probability provides a direct measure of the model's capability in identifying the actual choice. Nevertheless, in some researchers' opinions (Train 2009), choice models provide a list of predicted probabilities, rather than saying that the alternative with the highest probability must be selected. Therefore, it can be criticized that the fitted probability does not use the entire message that the model attempts to deliver.

Due to the above limitations of the pseudo R^2 and fitted probability, we consider a third metric for prediction accuracy, namely the area under the receiver operating characteristic curve (AUROC). The AUROC is a standard statistical tool to measure prediction accuracy for binary data (Fawcett 2006, Lowsky et al. 2013), and is therefore applicable to our setting in which each patient has binary outcomes: selected (positive) or not (negative). For a given threshold $\eta \in [0,1]$, patients in a choice set are marked as "selected" if their predicted probabilities are higher than η, and are marked as "not selected" otherwise. The method then calculates the true positive rate (percentage of correct predictions among the selected patients) and false positive rate (percentage of false predictions among the remaining patients). By varying η from 0 to 1, one may plot the receiver operating characteristic (ROC) curve whose X- and Y-coordinates correspond to the false and true positive rates for each η, respectively, and calculate the AUROC value. As a result, the average chance for a patient to be marked as selected for all $\eta \in [0,1]$ is proportional to her predicted probability. Therefore, the AUROC has effectively incorporated all the predicted probabilities into its assessment and is therefore better aligned with the estimation results compared to fitted probability.

We calculate the three prediction performance metrics for the *Urgency(only)-based* model with three functional forms of $f_w^{Tri}(\cdot)$ that we have considered: linear, cubic, and piece-wise linear (We did not test the constant and quadratic model, because the constant has a poor performance even for the study data, and the quadratic model is similar to the cubic). The comparison is summarized in Table 7. We find that the piece-wise linear model outperforms the other two with respect to both pseudo R^2 and AUROC. For fitted probability, the piece-wise linear model also outperforms in ED A, C, and D. In ED B, although the piece-wise linear model performs slightly worse than the cubic model, the p-value (=0.960) shows that the difference is not statistically significant. Therefore, the out-of-sample test shows that the piece-wise linear model achieves the best performance among the three models for all three test metrics, which demonstrates the robustness of the results.

Table 7 Out-of-Sample Test Statistics

ED	Marginal waiting cost function	Log-likelihood	Pseudo R^2	Fitted Probability (P-value)	AUROC (P-value)
A	Linear	-23584.4	0.065	24.4% (0.035)	0.723 (0.000)
	Cubic	-23338.4	0.075	24.4% (0.047)	0.731 (0.804)
	Piece-wise linear	-23304.4	0.076	24.9%	0.731
B	Linear	-13515.8	0.080	45.6% (0.000)	0.772 (0.000)
	Cubic	-12452.5	0.152	48.7% (0.960)	0.802 (0.000)
	Piece-wise linear	-11922.7	0.188	47.9%	0.812
C	Linear	-11445.8	0.137	39.8% (0.000)	0.781 (0.000)
	Cubic	-11053.4	0.167	41.1% (0.152)	0.794 (0.000)
	Piece-wise linear	-10783.4	0.187	41.6%	0.803
D	Linear	-10023.8	0.088	36.8% (0.204)	0.749 (0.000)
	Cubic	-10052.1	0.085	36.4% (0.062)	0.751 (0.000)
	Piece-wise linear	-9841.5	0.104	37.1%	0.756

P-value refers to significance of the difference from the piece-wise linear model.

The pseudo R^2 values reported in Table 7 are comparable to the pseudo R^2 values that we obtained from the study data estimation results (see Table 4). We have argued earlier that these pseudo R^2 values indicate reasonably good except for ED A. For the fitted probability, the average choice set sizes are 10.4, 5.5, 7.1, and 6.8 in the four EDs respectively, which corresponds to average fitted probabilities of 9.6%, 18.3%, 14.0%, and 14.7% by completely randomized draws. Our structural estimation framework significantly outperforms the randomized draws. Unlike the pseudo R^2 and fitted probability which are both sensitive to data structure (e.g., choice set sizes), the AUROC test provides a universally comparable metric for binary prediction performance. A five level performance accuracy classification is widely accepted in the statistics community: excellent (0.9-1.0), good (0.8-9.0), fair (0.7-0.8), poor (0.6-0.7), fail (0.5-0.6) regardless of the data and prediction sources (Tape, Pines et al. 2012). According to Table 7, the prediction accuracy of the piece-wise linear model is between good and fair, which supports the effectiveness of our framework.