Appendix A: End-Of-Shift Effect

As discussed in Observation 1 from Section 4.1, we find evidence of the ED decision maker's patient routing decisions being affected by the physicians' shift. The exact physician shift schedule is not available for this study. However, we can utilize the patient visit data structure to approximate the physician shift schedule. Since the data contains physician ID for each patient visit, we can sort patient visits first by physician ID and next by *selection time* of each patient (time of choice incident). This allows us to compute, for each and every patient choice incident, the time gap until the very next patient choice incident treated by the same physician. If the time gap is larger than 480 minutes, we consider that as a sign of change in the physician's shift—the former patient being the last to be treated in a shift and the latter patient being the first in a new shift. If the time gap is smaller than 480 minutes, the patients are considered to be treated in the same shift.

With the physician shift being approximated, we explore models controlling for various end-of-shift effect time windows. We hypothesize that shift changes will mainly affect patient routing decision across different triage levels rather than within triage levels. Thus, we introduce the end-of-shift effect by interacting the triage level-2 intercept, $Triage2_j$, with a time-dependent binary end-of-shift indicator, EOS_t. The coefficient of $Triage2_j \cdot EOS_t$ effectively captures the impact of the choice incident occurring at the end of a physician's shift compared to a non-end-of-shift choice incident on the prioritization behavior between triage level-2 and -3 patients. We explore five different time windows as the "end" of a shift. First, we consider only the very last choice incident of a physician shift as the end and all other choice incidents as normal time period. Next, we consider the very last choice incident and choice incidents that occurred within a certain time—ranging from 15 to 60 minutes—before that very last choice incident as the end of the focal physician's shift.

Table 5 End-Of-Shift Effect: BIC Score for Models with Varying End-Of-Shift Windows

End-of-shift window	ED A	ED B	ED C	ED D
Last choice incident only	233966.8	115758.8	119813.5	95983.8
15 mins from last choice incident	233976.2	115767.6	120094.4	96065.7
30 mins from last choice incident	233977.7	115763.7	120070.6	96048.3
45 mins from last choice incident	233978.3	115763.2	120052.6	96047.4
60 mins from last choice incident	233978.6	115763.1	120014.5	96047.2

Similar to our model selection in Section 5.1, we use the BIC score to compare models with different end-of-shift time windows. Table 5 reports the model fit by end-of-shift definitions. For all four EDs, the *last choice incident only* model consistently outperforms the wider time window models. The BIC performance is quite stable between 15 to 60 minutes time window models, while the gap between the *last choice incident only* model and the other models is much larger in all four EDs. Hence, we conclude that the end-of-shift effect exists only for the very last choice incident in a physician's shift and exclude those choice incidents from the main analysis.

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

Table 6 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 perform independence tests between the physician IDs[5] 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 6 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 D: 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 E: 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

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

where choice probability, $P(c(t)|\boldsymbol{\Sigma}(t))$ 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

Table 7 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.

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 7).

Appendix F: 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.

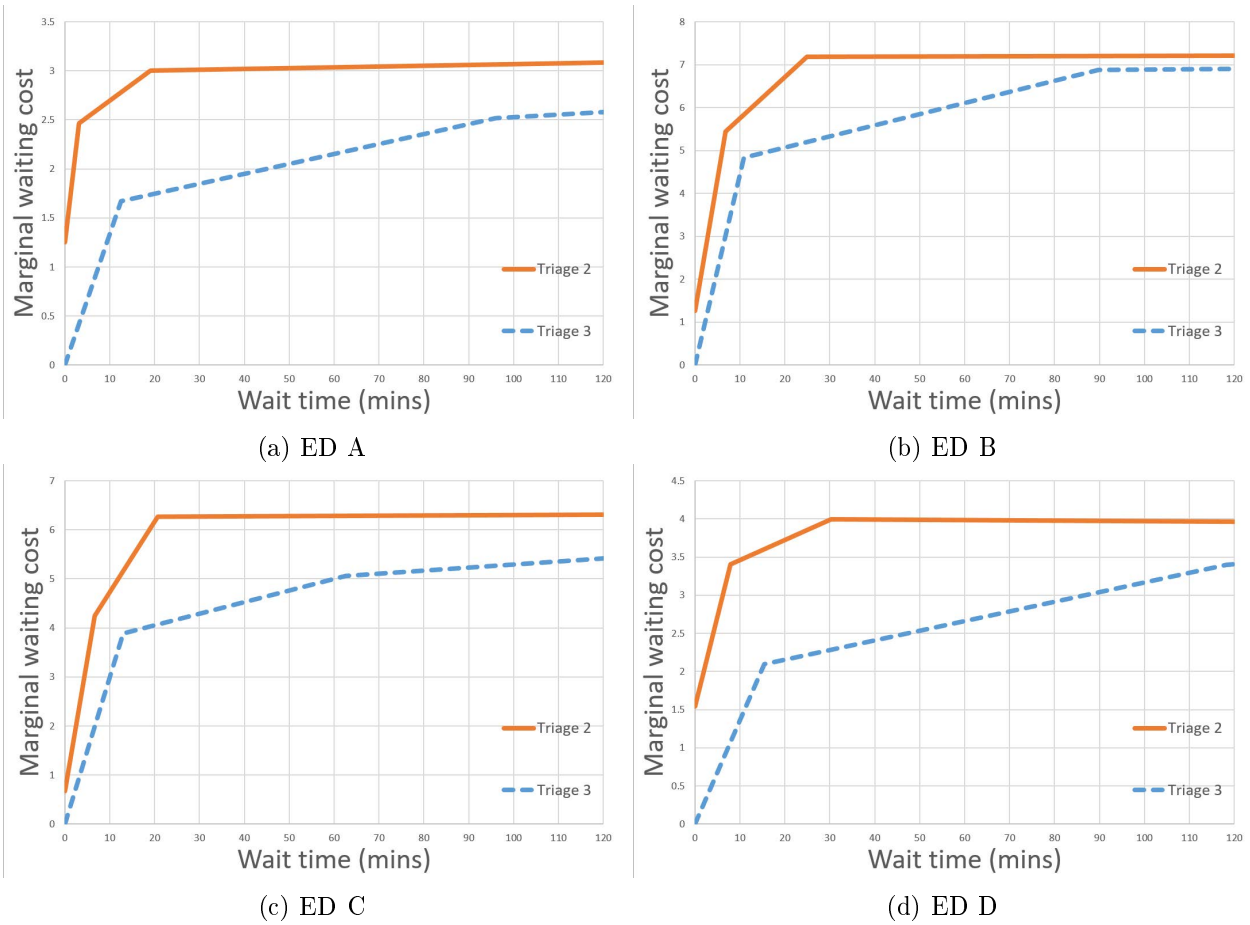
To test whether the IIA property is a reasonable assumption for the observed data, we investigate the mixed logit model, which has been discussed in Appendix D. The conditional logit model used in this paper is a special case of the mixed logit model when the random slopes and intercepts of the piece-wise 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 G: 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 H: 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 $(c, \mathbf{\Sigma})$[6]. 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[7](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 $\boldsymbol{\gamma}:=\{\gamma_b^{Tri}|b=1,\ldots,B,\ Tri=2,3\}$[8]. 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 $n\to\infty$.

Theorem H.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*

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

- *(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 $Tri(j)=2,3$[9].*
- *(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 $n\to\infty$,

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

Proof. By Theorem 2.1 of (Newey and McFadden 1994), to prove Equation (22), it suffices to show that (i) The true log-likelihood function $Q(\boldsymbol{\theta}) = \ln L(\hat{\boldsymbol{\theta}})$ is well defined and uniquely maximized at $\boldsymbol{\theta}$; (ii) Θ is compact; (iii) $\ln L(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$; (iv) $Q^n(\boldsymbol{\theta}) := \frac{1}{n} \ln L_n(\boldsymbol{\theta})$ converges uniformly in probability to $Q(\boldsymbol{\theta})$. We next prove each of the above conditions.

Note that it is common to define a likelihood function as in (6) using probabilities conditional on part of data (i.e, $\boldsymbol{\Sigma}$). In this case, the true log-likelihood function $\ln L(\hat{\boldsymbol{\theta}})$ is the expectation of the conditional probabilities with respect to the marginal distribution for that part of data (i.e., the limiting distribution of $\boldsymbol{\Sigma}$, π). That gives

$$Q(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\Sigma} \sim \boldsymbol{\pi}} \mathbb{E}_c \ln P(c|\boldsymbol{\Sigma}, \boldsymbol{\theta}).$$

We next prove that $Q(\boldsymbol{\theta})$ has a unique maximizer. To prove that, in view of Lemma 2.2 of (Newey and McFadden 1994), it suffices to prove that the MLE $\boldsymbol{\theta}$ is identified. That is, if $\boldsymbol{\theta}^1 \neq \boldsymbol{\theta}^2$, then when $n \to \infty$, with probability approaching to one, we will have a $\boldsymbol{\Sigma}$ in the choice data sequence such that

$$P(c|\boldsymbol{\Sigma}, \hat{\boldsymbol{\theta}}^1) \neq P(c|\boldsymbol{\Sigma}, \hat{\boldsymbol{\theta}}^2). \tag{23}$$

To prove the above statement, we first show that when $n \to \infty$, with probability approaching one,

$$\boldsymbol{\theta}^1 \neq \boldsymbol{\theta}^2 \Rightarrow V_{jt}(wait_j(t), \boldsymbol{X_j}|\hat{\boldsymbol{\theta}}^1) - V_{jt}(wait_j(t), \boldsymbol{X_j}|\hat{\boldsymbol{\theta}}^2) \not\equiv C, \tag{24}$$

where C denotes any constant. We next prove Equation (24) by considering the following two cases.

If $\boldsymbol{\alpha}^1 \neq \boldsymbol{\alpha}^2$, we must have

$$f_c(\boldsymbol{X_j}|\boldsymbol{\alpha}^1) - f_c(\boldsymbol{X_j}|\boldsymbol{\alpha}^2) = \boldsymbol{X_j}^T(\boldsymbol{\alpha}^1 - \boldsymbol{\alpha}^2) \not\equiv C \tag{25}$$

Because otherwise, we will have $\boldsymbol{X_j}^T(\boldsymbol{\alpha}^1 - \boldsymbol{\alpha}^2) - C \equiv 0$ for all j. The only solution to this linear equation is $\boldsymbol{\alpha}^1 - \boldsymbol{\alpha}^2 = 0$ due to the non-multicollinearity assumption (a3).

By (a6), when $n \to \infty$, with probability approaching one we can find choice sets in which a patient with attributes $\boldsymbol{X_j}$ has $wait_j(t) \in [0, \epsilon]$ for some $\epsilon > 0$. By choosing $\epsilon \to 0$, both $f_w^{Tri(j)}(wait_j(t)|\boldsymbol{\theta}^1)$ and $f_w^{Tri(j)}(wait_j(t)|\boldsymbol{\theta}^2)$ converge to zero. Thus, in those choice sets,

$$V_{jt}(wait_j(t), \boldsymbol{X_j}|\hat{\boldsymbol{\theta}}^1) - V_{jt}(wait_j(t), \boldsymbol{X_j}|\hat{\boldsymbol{\theta}}^2) \to (f_c(\boldsymbol{X_j}|\boldsymbol{\alpha}^1) - f_c(\boldsymbol{X_j}|\boldsymbol{\alpha}^2))\mu_j \not\equiv C, \tag{26}$$

where the inequality follows from Equation (25).

If $\boldsymbol{\alpha}^1 = \boldsymbol{\alpha}^2$, then $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ must differ in the $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ part. The two piece-wise polynomials are then different and there must exist a $\tau \in [0, \overline{W}^{Tri(j^1)}]$ such that $f_w^{Tri(j^1)}(\tau|\boldsymbol{\theta}^1) - f_w^{Tri(j^1)}(\tau|\boldsymbol{\theta}^2) \neq 0$. By (a6), we can find a choice set t^1 such that $wait_{j^1}(t^1)$ lies in a neighborhood of τ and thus $f_w^{Tri(j^1)}(wait_{j^1}(t^1)|\boldsymbol{\theta}^1) - f_w^{Tri(j^1)}(wait_{j^1}(t^1)|\boldsymbol{\theta}^2) \neq 0$. Still by (a6), we can find another choice set t^2 such that $wait_{j^2}(t^2)$ lies in a neighborhood of 0 and thus $f_w^{Tri(j^2)}(wait_{j^2}(t^2)|\boldsymbol{\theta}^1) - f_w^{Tri(j^2)}(wait_{j^2}(t^2)|\boldsymbol{\theta}^2) \to 0$. The existence of j^1, t^1, j^2, and t^2 implies that

$$f_w^{Tri(j)}(wait_j(t)|\boldsymbol{\theta}^1) - f_w^{Tri(j)}(wait_j(t)|\boldsymbol{\theta}^2) \not\equiv C. \tag{27}$$

Since $\boldsymbol{\alpha}^1 = \boldsymbol{\alpha}^2$, the f_c part has the same value under parameters $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$. Only the f_w^{Tri} term remains in the difference of $V_{jt}(wait_j(t), \boldsymbol{X_j}|\hat{\boldsymbol{\theta}}^1) - V_{jt}(wait_j(t), \boldsymbol{X_j}|\hat{\boldsymbol{\theta}}^2)$. Thus,

$$V_{jt}(wait_j(t), \boldsymbol{X_j}|\hat{\boldsymbol{\theta}}^1) - V_{jt}(wait_j(t), \boldsymbol{X_j}|\hat{\boldsymbol{\theta}}^2) = (f_w^{Tri(j)}(wait_j(t)|\boldsymbol{\theta}^1) - f_w^{Tri(j)}(wait_j(t)|\boldsymbol{\theta}^2))\mu_j \not\equiv C. \tag{28}$$

We have thus proved Equation (24) for both cases of $\boldsymbol{\alpha}^1 = \boldsymbol{\alpha}^2$ and $\boldsymbol{\alpha}^1 \neq \boldsymbol{\alpha}^2$.

We next prove Equation (23). Since $V_{jt}^1 - V_{jt}^2 \not\equiv C$, we can find two patients j^1 and j^2 in choice sets t^1 and t^2, respectively, such that

$$V_{j^1,t^1}(\hat{\boldsymbol{\theta}}^1) - V_{j^2,t^2}(\hat{\boldsymbol{\theta}}^1) > V_{j^1,t^1}(\hat{\boldsymbol{\theta}}^2) - V_{j^2,t^2}(\hat{\boldsymbol{\theta}}^2). \tag{29}$$

By (a5), there is a positive probability for two patients with attributes $\boldsymbol{X}_{j^1}$ and $\boldsymbol{X}_{j^2}$ to appear in the same choice set t'. Furthermore, by (a6), with a positive probability their wait times $wait_{j^1}(t')$ and $wait_{j^2}(t')$ are in a small neighborhood of $wait_{j^1}(t^1)$ and $wait_{j^2}(t^2)$, respectively. Therefore, the deterministic valuation of patient j^1 and j^2 in choice set t' are approximately given by $V_{j^1,t^1}(\hat{\boldsymbol{\theta}}^i)$ and $V_{j^2,t^2}(\hat{\boldsymbol{\theta}}^i)$, respectively, under parameters $\hat{\boldsymbol{\theta}}^i$ $(i = 1, 2)$. Then the strict inequality (29) implies that patient j^1 will be more favored under parameters $\boldsymbol{\theta}^1$ compared to $\boldsymbol{\theta}^2$. This, expressed by the odds ratio, gives

$$\frac{P(j^1|\boldsymbol{\Sigma}(t'), \boldsymbol{\theta}^1)}{P(j^2|\boldsymbol{\Sigma}(t'), \boldsymbol{\theta}^1)} > \frac{P(j^1|\boldsymbol{\Sigma}(t'), \boldsymbol{\theta}^2)}{P(j^2|\boldsymbol{\Sigma}(t'), \boldsymbol{\theta}^2)}. \tag{30}$$

The above inequality implies that either $P(j^1|\boldsymbol{\Sigma}(t'), \boldsymbol{\theta}^1) \neq P(j^1|\boldsymbol{\Sigma}(t'), \boldsymbol{\theta}^2)$ or $P(j^2|\boldsymbol{\Sigma}(t'), \boldsymbol{\theta}^1) \neq P(j^2|\boldsymbol{\Sigma}(t'), \boldsymbol{\theta}^2)$. In either case, we have shown that Equation (23) holds for $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(t')$. Therefore, when $n \to \infty$, with probability approaching one, Equation (23) is satisfied for some $\boldsymbol{\Sigma}$. Thus, $\boldsymbol{\theta}$ can be identified.

Condition (ii) follows from assumption (a2) directly.

To prove condition (iii), we mainly need to prove that $f_w^{Tri}(\cdot)$ is continuous in $\boldsymbol{\gamma}$, as its continuity with respect to $\boldsymbol{\beta}$ are straightforward. If a break point γ_b^{Tri} has changed by $\delta > 0$, then the polynomial value over all the subsequent intervals $[\gamma_{b'}^{Tri}, \gamma_{b'+1}^{Tri}]$ with $b' \geq b$ will change by at most

$$\sup\{\sum_{k=D+b}^{D+b'} \beta_k^{Tri}(x^D - y^D) | x, y \leq \overline{W}^{Tri},\ |x - y| \leq \delta\} \leq K\delta. \tag{31}$$

where K is a uniform upper bound for the absolute value of the derivative of polynomial $\sum_{k=D+b}^{D+B} \beta_k^{Tri} x^D$ over $[0, \overline{W}^{Tri}]$. Intuitively, if the function curve has been shifted horizontally by a distance δ, then the function value of $f_w^{Tri}(\cdot|\boldsymbol{\theta})$ changes by at most $K\delta$.

Now suppose $\|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2\| = \delta$[10], so the location of each break point can change by at most δ. Since there are B break points, the total change to the function value can be upper bounded as

$$|f_w^{Tri(j)}(wait_j(t), \boldsymbol{X_j}|\boldsymbol{\theta}^1) - f_w^{Tri(j)}(wait_j(t), \boldsymbol{X_j}|\boldsymbol{\theta}^2)| \leq BK\delta = BK\|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2\|. \tag{32}$$

That implies that the function $f_w^{Tri(j)}(\cdot|\hat{\boldsymbol{\theta}})$ is (Lipschitz) continuous in $\hat{\boldsymbol{\theta}}$ and condition (iii) gets proved.

Finally, to prove condition (iv), we first prove that $Q^n(\hat{\boldsymbol{\theta}}) \to Q(\hat{\boldsymbol{\theta}})$ for all $\hat{\boldsymbol{\theta}}$, and then prove that the convergence is uniform. We define the empirical distribution π_n for a sequence of n choice incidents as

$$\pi_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1(\boldsymbol{\Sigma}(t) \in A), \text{ for all } A \subseteq \overline{\Omega} \tag{33}$$

Equation (21) in (a1) implies that π_n weakly converges to π. Therefore,

$$\begin{aligned} Q^n(\hat{\boldsymbol{\theta}}) &= \tfrac{1}{n}\textstyle\sum_{t=1}^{n} \ln P(c(t)|\boldsymbol{\Sigma}(t), \hat{\boldsymbol{\theta}}) \\ &= \tfrac{1}{n}\textstyle\sum_{\{A\}} \text{ as a partition of } \overline{\Omega} \sum_i 1(\boldsymbol{\Sigma}(t) \in A,\ i \in ChoiceSet(t)) \ln P(i|\boldsymbol{\Sigma}(t), \hat{\boldsymbol{\theta}}) \\ &\xrightarrow{p} \mathbb{E}_{\boldsymbol{\Sigma}\sim\boldsymbol{\pi}}\mathbb{E}_c \ln P(c|\boldsymbol{\Sigma}, \hat{\boldsymbol{\theta}}) \\ &= Q(\hat{\boldsymbol{\theta}}). \end{aligned} \tag{34}$$

where the convergence follows from π_n weakly converges to π and the strong law of large number.

It remains to show that the above convergence is uniform. In view of Lemma 2.9 of (Newey and McFadden 1994), it suffices to show that $Q^n(\hat{\boldsymbol{\theta}})$ is Lipschitz continuous in $\hat{\boldsymbol{\theta}}$. Since the multinomial-type log-likelihood function $\ln \frac{\exp(z_i)}{\sum_k \exp(z_k)}$ is Lipschitz continuous in each z_k when all z_ks are bounded, we deduce that $Q^n(\hat{\boldsymbol{\theta}})$ is Lipschitz continuous in each $z_k = f_c(\boldsymbol{X_k}|\hat{\boldsymbol{\theta}}) + f_w^{Tri}(wait_k(t)|\hat{\boldsymbol{\theta}})$. Because $f_c(\boldsymbol{X_k}|\hat{\boldsymbol{\theta}})$ is affine and therefore must be Lipschitz continuous in $\hat{\boldsymbol{\theta}}$, and $f_w^{Tri}(wait_k(t)|\hat{\boldsymbol{\theta}})$ is Lipschitz continuous due to Equation (32), we deduce that $Q^n(\hat{\boldsymbol{\theta}})$ is Lipschitz continuous in $\hat{\boldsymbol{\theta}}$. ■

Discussion of Assumptions The assumptions (a1)-(a6) are all needed to show that the MLE $\boldsymbol{\theta}$ is identifiable. We next provide more details about those assumptions. The ergodic condition mentioned in (a1) is a weaker assumption than the commonly-used iid condition. We cannot assume $(c(t), \boldsymbol{\Sigma}(t))_{t=1,2,\ldots,}$ to be iid in our model, because the choice sets are actually not independent – patients in the current choice set are likely to appear in the next one (if not chosen). Also, the sequence is not stationary, as the distribution of patient attributes may vary due to the hour-of-day or day-of-week effect. Instead, in (a1) we assume that the sequence $\{\boldsymbol{\Sigma}(t)|t = 1, 2, \ldots\}$ is a positively recurrent and periodically stationary Markovian process, which leads to ergodicity. This assumption is based on the following logic. First, the waitlist in ED can be modeled as a $G_t\backslash G\backslash 1$ queue with time-varying arrival rates. The number of patients in each choice set is an embedded discrete-time Markov chain at the departure epochs of this queueing process. We allow the distribution of the fixed attributes $\boldsymbol{X_j}$ of arrived patients to vary with time, but has to be periodically stationary. Therefore, given $\boldsymbol{\Sigma}(t)$, the next state $\boldsymbol{\Sigma}(t+1)$ depends on the random number of the arrived patients between two successive departure (choice) epochs, the random draw of the new arrived patients' attributes, and the selection outcome of the last choice incidence, $c(t)$. Thus, $\boldsymbol{\Sigma(t)}$ is a Markovian process with time-varying transition probabilities. Second, we can assume that the arrival process is periodically stationary. This is supported by the Pearson's Chi-square test, which shows that the dependence of the main patient attributes on week is statistically insignificant (p-value > 0.05). Consequently, the sample paths of the Markov process $\boldsymbol{\Sigma}(t)$ have the same distribution in each week, which implies periodic stationarity. Finally, the positive recurrence of $\{\boldsymbol{\Sigma}(t)\}_{t=1,2,\ldots}$ follows from that the average arrival rate is smaller than the average service capacity in the long run, so that the queue length will not grow to infinity. This is certainly a reasonable assumption in the ED setting.

(a2) requires the parameter to be contained in a compact domain. This assumption is necessary because otherwise the MLE may not exist. To illustrate that, we consider a simplified model which has no fixed patient attributes, only one triage class, $\mu_j = 1$, and $f_w^{Tri}(wait_j(t)) = \beta_1 wait_j(t)$. Thus, $V_{jt}(wait_j(t), \boldsymbol{X_j}) = \beta_1 wait_j(t)$. If we observe that patients are FCFS in all choice sets, then the likelihood function will be maximized at $\beta_1 = +\infty$ so that the contribution of the stochastic term ϵ to $V_{jt}(wait_j(t), \boldsymbol{X_j})$ is minimized. This example shows that the MLE may not exists if we have not required $\boldsymbol{\Theta}$ to be compact.

(a3) is a standard assumption for the coefficients $\boldsymbol{\alpha}$ to be identified. (a4) is required to show boundedness and Lipschitz continuity of $f_w^{Tri}(\cdot|\hat{\boldsymbol{\theta}})$ with respect to $\hat{\boldsymbol{\theta}}$, which are used to prove uniform convergence of $Q^n(\hat{\boldsymbol{\theta}})$ to $Q(\hat{\boldsymbol{\theta}})$. (a5) is necessary for all parameters to be identifiable. Without (a5), one may consider an

example in which each choice set contains either all males or all females. Then the effect of gender on the choice probability can never be identified, because we have not observed any competition between different genders. To see the necessity of (a6), one may consider the case when the probability of the waiting times are zero over certain intervals for some patients, then we cannot fit the piece-wise polynomial during those intervals so $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ cannot be identified.

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.

Parameter Choice of D and B: At the end, we comment on how to identify the parameters D and B. The integers D and B give an upper bound for the highest degree and number of break points for the polynomial regression splines we use to fit $f_w^{Tri}(\cdot)$. From a theoretical perspective, we can always set D and B to be a sufficiently large value, so that the MLE returns a piece-wise polynomial with the best fit, which must be unique by Theorem H.1. The maximum degree and number of break points for that piece-wise polynomial gives the optimal parameters $D^* \in [0, D]$ and $B^* \in [0, B]$. By sending D and B to infinity, the optimal polynomial regression splines can always be identified in theory. However, in reality, the complexity of computing the MLE increases quickly with D and B. For our study data, in order to complete the computation in a reasonable time scale, we have to test several combinations with either a smaller B (such as the cubic model) or a small D (such as the piece-wise linear model) but not both, and then choose the combination with the largest log-likelihood.

Appendix I: 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, criticize that the fitted probability does not use the entire message that the model attempts to deliver. This limitation is shown in the following example with a choice set of three patients $\{1,2,3\}$. The choice probabilities predicted by Model A and Model B are, respectively, $\{0.5,0.4,0.1\}$ and $\{0.5,0.1,0.4\}$; while in the data patient 2 was chosen. Then the two models perform equally poorly according to the fitted probability as both models have chosen patient 1 instead of patient 2. However, the fitted probability does not take into account that Model A has assigned a much higher probability to the correct patient than Model B, and thus should deserve a better score. Another limitation of using the fitted probability is its dependence on the sizes of the choice sets (Li 2002). For this reason, one cannot use fitted probability to compare predictions made for difference data sets.

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)[11]. 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. To further illustrate that, consider the previous three-patient example. For all ηs between $0.1-0.4$, Model A will mark both patients 1 and 2 as selected, while Model B will select both patients 1 and 3 but still miss the correct pick. Therefore, Model A has a higher true positive rate as well as a lower false positive rate for those ηs. In other words, the AUROC metric successfully captures the advantage of Model A over Model B.

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[12]. The comparison is summarized in Table 8. 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.

The pseudo R^2 values reported in Table 8 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

Table 8 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.

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 8, the prediction accuracy of the piece-wise linear model is between good and fair, which supports the effectiveness of our framework.