Patient Sensitivity to Emergency Department Waiting Time Announcements

Eric Park, Huiyin Ouyang
Faculty of Business and Economics, The University of Hong Kong, ericpark@hku.hk, oyhy@hku.hk

Jingqi Wang
The Chinese University of Hong Kong, Shenzhen, wangjingqi@cuhk.edu.cn

Sergei Savin
The Wharton School, University of Pennsylvania, savin@wharton.upenn.edu

Siu Chung Leung, Timothy, H., Rainer
Li Ka Shing Faculty of Medicine, The University of Hong Kong, lsc676a@hku.hk, thrainer@hku.hk

Emergency department (ED) delay announcement systems are implemented in many countries. We answer three important questions pertaining to the operations and effectiveness of such systems by studying the public hospital network and ED waiting time (WT) announcement system in Hong Kong SAR: (1) How many patients are aware of (and sensitive to) the ED WT announcements? (2) How sensitive are these patients to the announced WT? (3) How can we improve the WT announcement system? We study over 1.3 million patient visits to 17 public EDs. Using a latent-class conditional logit model, we estimate the fraction of patients sensitive to the announced WT and their sensitivity. In the patient's ED choice decision, we estimate the tradeoff between the travel distance to an ED and the time waiting at the ED. We simulate the operation of the three EDs on Hong Kong Island for counterfactual policy study. We find that 2.5% of the patients are sensitive to the announced WT, and are willing to travel an additional 6.8 km to save 1 hour of waiting. Counterfactual analysis shows that the average WT and number of left without being seen patients can be reduced by 1.4% and 11.5%, respectively, by increasing the awareness (fraction of sensitive patients) to 15.8% and, simultaneously, reducing the announced WT update window to 1 hour from the current level of 3 hours. Improving awareness beyond a certain level without providing the most recent delay information may worsen system performance due to "oscillating effect" among EDs.

1. Introduction

Many communities across the world, including Australia,[1] Canada,[2] and the United States,[3] have adopted emergency departments (EDs) delay announcement systems to keep patients informed. If properly managed, such delay announcement systems can serve as a tool to manage patient flows by discouraging patients with mild conditions from going to an already overcrowded ED. In

[1] Perth, Western Australia: https://ww2.health.wa.gov.au/Reports-and-publications/Emergency-Department-activity/Data?report=ed_activity_now

[2] Vancouver, British Columbia: http://www.edwaittimes.ca/WaitTimes.aspx

[3] Dashboard for entire country where real time updates are collected from publicly available sources: https://ertrack.net/

networks of multiple EDs, the delay announcement systems may also help balance patient load across the EDs by influencing patients' ED choice.

The Hospital Authority in Hong Kong is a government body which manages all public hospitals in Hong Kong. The Hospital Authority uses a delay announcement system which collects and publicizes the waiting time data of all public EDs in Hong Kong.[4] Patients can access the waiting time information via a webpage[5] or the Hospital Authority smartphone app. For each ED, the waiting time of each patient (i.e., the time between patient's arrival and the time a patient is seen by a physician) recorded within the past three hours is collated, after which a single reference waiting time is computed and updated every 15 minutes. EDs in Hong Kong use a five-level triage system based on the patient "urgency." The Hospital Authority uses a waiting-time-based service level target for triage level 1 to 3 patients.

This paper answers three key questions pertaining to ED delay announcement system operations and its design by studying the Hong Kong public ED network. First, how many patients are currently aware of (and sensitive to) the ED waiting time announcement system in each triage category? Can we estimate the fraction of patients (penetration rate) that are sensitive to the delay announcements using patient ED visit data? Patients in critical conditions may not consider checking the waiting time announcements, especially given that they have high priority and the majority of them arrive to the ED via an ambulance. It is likely the less urgent patients who pay attention to the delay announcements and may base their decisions on the announced waiting time of different EDs. Second, how sensitive are these patients to the announced waiting time? How do patients make the trade-off between travel distance and waiting at the ED: is it worth traveling to a less congested but further away ED in favor of a closer but more congested one? Third, how should the waiting time announcement system be improved? What are the impacts of increasing the penetration rate of the delay announcement system among patients? Is it better to use more accurate or timely predictors of the waiting time in the announcement system? From the perspective of the policy maker, Hospital Authority in case of Hong Kong, answering these questions can help it evaluate the current effectiveness of the delay announcement system and may offer insights about how such delay announcement systems should be optimally designed and implemented.

We analyze data of more than 1.3 million unique public ED patient visits in Hong Kong during the full 2019 calendar year using a latent class conditional logit model. We structurally estimate the fraction of patients sensitive to the announced waiting time, their sensitivity to the announcements,

[4] In Hong Kong, EDs are referred to as accident & emergency (A&E). In the remainder of the paper, we shall use the term ED.

[5] `https://www.ha.org.hk/visitor/ha_visitor_index.asp?Content_ID=235504&Lang=ENG`

and patient characteristics that lead to higher sensitivity. Our results show that 0.8%, 3.1%, and 4.8% of triage level 3, 4, and 5 patients, respectively, (2.5% overall) are sensitive to the announced waiting time. These sensitive patients are willing to travel an additional 12.3, 4.6, and 3.5 km, respectively, to save one hour of waiting. We find that patients over 60 years old and those residing in the Kowloon district are much less likely to be sensitive to the announcement system. We follow up the empirical analysis with counterfactual analysis on policy changes where we simulate a virtual healthcare system with three EDs (to mimic the system on the Hong Kong Island) using parameters estimated from the latent class conditional logit model. We find that the average waiting time and number of left without being seen (LWBS) patients among the three EDs can be reduced by 1.1 minutes (1.4%) and 2.7 patients (11.5%) if one can increase the awareness (sensitive percentage) to 15.8%. Interestingly, our simulation results indicate that the operational system performance measures, including the average ED waiting time and LWBS patients, do not always improve as the awareness of the announcement system increases. One possible reason is the delay in the announced waiting time. When more patients make ED choices based on the announced waiting time, it may lead to waiting time oscillations, as mentioned by Dong et al. (2019). We also investigate the effect of smaller delays in making the announcements and find that using more recent information can lead to improvements in the system performance, which are more significant when more patients are sensitive to the announced waiting time.

Our contributions are threefold. First, we develop a framework to empirically identify the percentage of patients sensitive to delay announcement and their sensitivity level using ED visit data. Second, we show that less urgent patients are more likely to be sensitive to the delay announcement, but that the sensitivity decreases with the level of urgency. Third, we demonstrate that increasing the level of awareness among patients may not always be desirable and point out that making the announced waiting time information more recent is beneficial.

The results have direct practical implications for those who manage EDs, for example, the Hong Kong Hospital Authority. Our estimation results show that the current penetration level of the delay announcement system in Hong Kong is quite low, and that further promotion is needed to increase patients' awareness of the system. However, the management should closely monitor the overall awareness level because a very high percentage of sensitive patients may hurt the system performance due to the "oscillating effect" among EDs, unless the most recent delay information is provided.

2. Related Literature

The effects of delay announcements on customers' behavior have been extensively researched in the service literature. Sharing information about customers' waiting times can serve as a lever to manage customer demand flow, which, if properly used, may benefit the entire service system.

On the theoretical side, most existing studies focus on a setting with a single service provider (Naor 1969, Edelson and Hilderbrand 1975, Hassin 1986, Chen and Frank 2004, Shone et al. 2013), and there are few analytical studies of delay announcements in a network of multiple service providers in the extant literature. Singh et al. (2017) studies a setting with two service providers and find that service capacity plays a vital role in determining whether to disclose delay information to customers. More specifically, they show that the decision on whether or not to make delay announcement depends on customer sensitivity to delay, arrival rate, and the capacity of service providers.

Empirically, researchers have studied the impacts of delay announcements primarily in call center settings. Rather than assuming customers' joining or balking upon arrival, empirical studies examining customer behavior in call centers mostly focus on the impact of delay announcement on customer abandonment during waiting. Delay announcements influence customer decisions on whether to waiting or abandon, and, in turn, affect system performance. Structural estimation studies use data from call centers and are based on the analytical work by Whitt (1999), Jouini et al. (2009), Armony et al. (2009), and Jouini et al. (2011), among whom Armony et al. (2009) and Jouini et al. (2011) allow abandonment during waiting. Early empirical studies by Feigin (2006) and Mandelbaum and Zeltyn (2013) corroborate the impact of delay announcement on customer patience in queue using data from call center to demonstrate an increase in patience after receiving delay information. Based on the concept of the rational expectation equilibrium, Akşin et al. (2017) define the equilibrium in steady state as the one in which customer expectations about their waiting time matches their actual waiting time. They confirm that abandonment behavior changes as a function of customer characteristics, the delay announcement message, and the operating conditions in the call center. Their main conclusions are as follows: (i) delay information helps customers make better decisions in the sense that customers receiving long (short) delay announcement abandon more and faster (less and slower); (ii) the impact of the announcements is strongest when the state of the system is congested; and (iii) the increased granularity in delay information (the more exact delay information) leads to smoother change in customer behavior.

Considering the heterogeneity of customers, Yu et al. (2017) and Yu et al. (2018) develop frameworks characterizing strategic interactions between service providers and heterogeneous customers in call center settings. Allowing customer waiting costs to depend on delay announcement, Yu et al. (2017) empirically show that delay announcements affect both customer beliefs about the system and their waiting cost. In particular, they show that the cost-reward ratio decreases in the customer's expected waiting time before receiving service associated with the announcement. Yu et al. (2018) illustrate that delay announcements have two roles. Besides informing customers about the state of the system, delay announcements also have the potential to elicit information about customer types, such as customer valuation of service and patience, based on their response to the announcements. The service provider can use delay announcements as a tool to manage customer expectations and to prioritize customers. To detect the behavioral implications of delay announcements on customer decision making, Yu et al. (2021) and Webb et al. (2019) employ data from call centers to explore the loss aversion based on the reference point perceived by customers through announced delay. Yu et al. (2021) find that customers indeed exhibit loss aversion, independent of the correctness of the delay information provided. In the case of inaccurate delay information, customers use the observed average delay as a reference point instead. For those customers who do not abandon, Webb et al. (2019) find that additional time spent on waiting beyond the reference point leads to longer service time.

Delay announcement in the health care setting, particularly in EDs, are different from that in call centers in that the delay announcement in one ED affects nearby EDs as well, creating a linkage between multiple service systems. In addition, patients in EDs are typically classified into different priority groups and can be highly heterogeneous. Dong et al. (2019) conduct a pioneering work on this track of research. They find that hospital delay announcements indeed influence patient choices, creating load balancing and thus synchronization in the network. The authors also conduct a numerical study using parameters calibrated by real data to examine how patient sensitivity to delay, the load of the system, and the heterogeneity among hospitals change the effects of delay announcements on network synchronization. They demonstrate the importance of timely and accurate delay announcements in improving system performance and show that using historical averages as delay estimators can cause “oscillations” in the state of the system and result in higher waiting time. These observations prompted a search for better ED waiting time predictors, resulting in new predictors such as Q-Lasso (Ang et al. 2016) and weighted average of static and dynamic announcement (Bassamboo and Ibrahim 2021). However, although there already exist refined real-time delay estimators (see, for example, Ibrahim and Whitt (2009), Arora et al. (2020), and Ibrahim and Whitt (2011)), these are rarely used in practice (Dong et al. 2019). Instead, hospitals usually publish historic average waiting times (such as four-hour moving average), potentially creating oscillations in the system (Dong et al. 2019, Pender et al. 2016). With data at a more granular level on each individual patient’s ED choice, we are able to empirically estimate the proportion of patients who are sensitive to the delay announcement information.

Instead of having patients making the ED choice, this decision can be made in a centralized manner by a coordinating authority. Fatma and Ramamohan (2021) propose a patient diversion mechanism based on real-time delay predictions within a healthcare facility network. The simulation results show that the implementation of the diversion framework reduces congestion across the network and facilitates synchronization. However, the performance of the diversion mechanism deteriorates as the delay predictor becomes less accurate.

This paper differs from the extant literature by empirically estimating the percentage of patients sensitive to delay announcement and their sensitivity level in a interrelated network setting using ED visit data. We also offer prescriptive suggestions on the delay announcement system via simulation.

3. Data and Study Setting

Table 1 Summary Statistics: Raw Data

	Triage 1	Triage 2	Triage 3	Triage 4	Triage 5	Missing	Total
Average waiting time (min)	0.0	7.3	25.9	116.8	128.6	100.5	79.6
95-th percentile waiting time (min)	0	15	70	341	362	360	281
Average patient age (years)	68.7	63.0	58.3	44.5	42.2	41.8	50.1
Fraction of ambulance arrivals	87.6%	67.2%	48.0%	15.1%	1.4%	22.2%	28.5%
Fraction of female patient visits	43.2%	47.3%	50.8%	52.9%	53.5%	48.9%	51.9%
Total number of visits	22,032	52,687	742,477	1,186,135	72,230	4,802	2,080,363
Fraction of the total number of visits	1.1%	2.5%	35.7%	57.0%	3.5%	0.2%	

Notes. Waiting time is the actual time patients waited at the ED.

Public hospital EDs under the Hong Kong Hospital Authority use a triage system which classifies patients into five triage categories based on their urgency: level 1 (critical), level 2 (emergency), level 3 (urgent), level 4 (semi-urgent) and level 5 (non-urgent). For triage levels 1, 2, and 3 patients, the Hospital Authority has set waiting time based service targets: triage level 1 patient will be treated immediately; 95% of triage level 2 patients will be treated within 15 minutes; 90% of level 3 patients will be treated within 30 minutes.[6] Meanwhile, due to the limited alternative options, majority of the public in Hong Kong rely on the public health system for emergent healthcare. In a city of 7.5 million population, over 1.25 million patients visited a public ED amassing to a total of over 2 million visits in 2019. As a result, public EDs in Hong Kong have been suffering from overcrowding in general. The Hospital Authority even increased the ED visit fee from HK$100 to HK$180 in 2017 for the first time since 2003, encouraging appropriate use of public healthcare services.

The study data consists of over 2 million patient visits to the 18 public hospital EDs across Hong Kong during the 2019 calendar year. There are approximately 1.25 million unique patients according to the patient identifier. Table 1 summarizes the overall patient visit data at the individual visit level. The reported waiting time is computed as the difference between time of registration

[6] `https://www.ha.org.hk/visitor/ha_serviceguide_details.asp?Content_ID=10051&IndexPage=200066&Lang=ENG&Ver=HTML`

at the ED and the time to bed. Noticeably, triage level 1 patients do not wait for a bed upon arriving to the ED while over 87% of them arrive by an ambulance which has a pre-determined routing policy of delivering the patient to the nearest ED regardless of the congestion/waiting time situation. While the mode of arrival for the non-ambulance patients is unclear, considering their critical condition, they are also likely to be brought in to the closest ED irrespective of delays. Triage level 2 patients also wait for only a fairly short period of time, on average, with a large portion of patients arriving by ambulance.

In our study we limit our analysis to triage levels 3, 4, and 5 patients who did not arrive by ambulance. This focus is dictated by the fact that rest of the patients were almost always brought to the nearest ED rather than exercised their choice of the ED based on the delay information. Moreover, we focus on the patients that attended one of the 5 nearest EDs to their residential district. There are 138 unique residential districts, and we discuss the use of this information in more detail in Section 4.2. Out of the 18 public hospitals in Hong Kong, St. John Hospital is located on a remote island with a population of less than 25,000 and is not connected by road to any other public hospital in the city. We exclude all patient visits to this ED due to the lack of alternative hospital choice for the patients. Table 2 summarizes the study data by triage level and ED by order of closeness to the patient. The first and second block in this Table report the average statistics for the announced waiting time of each ED observed by the patient and the travel distance to them. For instance, across all three triage levels on average, the second-nearest ED is 5.1 km (7.8 km - 2.7 km) farther than the nearest ED from a patient while the waiting time were identical at the two EDs at 2.7 hours. Overall, 75.6% of the patients attended the nearest ED while 14.5% attended the second-nearest ED. When patients attended the second- or third-nearest ED, they had a shorter announced waiting time compared to the average, 2.6 vs 2.7 hrs for second-nearest and 2.6 vs 2.8 hrs for third-nearest ED (fourth vs second block). From the fifth block, we see that patients who attended the second- and third-nearest ED on average travelled 5.7 and 8.1 km respectively, much less than the average distance to the respectively ranked ED of 7.8 and 9.9 km from the second block. This shows that patients who attended the non-nearest ED attended a relatively closer nearby ED that was also slightly less congested which provides evidence of patients taking distance and announced waiting time into account in their ED choice decision. In Table 2 we also report the percentage of patients in each category who used the Hong Kong's Cross-Harbor Tunnel to get to the ED, thus overcoming a substantial additional obstacle in getting to their chosen care location.

The waiting time announce system is hosted on the Hospital Authority website (`https://www.ha.org.hk/visitor/ha_visitor_index.asp?Content_ID=235504&Lang=ENG`, accessed on February 23. 2022) which can also be accessed via a link from the Hospital Authority smartphone

Table 2 Summary Statistics: Study Data

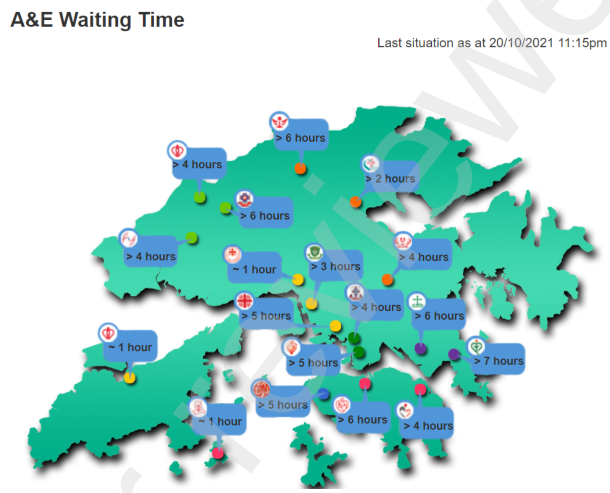
	Triage 3	Triage 4	Triage 5	Total
Average announced ED waiting time (hr)				
Nearest ED	2.9	2.6	2.4	2.7
2nd nearest ED	2.8	2.7	2.7	2.7
3rd nearest ED	2.8	2.8	2.8	2.8
4th nearest ED	3.0	3.0	3.1	3.0
5th nearest ED	2.7	2.7	2.6	2.7
Average travel distance to ED (km)				
Nearest ED	2.8	2.6	2.4	2.7
2nd nearest ED	7.2	8.1	7.0	7.8
3rd nearest ED	9.3	10.1	9.0	9.9
4th nearest ED	12.4	14.1	14.0	13.6
5th nearest ED	14.1	15.9	16.4	15.4
Attended ED				
Nearest ED	69.2%	77.9%	77.1%	75.6%
2nd nearest ED	18.4%	13.1%	14.0%	14.5%
3rd nearest ED	7.1%	5.1%	5.5%	5.6%
4th nearest ED	3.2%	2.2%	2.1%	2.4%
5th nearest ED	2.2%	1.7%	1.4%	1.8%
Average announced waiting time of attended ED (hr)				
Overall	2.9	2.5	2.4	2.6
Nearest ED	3.0	2.5	2.4	2.6
2nd nearest ED	2.9	2.5	2.3	2.6
3rd nearest ED	2.8	2.5	2.3	2.6
4th nearest ED	3.2	2.9	2.8	3.0
5th nearest ED	3.0	2.6	2.4	2.7
Average distance to attended ED (km)				
Overall	4.0	3.6	3.4	3.7
Nearest ED	2.7	2.5	2.2	2.5
2nd nearest ED	5.2	6.0	6.0	5.7
3rd nearest ED	8.0	8.2	7.7	8.1
4th nearest ED	10.7	11.3	10.4	11.0
5th nearest ED	14.3	13.5	12.1	13.7
Average patient age (years)	49.0	42.1	42.0	43.9
Fraction of female patient visits	51.5%	53.5%	54.0%	53.0%
Fraction of cross-harbor visits	0.3%	0.2%	0.2%	0.2%
Number of visits	351,895	912,436	65,778	1,330,109
Fraction of the total number of visits	26.5%	68.6%	4.9%	

Notes. Attended ED is the percentage of attended ED by ranking in closeness to patient location. Waiting time is the announcement observed by the patients before attending an ED.

Map view Last situation as at 20/10/2021 11:15pm

Hospital	Reference waiting time
Hong Kong Island	
Pamela Youde Nethersole Eastern Hospital	Over 4 hours
Queen Mary Hospital	Over 5 hours
Ruttonjee Hospital	Over 6 hours
Kowloon	
Caritas Medical Centre	Over 5 hours
Kwong Wah Hospital	Over 4 hours
Queen Elizabeth Hospital	Over 5 hours
United Christian Hospital	Over 6 hours
New Territories	
Alice Ho Miu Ling Nethersole Hospital	Over 2 hours
North District Hospital	Over 6 hours
North Lantau Hospital	Around 1 hour
Pok Oi Hospital	Over 6 hours
Prince of Wales Hospital	Over 4 hours
Princess Margaret Hospital	Over 3 hours
St John Hospital	Around 1 hour
Tin Shui Wai Hospital	Over 4 hours
Tseung Kwan O Hospital	Over 7 hours
Tuen Mun Hospital	Over 4 hours
Yan Chai Hospital	Around 1 hour

(a) Table



(b) Map

Figure 1 Hong Kong Hospital Authority Accident & Emergency Waiting Time Announcement System

app. The system reports the 95-th percentile of waiting time of all ED patients who entered care over the last 3 hours (or last 6 hours for the period between 3AM and 9AM). The announcements are updated every 15 minutes with the time stamp of the latest update noted. The announced waiting times are referred to as "Reference waiting time" only, while the announcement system states them as not the current estimated waiting time and does not provide how they are calculated to the public. The granularity of the announced waiting time has 9 levels: around 1 hour, over 1 hour, over 2 hours, ..., over 8 hours. The announcement system uses both the table format (Figure 1a), with the name and announced waiting time of each hospital in the three administrative regions, and the map (Figure 1b) with the announced waiting time and location of the hospitals shown.

4. Model of Attending Hospital Choice by ED Patients

In this Section we describe the general modeling framework we use, the latent class conditional logit, followed by the details of our model specification.

4.1. Latent class conditional logit framework

In the conditional logit framework, a patient evaluates the utility gained by going to each of the five nearest EDs at each choice incident. The patient then chooses the ED with the highest utility as the destination. In the non-latent class standard conditional logit model, the utility of patient n attending ED j at choice incident t, U_{njt}, can be expressed as

$$U_{njt} = V_{njt} + \epsilon_{njt} = \boldsymbol{\beta}\mathbf{x}_{njt} + \epsilon_{njt}, \tag{1}$$

where V_{njt} is patient n's own valuation of attending ED j at choice incident t. This valuation depends on several factors, $\mathbf{x}_{njt}$, that we discuss in detail in subsection 4.2. The error term, ϵ_{njt} which represents the external factors that affect the utility, is i.i.d. type I extreme value distributed. The column vector of preference (or taste) parameters, β, is the same for all decision makers—patients.

The latent class conditional logit model relaxes the last condition and allows decision maker heterogeneity in some, but not necessarily all, of their taste parameters. It assumes that there are C distinct types of decision makers, where type (or class) $c = 1, ..., C$ patients share the same taste parameter vector $\beta_{\mathbf{c}}$. In the context of our study, this represents groups of patients that have different sensitivity to announced waiting times. Conditional on patient n belonging to class c, the probability of observing her sequence of T choices can be computed as a product of logits

$$P_n\left(\boldsymbol{\beta}_c\right) = \prod_{t=1}^{T}\prod_{j=1}^{J}\left(\frac{\exp(\boldsymbol{\beta}_c\mathbf{x}_{njt})}{\sum_{k=1}^{J}\exp(\boldsymbol{\beta}_c\mathbf{x}_{nkt})}\right)^{y_{njt}}, \tag{2}$$

where y_{njt} is the binary outcome variable which equals 1 if patient n attended ED j or 0 if she did not attend ED j at choice incident t. Since a patient's class membership is unobserved, the unconditional probability of observing patient n's sequence of choices needs to be computed as the average of equation (2) weighted by the probability of patient n belonging to class c, π_{nc}. This weight can be computed as a fractional multinomial logit

$$\pi_{nc}(\boldsymbol{\Theta}) = \frac{\exp(\boldsymbol{\theta}_c\mathbf{z}_n)}{1 + \sum_{l=1}^{C-1}\exp(\boldsymbol{\theta}_l\mathbf{z}_n)}, \tag{3}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_{C-1})$ is the collection of parameters for class membership and $\mathbf{z}_n$ is the vector of explanatory variables pertaining to class membership selection. Note that $\boldsymbol{\theta}_C$ is normalized to $\mathbf{0}$ for identification.

The log-likelihood of observing the entire sequence of choices is then the sum of each patient's log-unconditional likelihood:

$$\ln L(\mathbf{B}, \boldsymbol{\Theta}) = \sum_{n=1}^{N}\ln\sum_{c=1}^{C}\pi_{nc}(\boldsymbol{\Theta})P_n(\boldsymbol{\beta}_c), \tag{4}$$

where $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_C)$ is the collection of taste parameters for all C classes. We use the lclogit2 command in STATA to obtain the maximum likelihood estimates of equation (4) (Yoo 2020).

4.2. Model specifications: patient's ED choice and membership classification

There are two choice models in the latent class conditional logit framework: patient's choice of which ED to attend (equation 2) and the classification of the patient's underlying taste class (equation 3). In this subsection, we discuss the specification for both models.

When a patient is seeking emergency medical attention, there are several factors that can affect the decision of which ED to attend. Generally, ED patients seek the shortest possible time to see a physician. This time can be split into two components: travel time to the ED from the patient's location and the waiting time to be seen by a doctor after arriving at the ED. There may be exceptions where a patient with a certain medical condition may prefer a specific hospital over other EDs that may have shorter time to the first consultation with a physician. With the help of modern technology, it is easy for patients to get a fairly accurate estimate of the travel time to an ED, whereas accurately estimating the waiting time at a specific ED is likely to be more difficult given the limited information available to the patients. However, for patients who are aware of the ED waiting time announcement system, the announced waiting time information can facilitate patient's analysis of the trade-off between travel time and waiting time at the ED. While we specify our choice model to capture this trade-off and the relative patient sensitivity to the two inputs, the estimation of the travel time to potential EDs the patient was facing at the time of the choice incident is complicated by the lack of information about the precise location from where she came from to the ED or about the traffic conditions at that time.

We address this issue by utilizing the patient's residential district information. We use the classification of 138 unique residential districts to approximate the patient's location. A key assumption we make is that patients went to the ED from their residential district whether it was exactly from home or from the neighborhood. Specifically, we use Google Maps to define an approximate center point for each district, and to identify the five nearest EDs (by driving distance) for each district. For any patient visit where the patient attended an ED outside of the five nearest ones, we assume that the patient was not at or near home at the time of deciding to attend an ED. These observations are dropped from the analysis as we cannot identify either the potential EDs the patient may have considered or the distance to them. Since we cannot estimate the expected travel time at the time of the incident due to the lack of traffic condition information, we instead use the shortest driving distance from the center point of the patient's residential district to each ED as the travel the patient had to consider.

While we expect most patients to be sensitive to the time to first be seen by a doctor, not all patients will experience the same waiting time after they arrive at the ED. Based on the triage system, urgent patients will be given priority while less urgent patients will have to wait longer. While patients may not know their exact triage class before being assessed at the ED, they may have a general sense of urgency of their case. This can affect their expectation of waiting time once they arrive at the ED. Therefore, we analyze the patients by their evaluated triage class. For the main result, we run the latent class conditional logit model for visits from each triage class separately. This allows us to compare the sensitivity to waiting time and travel across patients with different levels of urgency. We show that our results are robust under the combined analysis for patients across multiple triage classes (see subsection 6.2).

While the waiting time announcement system is publicly available, it is unclear how many patients are aware of the system and utilize it before visiting an ED. Hence, patients may exhibit heterogeneity in their sensitivity to the waiting time. Only patients who access the system to learn the latest waiting time at EDs can be sensitive to the waiting time compared to those who do not have such information at all. We adopt the latent class feature to identify the underlying type of the patients utilizing two patient characteristics available from data: age and geographic location. We expect younger patients who are generally more familiar with information technology to be more aware of the waiting time announcement system and exhibit preference for shorter waiting time, whereas older patients to be less aware of the announcement system and primarily be driven by travel distance in their ED choice decision. Hong Kong is divided into three administrative territories: Hong Kong Island, Kowloon, and the New Territories. The level of awareness of the waiting time announcement system may vary by territory and if so, identifying which territory needs further advertisement and education can be a vital information for policy makers. In our main analysis, we assume two classes of patients, $C = 2$, where the first class represents patients aware of and sensitive to the announced waiting time while the second class represent those that are neither aware of the system nor sensitive to the announced delay information. In particular, the latter class includes patients who do access the announcement system but do not take the delay information into account. As a robustness check (subsection 6.1), we estimate models with $C = 3$ classes, where one class is the "unaware" patients and where "aware" patients are divided into two classes according to the level of sensitivity to the announced waiting times. We show that the results of the three-class analysis are widely consistent with those for the two-class model, suggesting that the patient population can be divided dichotomously into sensitive and insensitive groups when it comes to announced waiting time sensitivity.

The utility of patient n belonging to class c in the patient classification model (equation 3) can be expressed as

$$\boldsymbol{\theta}_c \mathbf{z}_n = \boldsymbol{\theta}_1^{c\prime} \overline{Age}_n + \boldsymbol{\theta}_2^{c\prime} \overline{Territory}_n. \tag{5}$$

Here $\overline{Age}_n$ is the vector of age-group dummy variables for patient n categorized by 10-year age intervals, $(Age10_n, Age20_n, Age30_n, ..., Age70_n)$, where $Age10_n$ equals 1 if patient n's age is between

10 and 19, and $Age70_n$ equals 1 if patient n is older than 69. $Age0_n$ is dropped for identification purpose. $\boldsymbol{\theta}_1^c$ is the vector of coefficients for patient age groups belonging to type c class. $\overline{Territory}_n$ is the vector of territory dummy variables, $(Hong\ Kong\ Island_n, New\ Territories_n)$. We drop patient gender from the model as it has insignificant impact on class membership while adding computational burden.

In a model with single triage level, patient n who belongs to class c has a valuation of attending ED j and choice incident t that equals

$$\boldsymbol{\beta}_c \mathbf{x}_{njt} = \beta_1^c Wait_j(t) + \beta_2 Dis_{nj}(t) + \beta_3 X_j + \beta_4 CrossHarbour_{nj}(t). \tag{6}$$

$Wait_j(t)$ is the announced waiting time of ED j at the time of choice incident t. We map the waiting time of each ED that was announced two updating cycles before the patient registered at the ED of choice. For instance, if a patient registered at an ED at 13:47, the announcement at 13:15 is mapped as the waiting time observed by the patient for the focal incident. This is done to approximate the moment when the patient would have collected the waiting time information from the announcement system. $Dis_{nj}(t)$ denotes the distance from patient n to ED j at choice incident t, since some patients may have moved to a new residential district within the study period. We control for time-invariant fixed effect of ED j, $X_j = \{individual\ ED\ fixed\ effects\}$, and for the fact that patient n had to cross the harbor to attend ED j at choice incident t, $CrossHarbour_{nj}(t)$. β_1^c captures class c decision maker's sensitivity to the announced waiting time. In the estimation, we impose the insensitivity constraint on class 2, $\beta_1^2 = 0$. β_2 captures the class-independent travel distance sensitivity. The ratio between β_1^1 and β_2 represents the number of additional kilometers that class 1 patients, the waiting-time-sensitive group, are willing to travel to save one hour of waiting. The controls β_3 and β_4 are uniform across different classes.

Within our modeling framework, we assume that once a patient decides that she needs emergency medical attention, she will attend an ED and will not consider the outside option of not attending at all regardless of how long the waiting times are at the EDs.

5. Results

Table 3 reports the estimation results of the ED choice model (equation 6) on the left panel and membership classification model (equation 5) on the right panel. First, we find that there are two distinct classes of waiting time sensitivity within each triage level. For all three triage levels, the presence of waiting-time-sensitivity (class 1) patients is statistically significant at the 0.1% level. However, only a small fraction of patients are sensitive to the announced waiting times, 2.5% among triage level 3, 4, and 5 visits - this value represents class 1 membership shares weighted by visits per triage level. Further, we find that urgent patients have a smaller share of waiting-time-sensitive

Table 3 ED Patient Sensitivity to Announced Waiting Time and Travel Distance

	Triage 3	Triage 4	Triage 5		Triage 3	Triage 4	Triage 5
ED choice				Class 1 membership			
Class 1				Age10	0.291	-0.339***	-0.295
Waiting time (hr) (β_1^1)	-5.124***	-1.814***	-1.378***		(0.245)	(0.094)	(0.352)
	(0.213)	(0.034)	(0.079)	Age20	1.035***	0.602***	0.488
Membership shares	0.8%	3.1%	4.8%		(0.159)	(0.059)	(0.276)
Class 2				Age30	0.806***	0.377***	-0.127
Waiting time (hr)	0	0	0		(0.156)	(0.060)	(0.291)
	-	-	-	Age40	0.655***	-0.101	-0.490
Membership shares	99.2%	96.9%	95.2%		(0.162)	(0.067)	(0.299)
Travel distance (km) (β_2)	-0.415***	-0.396***	-0.399***	Age50	0.213	-0.399***	-0.780*
	(0.001)	(0.001)	(0.002)		(0.169)	(0.070)	(0.321)
N	1,759,475	4,562,180	328,890	Age60	-0.326	-0.963***	-1.889***
Visits	351,895	912,436	65,778		(0.187)	(0.089)	(0.489)
Patients	286,794	648,060	53,037	Age70	-1.164***	-2.391***	
Travel distance/waiting time (km/hr) ratio (β_1^1/β_2)					(0.220)	(0.194)	
Class 1	12.3	4.6	3.5	HKI	1.072***	0.904***	1.123**
					(0.172)	(0.075)	(0.358)
Class 2	0.0	0.0	0.0	NT	1.605***	1.302***	2.047***
					(0.140)	(0.056)	(0.251)

Notes. Standard errors in parentheses. ED fixed effects and cross harbor controls are reported in Table A1.

Triage level 5 has up to Age60 in the class membership model due to lack of elderly patients.

HKI: Hong Kong Island, NT: New Territories

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

patients, 0.8% for triage level 3 compared to 3.1% and 4.8%, respectively, for triage level 4 and 5. Meanwhile, the waiting time sensitivity increases as patients become more urgent. The travel-distance-to-waiting-time ratio, β_1^1/β_2, increases from 3.5 km/hr to 4.6 and 12.3 for "sensitive" triage level 5, 4, and 3 patients, respectively. Given that the nearest and second-nearest ED to patients are 5.1 km away on average (Table 2), the ratios suggest that the sensitivity to waiting time is large enough for the sensitive patients to divert away from the nearest ED to the next nearest EDs to trade-off additional travel with saving in waiting time once they reach an ED.

These findings suggest that patients that are not very urgent may be aware of the long waits once they arrive at the ED given their lower priority within the triage system, and more patients of that type check the waiting time announcement system to find EDs with potentially shorter waiting. In contrast, the urgent patients either have smaller "bandwidth" to check the time given the urgency of their condition or are aware of their likely high priority once they arrive at the ED. As a result, fewer of these patients are aware of/use the waiting time announcement system. In addition, the sensitive urgent patients are so sensitive to waiting that they are willing to travel even further as compared to their less urgent counterparts.

For the class membership, as expected, we find that younger patients, especially those in their 20s and 30s, are the most aware of the waiting time announcement system and are sensitive to waiting times. Among triage level 4 and 5 patients who compose over half of the patient body and experience the longest waiting (Table 1), older patients (those over 60) have a low awareness to the waiting time announcement system. For the triage level 5 model, age groups are considered only up Age60 as there are fewer patients over 70 and they are grouped together with those in their 60s. Geographically, patients residing in Kowloon have a lower awareness compared to those on Hong Kong Island or in New Territories. Hence, policy makers should focus on the older population, especially in Kowloon, in raising awareness of the ED waiting time announcement system.

The low awareness of the waiting time announcement system among the Hong Kong public suggests that the system may not be achieving its full potential. In Section 7, we perform counterfactual analyses and study opportunities to increase the effectiveness of the announcement system.

6. Robustness Analysis

In this Section, we perform various analyses to support the model choice and the robustness of the main results.

6.1. Patient heterogeneity

We find strong evidence of patient heterogeneity in their sensitivity to announced waiting time in ED choice decisions. We estimate a mixed logit model for each triage class visits where we assume patients' sensitivity to waiting time (β_1 in equation 6 but without the membership class) is a random coefficient. In terms of modeling patient heterogeneity, the spectrum of heterogeneity can be represented by the number of membership classes; one class represents a homogenous patient body with the least heterogeneity and as we increase the number of classes we allow more heterogeneity within the model. Table 4) reports the results of the mixed logit model which can be considered as the model allowing the highest level of patient heterogeneity with individual patients in their own class. We find strong evidence of heterogeneity in waiting time sensitivity in triage level 4 and 5 patients, which is in line with the substantial fraction of waiting time sensitive class patients in the main results, 3.1% and 4.8% (Table 3). Meanwhile, heterogeneity is not significant in triage level 3 patients, which is also in line with the main results with only a small fraction of sensitive class patients, 0.8%.

Table 4 Random Coefficient Mixed Logit: Patient Heterogeneity in Sensitivity to Announced Waiting Time and Travel Distance

	Tri 3	Tri 4	Tri 5
	Coefficient		
Waiting time (hr)	0.000	-0.031***	-0.038***
	(0.001)	(0.001)	(0.004)
Travel distance (km)	-0.403***	-0.385***	-0.390***
	(0.001)	(0.001)	(0.002)
	Std. Dev.		
Waiting time (hr)	0.024	0.126***	0.139***
	(0.014)	(0.002)	(0.009)
N	1,759,475	4,562,180	328,890
Visits	351,895	912,436	65,778
Patients	286,794	648,060	53,037

Note. Standard errors in parentheses. ED fixed effects and cross harbor controls are not shown.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

To further explore the level of patient heterogeneity, we estimate the latent class conditional logit model with three classes for each triage level (equation 4 with $C = 3$). Table 5 shows the result with three membership classes per triage level, the next level of patient heterogeneity in comparison to the main model (Table 3). While we find a third class, class 1 in Table 5, with significantly higher sensitivity to waiting time compared to the moderate sensitivity class, class 2, the proportion of class 1 is quite small, less than 0.6% for each triage level. Hence, we simplify the heterogeneity modelling by presenting the most parsimonious heterogeneity model with 2 classes in Table 3 as the main result.

6.2. Membership classification

For the main result, we estimate the choice and membership classification model for each triage class patients separately. This allows a flexible membership classification across triage levels. For instance, a patient that is classified into the insensitive group when she is a non-urgent patient, level 5 in the Hong Kong A&E triage system, could be classified as a sensitive patient when she is a semi-urgent patient, triage level 4. This allows urgency-dependent membership classification. As a robustness check, we analyze triage level 3, 4, and 5 visits all together assuming each patient belongs to a single class regardless of the triage level. This triage level-independent membership classification model assumes that each patient belongs to a single class regardless of their urgency but may have different sensitivity depending on their urgency. Table 6 reports the estimation results of the pooled latent class conditional logit model.

We find that 3.4% of the patients are sensitive to the announced waiting time in the urgency-independent membership model.

Table 5: Latent Class Conditional Logit: Three Membership Classes

	ED Choice			Class membership						
	Triage 3	Triage 4	Triage 5		Triage 3		Triage 4		Triage 5	
Class 1					Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Waiting time (hr)	-9.921***	-7.844***	-9.706***	Age10	1.237	0.001	0.624	-0.441***	0.355	-0.280
	(0.542)	(0.234)	(1.055)		(0.668)	(0.258)	(0.320)	(0.103)	(0.844)	(0.346)
Membership shares	0.4%	0.5%	0.6%	Age20	2.518***	0.131	2.082***	0.299***	1.378*	0.549*
Class 2					(0.461)	(0.193)	(0.261)	(0.068)	(0.695)	(0.272)
Waiting time (hr)	-2.107***	-1.329***	-1.217***	Age30	2.387***	-0.293	2.168***	-0.059	1.465*	-0.050
	(0.097)	(0.025)	(0.067)		(0.454)	(0.199)	(0.258)	(0.073)	(0.691)	(0.286)
Membership shares	0.8%	3.1%	5.2%	Age40	2.119***	-0.305	1.457***	-0.393***	1.642*	-0.387
Class 3					(0.462)	(0.203)	(0.267)	(0.078)	(0.674)	(0.291)
Waiting time (hr)	0	0	0	Age50	1.613***	-0.558**	1.268***	-0.738***	1.296	-0.666*
	-	-	-		(0.470)	(0.199)	(0.269)	(0.086)	(0.686)	(0.309)
Membership shares	98.7%	96.4%	94.2%	Age60	1.292**	-1.600***	0.803**	-1.373***	0.605	-1.751***
Travel distance (km)	-0.418***	-0.399***	-0.409***		(0.474)	(0.320)	(0.281)	(0.118)	(0.723)	(0.448)
	(0.001)	(0.001)	(0.003)	Age70	0.383	-2.028***	-0.583	-2.761***		
Travel distance/waiting time (km/hr) ratio					(0.499)	(0.297)	(0.388)	(0.246)		
Class 1	23.8	19.7	23.7	HKI	0.983***	0.796	1.289***	0.568***	0.507	1.205***
Class 2	5.0	3.3	3.0		(0.215)	(0.500)	(0.194)	(0.108)	(0.616)	(0.328)
Class 3	0.0	0.0	0.0	NT	0.837***	2.424***	1.478***	1.179***	1.885***	2.072***
					(0.199)	(0.355)	(0.179)	(0.071)	(0.465)	(0.238)

Notes. Standard errors in parentheses. ED fixed effects and cross harbor controls are not shown in the ED choice model. Number of observations for each triage level model are identical to those used in Table 3.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6 Pooled Latent Class Conditional Logit: Triage Level-Independent Membership Classification

ED choice	Class 1	Class 2	Class 1 membership	
Waiting time (hr)	-1.782***	0	Age10	-0.111
	(0.047)	-		(0.073)
Waiting time X Triage 3	1.037***	0	Age20	0.623***
	(0.062)	-		(0.050)
Waiting time X Triage 4	0.039	0	Age30	0.354***
	(0.051)	-		(0.051)
Travel distance (km)	-0.397***		Age40	-0.038
	(0.001)			(0.056)
Waiting time + Waiting time X Triage 3	-0.745***	0	Age50	-0.325***
	(0.044)	-		(0.058)
Waiting time + Waiting time X Triage 4	-1.743***	0	Age60	-0.927***
	(0.026)	-		(0.074)
Membership shares	3.4%	96.6%	Age70	-2.100***
Travel distance/waiting time (km/hr) ratio				(0.131)
Triage 3	1.9	0.0	HKI	0.685***
Triage 4	4.4	0.0		(0.062)
Triage 5	4.5	0.0	NT	1.147***
N	6,650,545			(0.042)
Visits	1,330,109			
Patients	883,261			

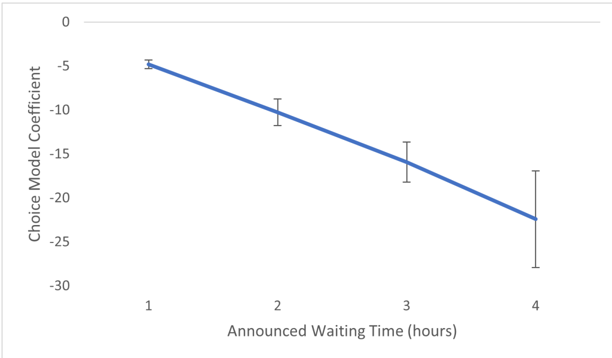
Notes. Standard errors in parentheses. ED fixed effects and cross harbor controls are not shown in the ED choice model.

HKI: Hong Kong Island, NT: New Territories

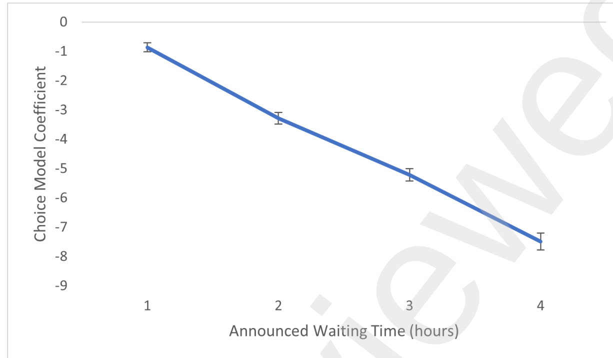
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6.3. Potential nonlinearity in waiting time sensitivity

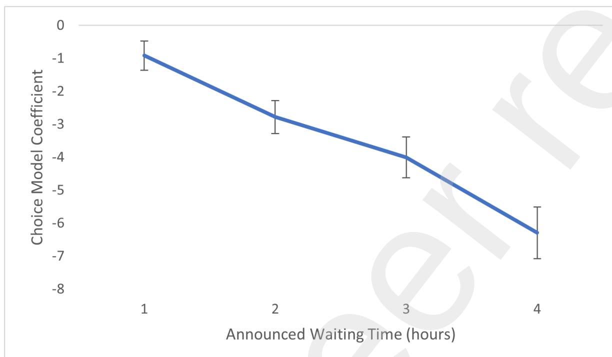
While the main model assumes linear waiting time sensitivity, it has been known that patients do not always exhibit constant marginal waiting cost (Ding et al. 2019). We explore potential nonlinearity in waiting time sensitivity in the ED choice behavior by estimating a nonlinear waiting time specification of equation 6. We capture the nonlinearity through step-wise functions which allow flexible marginal waiting time sensitivity. Waiting time is divided roughly into quartiles excluding the 0 hour waiting, where the highest quarter includes 4 to 8 hours, and the next three quarters are captured by 3, 2, and 1 hour alone respectively. The results are presented in Figure 2. For all three triage levels, we find that the waiting time sensitive patients, class 1, exhibit a somewhat linear waiting time sensitivity. Triage 4 shows the clearest linear pattern, while triage 5 shows some steepening after 3 hours. We conclude that we do not find significant evidence of nonlinearity in patients' sensitivity to announced waiting time information.



(a) Triage level 3, Class 1



(b) Triage level 4, Class 1



(c) Triage level 5, Class 1

Figure 2 Linearity in Waiting Time Sensitivity by Triage Level

6.4. Weekend and peak period effects

We explore any potential effect of patients visiting an ED during the weekend and/or in peak daytime period, defined as from 8AM to 6PM. We find that urgent and semi-urgent (triage level 3 and 4) patients are more sensitive to the announced waiting time during peak periods than non-peak periods while the opposite holds for non-urgent (triage level 5) patients. There is no significant difference between weekend and weekdays. For instance, urgent patients in the sensitive class (first column in Table 7) are willing to travel an additional 15.1 km to save 1 hour in waiting at the ED on weekdays during peak hours which is twice than during non-peak hours, 7.4 km. We can possibly attribute this to urgent patients avoiding farther travel during high traffic congestion periods.

6.5. Ambulance routing policy

We estimate a simple conditional logit model without latent classes, which assumes a single class of patient, for ambulance and walk-in patients separately for each triage level (Table 8). Ambulance patients are not sensitive to the announced waiting time as expected and prefer closer ED more than walk-in patients. Hence, our results provide evidence of the ambulance routing policy in effect which describes patients to be delivered to the nearest ED regardless of road and ED situation. The

Table 7 Latent Class Conditional Logit: Weekend and Peak Period Effect

	Triage 3		Triage 4		Triage 5	
	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Waiting time (hr)	-3.065***	0	-1.527***	0	-1.572***	0
	(0.179)	-	(0.039)	-	(0.143)	-
Waiting time X Weekend	0.523	0	0.045	0	-0.521	0
	(0.289)	-	(0.061)	-	(0.367)	-
Waiting time X Peak	-3.197***	0	-0.668***	0	0.352*	0
	(0.394)	-	(0.063)	-	(0.156)	-
Waiting time X Weekend X Peak	1.399*	0	-0.080	0	0.123	0
	(0.592)	-	(0.105)	-	(0.401)	-
Travel distance (km)	-0.416***		-0.396***		-0.399***	
	(0.001)		(0.001)		(0.002)	
Waiting time (weekend non-peak)	-2.541***	0	-1.482***	0	-2.093***	0
	(0.234)	-	(0.053)	-	(0.359)	-
Waiting time (weekday peak)	-6.261***	0	-2.195***	0	-1.220***	0
	(0.379)	-	(0.059)	-	(0.088)	-
Waiting time (weekend peak)	-4.339***	0	-2.230***	0	-1.618***	0
	(0.426)	-	(0.073)	-	(0.172)	-
Membership shares	1.0%	99.0%	3.1%	96.9%	4.5%	95.5%
N	1,759,475		4,562,180		328,890	
Visits	351,895		912,436		65,778	
Patients	286,794		648,060		53,037	
Travel distance/waiting time (km/hr) ratio						
Weekday non-peak	7.4	0.0	3.9	0.0	3.9	0.0
Weekend non-peak	6.1	0.0	3.7	0.0	5.2	0.0
Weekday peak	15.1	0.0	5.5	0.0	3.1	0.0
Weekend peak	10.4	0.0	5.6	0.0	4.1	0.0

Notes. Standard errors in parentheses. Only the ED choice model results without the ED fixed effects and cross harbor controls are shown. The omitted class membership model is robust to results in Table 3.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

high pseudo-R^2 statistics across all models suggest that the conditional logit framework successfully represents the ED choice decision process for all type of patients.

6.6. Frequent visitors

We also study the patient characteristics of frequent visitors, who visit the ED frequently and may have a better understanding of the congestion levels at the EDs from experience. We model this possible learning effect as the total number of visits to any public ED in the calendar year of 2019

Table 8 Conditional Logit: Walk-in vs Ambulance Patients

	Triage 3		Triage 4		Triage 5	
	Walk-in	Ambulance	Walk-in	Ambulance	Walk-in	Ambulance
Waiting time (hr)	0.000	-0.002	-0.027***	-0.004	-0.033***	0.003
	(0.001)	(0.002)	(0.001)	(0.002)	(0.004)	(0.035)
Travel distance (km)	-0.403***	-0.590***	-0.380***	-0.452***	-0.384***	-0.662***
	(0.001)	(0.002)	(0.001)	(0.002)	(0.003)	(0.048)
N	1,759,475	1,645,060	4,562,180	782,060	328,890	4,335
Visits	351,895	329,012	912,436	156,412	65,778	867
Patients	286,794	216,459	648,060	131,591	53,037	715
Pseudo-R^2	0.493	0.590	0.558	0.584	0.573	0.668

Notes. Standard errors in parentheses. ED fixed effects and cross harbor controls are not shown.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

whether the visit happened before or after the focal visit by the patient. The average total number of visits are 2.8, 3.3, and 6.2 for triage level 3, 4, and 5 patients respectively while the median is 2 visits for all triage levels. We find that the number of visits to an ED has a significant impact on class membership among triage level 4 and 5 patients, but the magnitude is negligible compared to the impact of age and residential district (Table 9).

6.7. Geographical districts

In the main results, Table 3, we capture the effect of the patient's geographic location on class membership at the administrative territory level with three categories. We delve into a more granular District of Residence Cluster level which the Hospital Authority manages hospitals and deliver public healthcare services through seven hospital clusters: Hong Kong East, Hong Kong West, Kowloon Central, Kowloon East, Kowloon West, New Territories East, and New Territories West. We find that Kowloon Central and East cluster patients have especially low awareness to the announcement system in triage level 4 and 5 patients (Table A2).

7. Counterfactual Analysis: Simulation on Patient Awareness and Announcement Update

In this Section, we perform counterfactual analyses on two key system parameters: the percentage of the public that are aware of and sensitive to the ED waiting time announcement system and the length of the window used to calculate the announced waiting time. Based on the discrete choice framework of the main empirical model (equation 4), we simulate a virtual healthcare network of three EDs. We analyze how the system parameters affect the overall system performance measures: the average actual waiting time at the ED and the number of left without being seen (LWBS) patients.

Table 9 Latent Class Conditional Logit: Frequent Visitor's Class Membership

	Triage 3	Triage 4	Triage 5		Triage 3	Triage 4	Triage 5
ED choice				Class 1 membership			
Class1				Age10	0.310	-0.330***	-0.279
Waiting time (hr)	-5.183***	-1.798***	-1.358***		(0.247)	(0.093)	(0.348)
	(0.208)	(0.033)	(0.076)	Age20	1.063***	0.602***	0.474
Membership shares	0.8%	3.1%	4.7%		(0.161)	(0.059)	(0.274)
Class2				Age30	0.831***	0.374***	-0.159
Waiting time (hr)	0	0	0		(0.158)	(0.060)	(0.290)
	-	-	-	Age40	0.683***	-0.109	-0.530
Membership shares	99.2%	96.9%	95.3%		(0.164)	(0.067)	(0.298)
Travel distance (km)	-0.415***	-0.395***	-0.398***	Age50	0.252***	-0.408	-0.830**
	(0.001)	(0.001)	(0.002)		(0.171)	(0.071)	(0.322)
				Age60	-0.283	-0.971***	-2.067***
Model data statistics					(0.188)	(0.089)	(0.545)
N	1,759,475	4,562,180	328,890	Age70	-1.101***	-2.421***	
Visits	351,895	912,436	65,778		(0.222)	(0.196)	
Patients	286,794	648,060	53,037	HKI	1.081***	0.899***	1.047***
					(0.172)	(0.075)	(0.371)
Travel distance/waiting time (km/hr) ratio				NT	1.611***	1.297***	2.015***
Class 1	12.5	4.5	3.4		(0.141)	(0.056)	(0.250)
				No. of Visits	0.037	-0.014***	-0.013***
Class 2	0.0	0.0	0.0		(0.020)	(0.002)	(0.003)

Notes. Standard errors in parentheses. ED fixed effects and cross harbor controls are not shown.

HKI: Hong Kong Island, NT: New Territories

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We study the public hospital network on the administrative territory of Hong Kong Island, where there are three EDs with different capacity and reputation, namely Queen Mary Hospital (ED 1), Pamela Youde Nethersole Eastern Hospital (ED 2), and Ruttonjee Hospital (ED 3). Patients residing on Hong Kong Island need to cross the harbor either via a tunnel or ferry to reach any other public ED that is not on the island. Hence, the three EDs and its patients can be considered as a relatively closed network. The majority of visits to these three EDs are by Hong Kong Island residents, with residential address belonging to one of the 38 districts on the island (denoted by $m = 1, \ldots, 38$). For patients who reside outside of the island, we denote the district as 0. We simulate patient arrivals with a two step process. First, patient arrivals are simulated as a non-homogeneous Poisson process with hourly arrival rates estimated from the patient visit data. This allows us to capture the intraday patient traffic pattern. Second, we simulate the following attributes for each patient visit by random generation: residential district ($m = 0, 1, 2, \ldots, 38$), triage level ($l =$

$1,2,3,4,5$), and arrival mode (either walk-in or by ambulance), with each probability distribution also estimated from the data.

We follow the practice of waiting time announcement in Hong Kong to simulate the announced waiting time AWT_j for ED $j = 1,2,3$, using the 95th percentile of waiting times of patients who finished waiting (at the time a patient first accessed by a physician) in the past three hours, and update the information every 15 minutes. For each arrival patient, we use the AWT_j 30 minutes before the patient's actual arrival time as the observed waiting time $Waiting_j$. Each ED is modeled as an $M(t)/M(t)/s(t)$ queue with abandonment. We estimated the arrival rates, the service rates, and number of physicians on duty for each hour-of-day from the data, and we assume that patients who wait longer than a tolerance time threshold will leave the system with a reneging probability. We tune the values of the reneging probability and the tolerance time so that the output performance is consistent with the data (See subsection 7.1).

We take the structure of the latent class conditional logit model (equation 4) and estimates from Table 3 to simulate the ED choices of patients that satisfy the following conditions: walk-in, triage level $l \geq 3$ and resident district $m > 0$. Similar to equation 1, for each patient, the utility of attending ED j, U_j, is obtained from equation 1 which captures the ED fixed effect, distance to each ED, and the announced ED waiting time which reflects the ED choice and the arrival to each ED from previous patients in the simulation. The patient will attend ED $i \in \{1,2,3\}$ with probability $e^{U_i}/\sum_{j=1}^{3} e^{U_j}$. The fixed effects of three EDs are provided in the Appendix (Table A1). ED choice of patients excluded from the empirical study data, those who either arrive by ambulance, or belong to triage levels 1 or 2, or reside in a district outside of the island ($m = 0$), were simulated based on the district level empirical data. For instance, for an ambulance patient from Wan Chai district, we first compute the overall percentage of ambulance patients from Wan Chai attending ED 1, 2, and 3 respectively in the data. We simulate the ED choice for the focal patient based on this empirical distribution.

The latent class conditional logit model assumes that patients have different sensitivity to the announced waiting time denoted as their class. In the counterfactual study, we also simulate each patient's sensitivity class c, based on his/her triage level. We keep consistency with the estimated results and parameterize the proportion of patients who are sensitive (denoted by p_l for triage level l) at 0.8%, 3.1%, and 4.8% for levels 3, 4, and 5 respectively. The probability of belonging to the sensitive class is denoted as p_l.

7.1. Simulation model validation

We set the sensitive proportion of patients, p_l, to the values estimated from data, i.e., $p_3 = 0.8\%, p_4 = 3.1\%$ and $p_5 = 4.8\%$. We simulate the sequence of ED choices for 100 replications, each with a run length of 396 days. The first 31 days serve as warm up and the simulation performance is calculated from the remaining 365 days of visit simulations. We compare the number of daily patient visits, number of LWBS patients, and average waiting time between the simulation and empirical data. (Table 10).

Table 10 Simulation Model Validation: Aggregate Comparison with Data

	Data				Simulation			
	ED 1	ED 2	ED 3	Total	ED 1	ED 2	ED 3	Total
Patient visits	334.3	346.8	188.9	867.0	337.0	344.7	188.1	869.8
Patients served	328.4	333.8	183.0	845.2	330.5	331.1	182.6	846.3
Left without being seen patients	5.9	12.9	5.8	24.7	6.6	13.7	5.5	23.5
Average waiting time (hours)	1.0	1.5	1.4	1.3	1.0	1.4	1.4	1.2

Notes. Number of patients are daily average. Waiting time is the actual time patients have waited at the ED they attended.

We simulate the total number of arrivals (sum of all three EDs) as a non-homogeneous Poisson process with arrival rates estimated from the data, thus the consistency in total daily patient visits with the data is expected. However, it is important to note that the arrivals to each ED in the simulation model depends on the patient choice model. The consistency in daily patient visits to each individual ED between the simulation and data shows that our latent class conditional logit model can effectively replicate the patient choice decisions. The values of average waiting time and number of LWBS patients demonstrate that the simulated queue can produce consistent performance measures compared to the data.

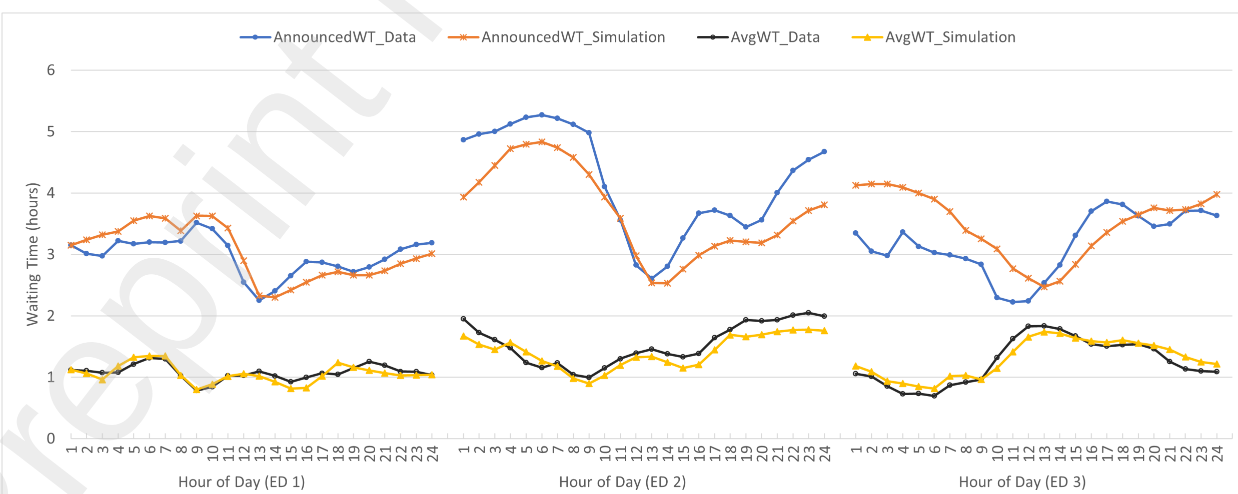


Figure 3 Simulation Model Validation: Intraday Pattern for Average Waiting Time (AvgWT) and Announced Waiting Time (AnnouncedWT) by Hour of Day

We also validate the simulation model's capability to capture intraday patterns by comparing average actual waiting time and average announced waiting time by hour-of-day for each ED. More specifically, we calculate the average waiting time for patients who arrived in the same hour-of-day, and the average announced waiting time in different hours-of-day for each of the three EDs from the simulation model, and compare those with data. The results are shown in Figure 3. We observe that the simulated hourly average waiting time for each ED matches closely with the data. While the simulated average announced waiting time matches closely with the data for ED 1, in ED 2 and 3 the simulation results exhibit mild deviation from the data. This gap can be attributed to the higher difficulty in estimating tail percentiles (announced waiting times are 95th percentile values) compared to that of estimating mean values. However, the overall trend and relative scale of the simulated announced waiting times are consistent with the data.

In conclusion, our simulation framework is capable of capturing the important feature of the study: patient choice behavior, and can provide a reasonably good estimation of the overall system performance measures: the average waiting time and the total number of LWBS patients.

7.2. Simulation results

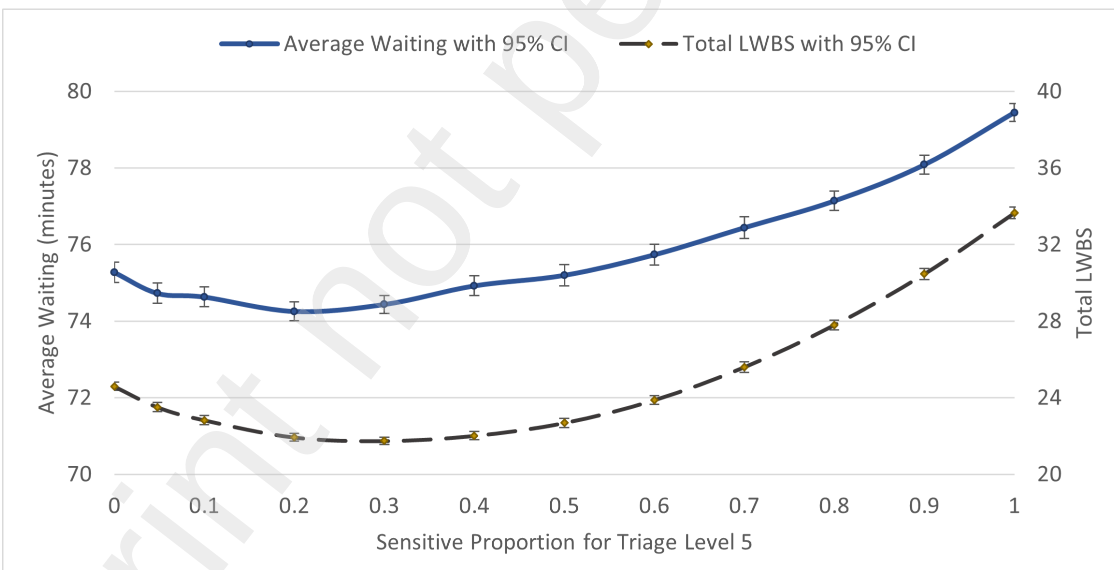


Figure 4 Impact of Patient Awareness on Average Waiting Time and Left Without Being Seen (LWBS) Patients

	Sensitive proportion parameters											
Triage level 3 (p_3)	0.000	0.008	0.017	0.034	0.051	0.068	0.084	0.101	0.118	0.135	0.152	0.169
Triage level 4 (p_4)	0.000	0.031	0.065	0.129	0.194	0.258	0.323	0.388	0.452	0.517	0.581	0.646
Triage level 5 (p_5)	0.000	0.048	0.100	0.200	0.300	0.400	0.500	0.600	0.700	0.800	0.900	1.000
Overall	0.000	0.025	0.053	0.105	0.158	0.211	0.263	0.316	0.368	0.421	0.474	0.526

Notes. Overall is the average of the three triage classes weighted by the unique number of patients in each class.

In the main results, we find that only a small fraction of patients are aware of and sensitive to the waiting time announcement system. From the policy makers perspective, a small awareness among the public isn't ideal as it is most likely limiting the pooling effect of the announcement system. Hence, a key question to be answered is how much of the announcement system's effectiveness is being limited due to the lack of patient awareness.

Using the simulation framework, we first explore how the level of patient awareness of the waiting time announcement system would affect the operational performance of the ED network measured by the average waiting time and number of LWBS patients. We parameterize patient awareness as the proportion of triage level 5 patients sensitive to the waiting time announcement system. We simulate counterfactual networks of the three EDs by increasing the sensitive proportion of triage level 5 from 0 to 1, and increase the sensitive proportions of triage levels 3 and 4 proportionally based on the awareness from the empirical model estimation. For instance, if 100% of triage level 5 patients are aware of the announcement system the maximum awareness in the simulation study, 16.9% and 64.6% of triage level 3 and 4 patients will be aware. This is based on the assumption that the advertising effect of promoting the announcement system will have a consistent effect across the different triage levels.

Figure 4 shows that when the sensitive proportion increases from 0 to a moderate level, both the average waiting time and the number of LWBS patients decrease. This result indicates that the announcement system can help balance patient traffic across the EDs and achieve the intended pooling effect as patients start to become aware of and sensitive to the announced waiting time information. Hence, implementing a waiting time announcement system is beneficial to a network of EDs without one, initially. However, when the sensitive proportion reaches a certain level, further increasing the proportion of sensitive patients can hinder the ED operations and negatively affect the patients with longer waits and more leaving without receiving treatment. One possible reason of the negative impact when more patients respond to the announcement is the delayed nature of the information, i.e., sensitive patients are not responding to the most up-to-date crowding conditions of each ED when making the ED choice decisions. Since the provided delay information is the 95th percentile in the past three hours, it may not accurately represent the latest crowding situations. And if more patients become sensitive and respond to the announcements, they may cause immediate large congestion at a previously less congested (short announced waiting time) ED by a huge flow into it and away from an ED with a long announced waiting time which may have been highly congested hours ago and not recently. Hence, when more and more patients choose to avoid long waiting by referring to the delayed information, it may create waiting time oscillations as mentioned in Dong et al. (2019). Similar oscillation phenomena have also been found in the transportation literature (Mahmassani and Jayakrishnan 1991, Yoshii et al. 1996), where the navigation system serves a similar role of directing demand as waiting time information in our setting. The transportation literature further shows that there exists an optimal penetration rate of navigation systems (Dai et al. 2017, Emmerink et al. 1995, Litescu et al. 2015).

Therefore, secondly, we investigate what if the information is less delayed. In the current practice of Hong Kong, announced waiting time is calculated based on the 95th percentile of patients in the past three hours (AWT window = 3 Hours). As the counterfactual, we simulate patient choices in the three EDs network with waiting time announcements calculated with shorter time windows. We simulate the network with 1 hour and 2 hours AWT window for 100 replications each and plot the average waiting time and total number of LWBS patients in Figure 5.

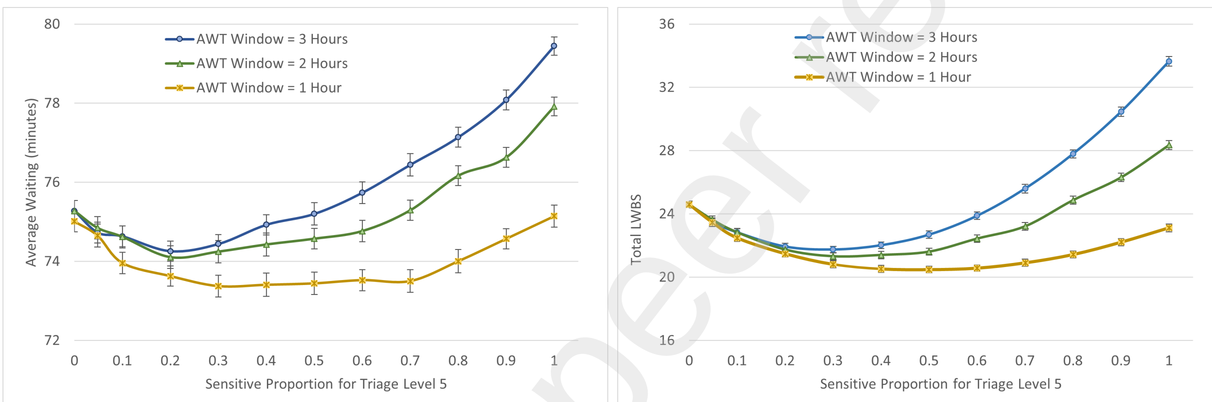


Figure 5 Impact of Announced Waiting Time (AWT) Update Window and Patient Awareness on Average Waiting Time and Left Without Being Seen (LWBS) Patients

We find that when the AWT window is shorter, the average waiting time and number of LWBS patients are smaller, and this positive effect is more significant when patient awareness is larger. At the current level of awareness, 4.8% among triage level 5 patients or 2.5% across triage level 3, 4, and 5 patients, the benefit of providing more recent congestion information only is insignificant. So there is no immediate need to update the system by shortening the update window. However, once more patients become sensitive to the announced information, the benefit of shortening the update window is substantial as it reduces the oscillation of patients across EDs. While it may not be desirable for the policy makers to advertise the announcement system more beyond a certain point, roughly 20 to 30% of awareness, when using a system with a long update window of 3 hours, they need not be worried of high awareness when the system has a short window. In the case of Hong Kong Island, it would be ideal to increase the overall awareness between 15.8% and 31.5% (or 30% to 60% among triage level 5 patients) and reduce the announced waiting time update window to 1 hour where both average waiting time and LWBS patients can be improved from the current operations. For instance, if overall awareness increases to 15.8% and the announcement update window reduces to 1 hour, policy makers can expect a reduction in the average waiting time and number of LWBS patients by 1.1 minutes (1.4%) and 2.7 patients (11.5%) respectively.

Overall, the counterfactual analysis shows that increasing the level of awareness among the patients may not always benefit the system and there may exist a sweet-spot of awareness that depends on the update window of the announcement system. On the other hand, shortening the update window can significantly improve the system performance if more patients become aware of the announcement system and utilize it in their ED visit decisions.

8. Conclusion and Future Research

While many communities across the world have adopted ED delay announcement systems, the extent to which those systems have been used by the patients or their effect on patient ED choice decisions have not been well studied. This paper addresses important questions pertaining to the operation and design of ED delay announcement systems and coordination of EDs within the urban hospital network. We model the individual patient's decision of which ED to attend as a latent class conditional logit choice where each patient belongs to an underlying class of sensitivity to the delay announcement. The model assumes that patients in the same class have a constant sensitivity to the announced waiting time while this sensitivity may vary across different classes. We structurally estimate how many patients are aware of and sensitive to the delay announcement system as well as the patients' sensitivity to the ED waiting time and the travel distance to the ED. We also develop a simulation framework based on the empirical model and estimated system parameters, and analyze the sensitivity of operational ED performance to system parameters and the information provided in the delay announcement.

Applying the framework to a patient visit-level data set from 17 public hospital EDs in Hong Kong SAR, we find that 0.8%, 3.1%, and 4.8% of triage level 3, 4, and 5 patients respectively (2.5% being the overall average across three triage levels) are aware of the delay announcement system and sensitive to the announced waiting time. While patients are sensitive to the travel distance to the ED in general, we find that the few but sensitive triage level 3, 4, and 5 patients are willing to travel an additional 12.3, 4.6, and 3.5 km, respectively, (6.8 km being the overall average across three triage levels) to save one hour of waiting at the ED after arrival. These results highlight that urgent patients are less likely to search for and respond to the announced waiting time but those who do are more sensitive to the waiting at the ED, to the point that they rather travel farther, hence, spend more time on travel to save the time they wait at the ED. While we find that patients are heterogeneous in their reaction to the delay announcements, we also find that this heterogeneity depends on the level of urgency. This finding is also supported by our robustness analysis result showing that patient sensitivity to the announced waiting time may vary with her clinical condition.

This provides an opportunity for future research on designing ED delay announcement systems with multiple types of delay information shared. Policy makers may consider offering reference waiting time for multiple classes of urgency, i.e., both urgent and non-urgent patients (triage level 3 and 5), simultaneously so patients can get a better estimate of their waiting time based on the self-assessed urgency level of their condition. Current delay announcement system designs in most countries announce only one kind of waiting time information, which is typically a good reference for only a small proportion of the entire patient population. Meanwhile, in the Perth area of Western Australia, the local government announces the average waiting time of triage level 4 patients along with two pieces of census information, the number of patients waiting to be seen in ED and the total number of patients in EDs for a network of 10 EDs.[7] It will be interesting to study how patients respond to different operational information—waiting time, queue length, and total census—when these informational components are offered collectively or separately. Such a study can advance our understanding of what type of information is most effective in ED delay announcement systems.

In our paper, we also study the effect of the different levels of "recency" of waiting information that should be provided to improve ED performances. Our simulation study shows that increasing the awareness of announced system to the public improves system congestion and service quality with fewer LWBS patients. However, a high awareness among the patients may hurt the patients and worsen overall system performance. Our additional analysis suggests that providing information with smaller time lag can reduce both average waiting time and number of LWBS patients. This also increases the system robustness to varying awareness levels. How patients respond to different lags of information is an area that needs to be studied as the inaccuracy of the lagged information may be a reason why not many patients take the delay announcement into account during their decision process. It may be that more patients are aware of the announcement system but do not take the information into account due to the lack of accuracy. Linking patient visit records with surveys on the patient awareness of the delay announcement system can answer such question. Further progress in this direction could be achieved by implementing more accurate predictors of ED waiting time developed by numerous researchers and studying the resulting patient response.

In our analysis, we also find that older patients are less likely to be responsive to the delay announcements, and that there exists a certain geographic effect on the level of awareness/sensitivity to the delay announcements. The Hong Kong government can take these results into account when they plan to increase the awareness of the ED wait time announcement system.

[7] `https://ww2.health.wa.gov.au/Reports-and-publications/Emergency-Department-activity/Data?report=ed_activity_now`

Overall, we provide a framework that can be applied to other delay announcement systems and the customer (or patient) visit records. We hope that our study can encourage future research in understanding patient's service provider choice behavior in delayed-affected networks of healthcare service providers, including not only emergency medicine, but also surgical and outpatient services.

References

Z. Akşin, B. Ata, S. M. Emadi, and C.-L. Su. Impact of delay announcements in call centers: An empirical approach. *Operations Research*, 65(1):242–265, 2017.

E. Ang, S. Kwasnick, M. Bayati, E. L. Plambeck, and M. Aratow. Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management*, 18(1):141–156, 2016.

M. Armony, N. Shimkin, and W. Whitt. The impact of delay announcements in many-server queues with abandonment. *Operations Research*, 57(1):66–81, 2009.

S. Arora, J. W. Taylor, and H.-Y. Mak. Probabilistic forecasting of patient waiting times in an emergency department. *Available at SSRN 3614760*, 2020.

A. Bassamboo and R. Ibrahim. A general framework to compare announcement accuracy: Static vs. les-based announcement. *Management Science*, 67(7):4191–4208, 2021.

H. Chen and M. Frank. Monopoly pricing when customers queue. *IIE Transactions*, 36(6):569–581, 2004.

R. Dai, Y. Lu, C. Ding, and G. Lu. The effect of connected vehicle environment on global travel efficiency and its optimal penetration rate. *Journal of Advanced Transportation*, 2017, 2017.

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

J. Dong, E. Yom-Tov, and G. B. Yom-Tov. The impact of delay announcements on hospital network coordination and waiting times. *Management Science*, 65(5):1969–1994, 2019.

N. M. Edelson and D. K. Hilderbrand. Congestion Tolls for Poisson Queuing Processes. *Econometrica*, 43(1): 81–92, 1975. ISSN 0012-9682. doi: 10.2307/1913415. URL `https://www.jstor.org/stable/1913415`. Publisher: [Wiley, Econometric Society].

R. H. Emmerink, K. W. Axhausen, P. Nijkamp, and P. Rietveld. Effects of information in road transport networks with recurrent congestion. *Transportation*, 22(1):21–53, 1995.

N. Fatma and V. Ramamohan. Patient diversion using real-time delay prediction across healthcare facility networks, 2021.

P. Feigin. Analysis of customer patience in a bank call center. In *Working paper, The Technion*, 2006.

R. Hassin. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica: Journal of the Econometric Society*, pages 1185–1195, 1986.

R. Ibrahim and W. Whitt. Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science*, 55(10):1729–1742, 2009.

R. Ibrahim and W. Whitt. Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals. *Production and Operations Management*, 20(5):654–667, 2011.

O. Jouini, Y. Dallery, and Z. Akşin. Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics*, 120(2):389–399, 2009.

O. Jouini, Z. Akşin, and Y. Dallery. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management*, 13(4):534–548, 2011.

S. Litescu, V. Viswanathan, M. Lees, A. Knoll, and H. Aydt. Information impact on transportation systems. *Journal of Computational Science*, 9:88–93, 2015.

H. S. Mahmassani and R. Jayakrishnan. System performance and user response under real-time information in a congested traffic corridor. *Transportation Research Part A: General*, 25(5):293–307, 1991.

A. Mandelbaum and S. Zeltyn. Data-stories about (im) patient customers in tele-queues. *Queueing Systems*, 75(2):115–146, 2013.

P. Naor. The Regulation of Queue Size by Levying Tolls. *Econometrica*, 37(1):15–24, 1969. ISSN 0012-9682. doi: 10.2307/1909200. URL `https://www.jstor.org/stable/1909200`. Publisher: [Wiley, Econometric Society].

J. Pender, R. H. Rand, and E. Wesson. Managing information in queues: the impact of giving delayed information to customers. *arXiv preprint arXiv:1610.01972*, 2016.

R. Shone, V. A. Knight, and J. E. Williams. Comparisons between observable and unobservable m/m/1 queues with respect to optimal customer behavior. *European Journal of Operational Research*, 227(1): 133–141, 2013.

S. P. Singh, M. Delasay, and A. A. Scheller-Wolf. Evaluating the First-Mover's Advantage in Announcing Real-Time Delay Information. *SSRN Electronic Journal*, 2017. ISSN 1556-5068. doi: 10.2139/ssrn.3081583. URL `https://www.ssrn.com/abstract=3081583`.

E. Webb, Q. Yu, and K. Bretthauer. Linking delay announcements, abandonment, and service time. *Abandonment, and Service Time (March 25, 2019)*, 2019.

W. Whitt. Improving service by informing customers about anticipated delays. *Management science*, 45(2): 192–207, 1999.

H. I. Yoo. lclogit2: An enhanced command to fit latent class conditional logit models. *The Stata Journal*, 20(2):405–425, 2020.

T. Yoshii, H. Akahane, and M. Kuwahara. Impacts of the accuracy of traffic information in dynamic route guidance systems. In *Proceedings of The 3rd Annual World Congress on Intelligent Transport Systems (CD-ROM)*, volume 87, page 98. Citeseer, 1996.

Q. Yu, G. Allon, and A. Bassamboo. How do delay announcements shape customer behavior? an empirical study. *Management Science*, 63(1):1–20, 2017.

Q. Yu, G. Allon, A. Bassamboo, and S. Iravani. Managing customer expectations and priorities in service systems. *Management Science*, 64(8):3942–3970, 2018.

Q. Yu, G. Allon, and A. Bassamboo. The reference effect of delay announcements: A field experiment. *Management Science*, 2021.

Appendix

Table A1 Fixed Effects of Table 3 ED Choice Model

	Triage 3	Triage 4	Triage 5
Cross harbor	-3.811***	-3.282***	-3.183***
	(0.048)	(0.031)	(0.104)
Caritas Medical Centre	2.549***	1.113***	1.424***
	(0.030)	(0.018)	(0.112)
Kwong Wah Hospital	3.273***	0.631***	0.247*
	(0.030)	(0.018)	(0.113)
North District Hospital	-0.683***	-1.546***	-1.094***
	(0.024)	(0.012)	(0.060)
North Lantau Hospital	-3.786***	-1.667***	-3.951***
	(0.054)	(0.049)	(0.249)
Princess Margaret Hospital	3.162***	0.738***	0.484***
	(0.029)	(0.018)	(0.112)
Pok Oi Hospital	-1.233***	-1.684***	-0.656***
	(0.039)	(0.022)	(0.094)
Prince of Wales Hospital	2.446***	0.841***	-1.132***
	(0.024)	(0.015)	(0.108)
Pamela Youde Nethersole Eastern Hospital	1.981***	0.231***	0.184
	(0.046)	(0.026)	(0.130)
Queen Elizabeth Hospital	3.618***	0.781***	0.717***
	(0.030)	(0.018)	(0.113)
Queen Mary Hospital	2.665***	0.800***	-0.001
	(0.053)	(0.029)	(0.137)
Ruttonjee Hospital	0.412***	-0.537***	-0.325**
	(0.049)	(0.027)	(0.130)
Tseung Kwan O Hospital	2.788***	0.819***	-0.066
	(0.030)	(0.020)	(0.119)
Tuen Mun Hospital	-0.289***	-1.857***	-1.691***
	(0.040)	(0.023)	(0.101)
Tin Shui Wai Hospital	-2.314***	-2.191***	-0.373***
	(0.042)	(0.023)	(0.096)
United Christian Hospital	2.696***	0.639***	0.987***
	(0.029)	(0.018)	(0.113)
Yan Chai Hospital	1.995***	0.803***	-0.171
	(0.028)	(0.017)	(0.111)

Notes. Standard errors in parentheses. Alice Ho Miu Ling Nethersole Hospital is excluded for identification.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A2 Latent Class Conditional Logit: Class Membership by 7 Geographical Districts

	Triage 3	Triage 4	Triage 5
ED choice			
Class1			
Waiting time (hr)	-4.928***	-1.716***	-1.321***
	(0.198)	(0.032)	(0.079)
Membership shares	0.9%	3.5%	4.7%
Class2			
Waiting time (hr)	0	0	0
	-	-	-
Membership shares	99.1%	96.5%	95.3%
Travel distance (km)	-0.416***	-0.395***	-0.398***
	(0.001)	(0.001)	(0.002)
Model data statistics			
N	1,759,475	4,562,180	328,890
Visits	351,895	912,436	65,778
Patients	286,794	648,060	53,037
Travel distance/waiting time (km/hr) ratio			
Class 1	11.9	4.3	3.3
Class 2	0.0	0.0	0.0

	Triage 3	Triage 4	Triage 5
Class 1 membership			
Age10	0.307	-0.235**	-0.403
	(0.247)	(0.091)	(0.315)
Age20	1.063***	0.641***	0.258
	(0.159)	(0.060)	(0.249)
Age30	0.846***	0.431***	-0.345
	(0.157)	(0.061)	(0.266)
Age40	0.682***	-0.048***	-0.700**
	(0.164)	(0.067)	(0.272)
Age50	0.262	-0.330***	-0.943***
	(0.169)	(0.070)	(0.291)
Age60	-0.249	-0.870***	-2.125***
	(0.185)	(0.088)	(0.478)
Age70	-1.112***	-2.190***	
	(0.218)	(0.183)	
HKE	1.157***	0.408***	-0.205
	(0.285)	(0.096)	(0.461)
HKW	1.137***	0.688***	0.643
	(0.311)	(0.111)	(0.437)
KC	-0.883	-0.605***	-1.801*
	(0.481)	(0.144)	(0.745)
KE	0.565*	-0.631***	-1.218**
	(0.284)	(0.107)	(0.439)
NTE	1.348***	0.631***	1.295***
	(0.260)	(0.077)	(0.300)
NTW	1.958***	1.111***	0.888**
	(0.253)	(0.072)	(0.281)

Notes. Standard errors in parentheses. ED fixed effects and cross harbor controls are not shown.

HKE: Hong Kong Island East, HKE: Hong Kong Island West, KC: Kowloon Central, KE: Kowloon East, NTE: New Territories East, NTW: New Territories West

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$