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

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

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

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

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

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

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

Appendix C: Decision Maker Heterogeneity

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

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

Appendix D: Unobserved Patient Heterogeneity

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

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

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

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

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

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

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

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

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

Standard errors in parentheses.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

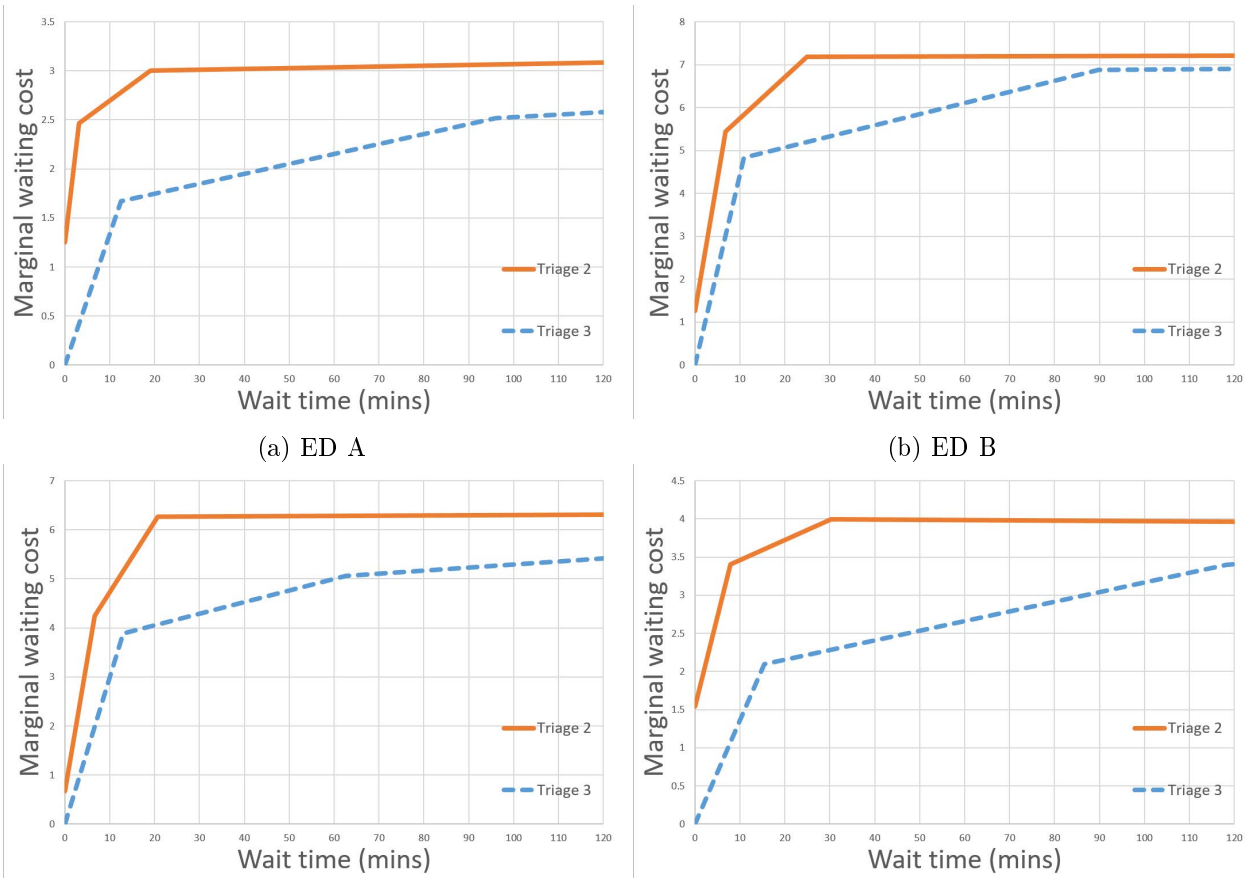
To test whether the IIA property is a reasonable assumption for the observed data, we investigate the mixed logit model, which has been discussed in Appendix C. The conditional logit model used in this paper is a special case of the mixed logit model when the random slopes and intercepts of the piece-wise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

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

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

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

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



(a) ED A (b) ED B

(c) ED C (d) ED D

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

Then when $n\to\infty$,

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

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

Appendix H: Out-of-Sample Test

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

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

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

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

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

Table 7 Out-of-Sample Test Statistics

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

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

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

References

Antoniadis A, Gijbels I, Nikolova M (2011) Penalized likelihood regression for generalized linear models with non-quadratic penalties. *Ann. Inst. Stat. Math.* 63(3):585–615.

Dierckx P (1995) *Curve and Surface Fitting with Splines* (Oxford University Press).

Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.* 27(8):861–874.

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

Li T (2002) Econometric analysis of cross section and panel data (book). *J. Econom. Lit.* 40(4):1239–1241.

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

Lowsky D, Ding Y, Lee D, McCulloch C, Ross L, Thistlethwaite J, Zenios S (2013) A k-nearest neighbors survival probability prediction method. *Statist. Medicine* 32(12):2062–2069.

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

McDonald JH (2009) *Handbook of Biological Statistics*, vol. 2 (Sparky House Publishing, Baltimore).

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

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

Newey WK, McFadden D (1994) Large sample estimation and hypothesis testing. *Handbook of Econometrics*, vol. 4, 2111–2245.

Pardoe I, Simonton DK (2008) Applying discrete choice models to predict Academy Award winners. *J. Royal Statist. Soc. Ser. A* 171(2):375–394.

Pines JM., Carpenter CR, Raja AS, Schuur JD (2012) The epidemiology and statistics of diagnostic testing. *Evidence-Based Emergency Care: Diagnostic Testing and Clinical Decision Rules*, 2nd ed., 19–35.

Ruppert D, Carroll RJ (1999) Penalized regression splines. Technical Report, Cornell University Operations Research and Industrial Engineering.

Sung M-C, McDonald DC, Johnson JE (2016) Probabilistic forecasting with discrete choice models: Evaluating predictions with pseudo-coefficients of determination. *Eur. J. Oper. Res.* 248(3):1021–1030.

Tape TG The area under an AUC curve. University of Nebraska Medical Center, http://gim.unmc.edu/dxtests/Default.htm.

Train K (2000) *Halton Sequences for Mixed Logit* (Department of Economics, UCB).

Train KE (2009) *Discrete Choice Methods with Simulation* (Cambridge University Press).