THE UNIVERSITY OF CHICAGO

MACHINE LEARNING FOR QUEUE PRIORITIZATION: APPLICATIONS TO THE EMERGENCY DEPARTMENT

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
GIZEM YILMAZ

CHICAGO, ILLINOIS
AUGUST 2022

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to express my deepest gratitude to my advisor and chair of my committee Prof. Dan Adelman for his enormous support and guidance during my academic journey at Chicago Booth. Thanks to him for being the most understanding advisor, allowing me to express my opinions freely, and valuing me. Without him, this endeavor would not be possible.

I am extremely grateful to my committee members for supporting my dissertation. Dr. Thomas Spiegel was always generous in sharing his professional experience and knowledge in emergency department operations. Prof. Varun Gupta has been a great mentor since the beginning of my Ph.D. journey. Prof. Rad Niazadeh provided me with his expertise in machine learning. I sincerely thank Kanix Wang for being at every step of this dissertation and Kate Haas LaVigne for assisting me with the technicalities.

I am also grateful to Chicago Booth, Ph.D. Office for making my Ph.D. journey as smooth as possible. I cannot express my gratitude to Cynthia Hillman for supporting me at every step of the job search process and motivating me whenever I needed it. Thanks to Kimberly Mayer for organizing the best coffee hours and ordering the delicious sweets that improve my day. Thanks to Amity James for handling all the paperwork that allows me to be here. Major thanks to Malaina Brown, who was ready to listen whenever I needed advice at the hardest times of my academic journey at Chicago Booth.

At Chicago Booth, I had the pleasure of being friends with Ali Cem Randa, Cagla Keceli, Charlie Hannigan, Deniz Akturk, and Zeynep Kahveci, of whom I have fond memories. They have been my rays of sunshine throughout this journey. I would be remiss in not mentioning Zuguang Gao, Nasser Barjesteh, and my beloved roommate Gulin Tuzcuoglu. Thanks to them for not leaving me alone during the pandemic.

I would like to express my special thanks to my dearest friend and biggest supporter, Amir Alwan. He has been there for me whenever I felt down and cherished my happy moments. Words cannot express my gratitude for his friendship.

Last but not least, I am thankful that I have a loving and supporting family that I can always count on. Major thanks to my mom for just being the best mom for me and for all the sacrifices she made to raise me.

This project was formally determined to be a quality improvement project, not human

subjects research, and was therefore not overseen by the Institutional Review Board, per institutional policy.

ABSTRACT

Queue prioritization is a common practice that allocates limited resources to heterogeneous customers to improve operational outcomes and customer satisfaction in service systems such as call centers and emergency departments. Ideally, the decision-maker has perfect knowledge of customer priority and allocates the resources to the customer with the highest priority. In practice, the decision-maker makes an educated guess regarding customer priority based on the available information and might make a suboptimal decision.

The first part of the dissertation studies a queue prioritization problem with two homogeneous customer segments. The service provider does not know the customer type, i.e., from which segment the customer is, upon a customer's arrival. Instead, they utilize a binary classification model to estimate the probability of being a high-importance customer. The customer is assigned to the high priority group if the likelihood is above a certain threshold. The ROC curve shows the performance of the binary classification algorithm in terms of sensitivity and specificity at various thresholds. Changing the threshold usually impacts the classification algorithm's sensitivity and specificity in opposite directions, i.e., there is a trade-off between sensitivity and specificity. The traditional threshold selection method tends to optimize a ROC curve-based metric and does not consider the operational externalities. This dissertation analyzes the optimal threshold policy in terms of the ROC curve, i.e., sensitivity and specificity, by considering the operational nature of the service systems.

Chapter 2 studies a non-preemptive M/G/1 queueing system with two customer types under imperfect information. If the customer is assigned to the high priority group, they are provided with priority service faster and not too much more variable than non-priority service. We model imperfect information by approximating the ROC curve with an increasing, concave, piecewise linear function and show that the optimal threshold policy is unique. It is optimal to trade off a loss in specificity for a higher gain in sensitivity.

The second part of the dissertation is an empirical study on queue prioritization where customer priority depends on other customers' characteristics and the system status, focusing on application to emergency departments (EDs). We model the patient prioritization problem using the discrete choice framework, particularly the conditional logit model (CLM).

Chapter 3 and Chapter 4 explore how the ED system attributes affect the patient pri-

oritization rule by utilizing a tree-based segmentation algorithm that generates ED system clusters where a similar patient prioritization rule is observed. We first test the performance of this approach on an artificial dataset generated by simulating an ED that always uses the optimal deterministic policy. Since the decisions in the simulated environment are deterministic, any plausible algorithm performs well in this artificial dataset. The tree-based algorithm can predict the bed assignment decisions with almost perfect accuracy.

Chapter 4 applies the tree-based algorithm to a dataset that includes patient encounters at the Emergency Department at the University of Chicago Medicine (UCM). According to the algorithm, room type, waiting room census, and time of the day are the most important system-level attributes; ESI score and waiting time are the most important patient-level attributes for patient prioritization. High acuity patients are prioritized for the main service area that includes most of the rooms, while low acuity patients are prioritized for the fast track area in the ED. The First-Come-First-Served principle is generally followed within the same urgency class. As the waiting room becomes more crowded and resource utilization increases, the adherence to urgency-based prioritization increases.

The discrete choice framework assumes that the choice options come from a finite set. However, the emergency department dataset, by its nature, does not have a finite set of choice options because each patient is unique. In Chapter 5, we develop a tree-based algorithm that segments patients into a finite number of clusters, leaf nodes, based on their attributes and incorporates the leaf membership as alternative specific constants into the model to capture the alternative's inherent characteristics and interaction effects.

Finally, we summarize our results and discuss future directions in Chapter 6.

CHAPTER 1
INTRODUCTION

Queue prioritization is common in many service systems where limited resources serve heterogeneous customers, such as call centers and emergency departments. Customer service departments of many companies move the high-value customers in front of the line and assign them to the most skilled agents. In emergency departments, patients are triaged upon arrival to provide timely medical care to those who need urgent care. Queue prioritization improves the operational outcomes and customer/patient satisfaction by directing limited resources to the highest priority customers. Ideally, the service provider would perfectly observe whether an arriving customer has a priority condition and would give preferential treatment to them. In practice, the service provider has imperfect information on customers and must deploy a queue prioritization based on the available information.

There is a vast literature on queue prioritization, particularly under the perfect information, in operations. It has been proven that the optimal decision rule for a multiclass queuing system is to serve the customer class with the highest $c\mu$ index. The implicit assumption behind the analysis of such queuing systems is that customer priority is inherent, i.e., information on other customers' characteristics and resource limitations do not impact the customer's perceived importance. This definition of priority is reasonable for situations where (i) customers can be segmented so that those within the same segment are similar enough and (ii) changes in the system-level characteristics are irrelevant from an operational perspective. Then, the queue prioritization problem boils down to predicting the inherent customer type in these situations.

Many service systems frequently use classification algorithms to predict customer type. In business applications, customer type may refer to the answer to a yes-no question. Is this caller a target customer? Will this customer churn? Will this emergency department patient be admitted to the inpatient stay? Will this patient be accepted to an intensive care unit? A binary classification algorithm can determine the likelihood of a "yes" outcome. The algorithm predicts a yes to the binary question if this likelihood exceeds a predetermined threshold. The threshold selection is an integral part of the prediction. The common practice

in the classification literature is to select a threshold that optimizes a non-operational metric without acknowledging the operational externalities that customers have on each other. In practice, resources are limited and shared across customers. A customer's utilization of a scarce resource creates operational externalities on others. Queuing theory provides a useful framework for studying resource allocation problems with heterogeneous customers.

The queuing literature assumes that the service rate depends on the inherent customer type rather than the classification. Thus, the classification algorithm impacts only the customer scheduling, i.e., the order that customers are processed. In some service systems, the service rate is also determined by the classification outcome. In call centers, "important calls" are directed to the most skilled agents who resolve calls more efficiently than their peers. In a pilot program conducted at the Emergency Department of the University of Chicago Medicine (UCM), the treatment process is accelerated for patients predicted to be admitted to an inpatient stay. In these systems, the classification algorithm impacts not only customer scheduling but also service time. Therefore, it becomes more crucial for a binary classification algorithm to acknowledge the operational externalities in the optimal threshold selection. Chapter 2 incorporates the operational nature of the problem by modeling the service system as an M/G/1 multiclass queuing system. We also diverge from a common assumption in the queuing literature that service rates depend on the inherent customer type. To our best knowledge, we are the first to study the optimal threshold selection problem in a multiclass queuing system where service rates depend on the customer type prediction rather than the true type.

So far, we have defined priority as an inherent characteristic of the customer. The inherent priority definition is reasonable when the customers can be segmented into homogeneous classes, and the decision environment is stable or irrelevant to the service provider. What if the priority is relative rather than inherent, i.e., depends on other customers' characteristics and the system status? In the emergency department, patient prioritization is necessary since the patients drastically differ from each other, and the emergency department resources, specifically ED beds, are limited. Patients are "triaged," in which they are sorted into five different urgency categories to help with patient prioritization. However, this categorization is only a proxy for the perceived urgency of the patient. Other factors may go into patient

prioritization, such as the available resources. For example, suppose a patient requires a specific resource that is unavailable at the time. In that case, another patient with a less urgent condition can be prioritized over the former patient.

Chapter 3 utilizes a machine learning algorithm that can be deployed to support patient routing decisions for ED bed allocation and drive insights into emergency department operations. Given that the patient priority is relative to other patients and depends on the status of the ED, we model the patient prioritization problem using the discrete choice framework, in particular the conditional logit model (CLM). A decision epoch corresponds with the times that a decision is made to allocate an ED bed to a patient, and the choice set is patients waiting to be assigned to a bed. The CLM allows us to calculate choice probabilities for each patient in the choice set. These probabilities reflect the relative priority of each patient in the waiting room during decision-making. The status of the ED changes over the day and affects the queue prioritization rule. For example, the fast-track at the Emergency Department of UCM is usually used to treat low acuity patients and is not open between 6 am and 12 pm. During this period, the waiting time becomes more important, and triage levels become less important to determine the relative priority of patients for an ED bed compared to the rest of the day.

To understand how the ED status, as given by the number of boarders, waiting room census, number of available beds, etc., affects the patient prioritization rule, we utilize *Marketing Segmentation Trees* (MST), a tree-based segmentation algorithm recommended by Aouad et al. [2019]. The MST algorithm allows us to segment attributes that define an ED status to clusters where a similar patient prioritization rule is observed. Once we segment the ED status into clusters, we implement a CLM model at each cluster to predict the patient prioritization decisions. We test the performance of this approach on an artificial dataset before applying it to the UCM ED dataset. Chapter 3 also introduces a Markov Decision Process framework to model emergency departments at a high level and numerically obtains the optimal policy. Then we generate an artificial dataset by simulating an ED that always makes decisions based on the optimal policy. Since the simulated environment decisions are always deterministic, any plausible algorithm performs well in this artificial dataset. The MST algorithm and conditional logit model with all pairwise interaction terms can predict

the bed assignment decisions with almost perfect accuracy.

Chapter 4 describes the UCM ED dataset and applies the MST algorithm to the dataset to obtain insights on patient prioritization rule for a bed assignment in the ED. The algorithm discovers that room type, waiting room census, and time of the day are the most important factors impacting patient routing decisions. The prioritization rule at each cluster mainly depends on the ESI score and the waiting time. For the urgency room, high-acuity patients are prioritized, while low-acuity patients are prioritized for the non-urgency room. As the waiting room becomes more crowded and resource utilization increases, it becomes more likely that a patient with a higher level of urgency will be selected than a patient with a lower level of urgency.

The standard CLM framework is usually applied when the decision-maker chooses from a choice set that can change from decision epoch to decision epoch, but its elements always come from a predefined set. For example, the CLM model is extensively used in marketing and transportation literature. The mode options are usually limited to bicycle, automobile, train, bus, walking, carpool, and two-wheeler in transportation choice. In assortment optimization, the choice set includes a finite number of brands. The convenience of a finite choice set is that we can incorporate alternative specific constants in the utility model. The alternative specific constant terms can improve the model performance by capturing the alternative's inherent characteristics and attribute interactions, which are not captured by the other terms in the utility function.

In the emergency department, the choice set, by its nature, comprises patients who drastically change from each other, i.e., there is no finite set that includes all possible alternatives in the patient selection problem in the emergency departments. Even the same patient does not appear as the same alternative during a decision horizon due to the time-sensitive nature of health and, trivially, waiting time. We attempt to address this issue in Chapter 5 by developing a tree-based algorithm that segments patients into finite number groups based on their attributes and incorporates the group membership as alternative specific constants in the utility function. The patient segmentation and model fit are simultaneously performed rather than sequentially.

CHAPTER 2
QUEUE PRIORITIZATION WITH CLASSIFICATION DEPENDENT SERVICE RATES UNDER IMPERFECT INFORMATION

2.1 Introduction

Customers differ in their delay sensitivity and financial contribution in many service systems. Given the customer heterogeneity, serving customers in order of their arrival is not always the best strategy. In practice, many service systems segment their customers into different priority classes. Often, prioritized customers have a separate line that allows them to jump over non-prioritized customers to receive faster service. Optimal prioritization policies for multiclass queuing systems are extensively studied under perfect information where a job's true type can be immediately observed upon arrival. It has been proven that the "$c\mu$-rule" is the optimal policy for many such scenarios. However, the decision-maker does not always observe the job's true type.

With the growing popularity of artificial intelligence, many decision support models are being developed to predict customer types. Often, these classification algorithms first predict the likelihood of each class based on the available information and then use these estimated likelihoods as the basis for classification. For binary classification tasks with two classes, one class is referred to as the positive class, and the other class is referred to as the negative class. For this paper, being positive refers to having a priority condition, such as being a target customer or a patient likely to be admitted to the hospital; negative refers to not having a priority condition. If a customer's likelihood of having a priority condition is above a predetermined threshold, the customer is assigned to the positive class; otherwise, they are assigned to the negative class.

The performance of a binary classifier can be visualized by a confusion matrix that formulates the four outcomes of the classification process: true positive, true negative, false positive, and false negative. It is common to characterize the classification performance in terms of sensitivity and specificity in practice. Sensitivity refers to the true positive rate,

i.e., the percentage of correctly identified positives among all true positives. On the other hand, specificity refers to the true negative rate, i.e., the percentage of correctly identified negatives among all true negatives. Increasing the threshold value lets the classifier assign more customers to the negative class, increasing the number of correctly identified negatives, thereby increasing the specificity. However, assigning more customers to the negative class can also increase the number of false negatives, i.e., decrease the number of true positives, thus decreasing the sensitivity. A receiver operating characteristic (ROC) curve plotted based on the training data evaluates the performance of the binary classifier at different thresholds. On the ROC curve, the true positive rate (sensitivity) is displayed on the vertical axis, and the false positive rate (1-specificity) is displayed on the horizontal axis (see James et al. [2014] for details).

The threshold selection is an integral part of a binary classification algorithm. The common practice in the classification literature is to select a threshold that optimizes a non-operational metric derived from the confusion matrix, such as Youden's index, i.e., sensitivity + specificity - 1. There are also cost-based approaches that determine the optimal threshold value based on the cost analysis of the four possible outcomes of the classification process. However, these approaches assume that costs are known and independent of the threshold value and that a customer's classification outcome does not have an externality on other customers.

In many service systems, resources are limited. The utilization of shared resources by a customer creates operational externalities for other customers. For example, the classification of false priority customers can exhaust the resources allocated for true priority customers and thus potentially increase their waiting time by creating additional congestion. Despite the real-life prevalence of multiclass service systems with imperfect information, there is scant literature on improving the job flow in such systems. Argon and Ziya (Argon and Ziya [2009]) study an M/G/1 queuing system where customers can be one of two possible types. They assume that service time and waiting cost depend on the true customer type. The $c\mu$-rule determines the prioritized type. Customer types are unknown upon arrival, but each customer arrives with an imperfect signal corresponding to the probability of being the prioritized type. They propose a threshold policy that prioritizes a customer if their signal

exceeds a certain threshold. Their main interest is to optimize the threshold to minimize the expected waiting cost.

In some service systems, the classification determines the service time rather than the true customer type. For example, The Emergency Department at the University of Chicago Medicine (UCM) designed the "Dr. Admit" program to identify patients likely to be admitted to the hospital early upon their arrival. The program strives to reduce the time to admission - defined as the time from arrival to the time that an admission order is made - by partnering and coordinating with other service lines to prioritize "Dr. Admit" patients' care needs. As another example, in call centers, agents differ in their skill level and experience. The most skilled agents can resolve calls more effectively and faster than the less skilled agents. Not all calls can be assigned to highly skilled agents because there may not be a sufficient number of highly skilled agents to answer all calls within a reasonable wait time. Moreover, call centers may want to keep the workload fair between agents. Due to the limited number of highly skilled agents, only calls deemed more critical can get routed to a highly skilled agent and receive faster service than less critical calls.

We model a service system as a non-preemptive M/G/1 queue with two customer types: type-A and type-B. Type-A customers have higher delay sensitivity than type-B customers and thus should be labeled as a high priority. However, the decision-maker does not have perfect customer information. Rather, they assign the customer to either high priority or low priority class based on the available information upon their arrival. The service rate depends on the customer classification outcome rather than their type. Thus, customers predicted as high priority receive better service in terms of service time. This is the first study analyzing service systems with classification-dependent service rates under imperfect information, to the best of our knowledge. In other works, service rates depend on the true customer types; thus, changing the classification affects the customer service processing order but does not affect the system workload, i.e., the ratio of the arrival rate to the service rate. In contrast, we assume that service rates depend on the classification outcome of the customer type; thus, changing the classification affects both the customer service order and server utilization by each customer type.

We first study service systems where customer types are perfectly observed and show

that misclassification of some type-B customers into the priority class might decrease average waiting costs by allowing these customers to receive faster service, thus decreasing the waiting time of other type-B customers who have not been served yet. This might increase the waiting time of true type-A customers. However, if the positive externality on the type-B customers can compensate for the increase in the waiting time of true type-A customers, average waiting costs decrease. It might even be possible that some misclassification of type-B customers can decrease the true type-A customers' waiting time. This result may seem surprising at first, but it makes intuitive sense due to the non-preemptive nature of the queuing system. Non-preemption implies that a priority customer cannot interrupt the service of a non-priority customer and has to wait until their service is completed. This expected remaining service time of the customer is called residual time. Intuitively, if more customers are assigned to a priority class, it is more likely that a given customer in service is being served at the priority speed. If the priority service time variability is not much higher than that in the non-priority service time, the expected remaining time in service would be shorter, positively impacting both type-B and type-A customers' waiting times. Depending on the system parameters, this positive impact on type-A customers' waiting time can dominate the negative externality of the additional congestion in the priority line.

A system with imperfect information calculates the likelihood of being type-A according to a machine learning algorithm with an already established ROC curve. We aim to find an optimal threshold policy, i.e., an optimal sensitivity and specificity trade-off, to minimize patients' waiting costs under imperfect information. We assume that waiting costs are linear in time. By approximating the ROC curve as an increasing, concave, and piecewise linear function, we show that the optimal threshold policy is unique and always trades off a lower loss in specificity for a higher gain in sensitivity. In other words, moving right on the ROC curve is optimal as long as the marginal gain in sensitivity is higher than the marginal loss in specificity. Then we approximate the ROC curve using a simple piecewise linear function with a single breakpoint to obtain analytic results on the expected value of the perfect information (EVPI).

The rest of this chapter is organized as follows. Section 2.2 reviews common practices for threshold selection in binary classification and the relevant literature in priority queues. In

Section 2.3, we introduce notation and the queuing model. In Section 2.4, we solve the model for the perfect information case and show that some misclassification of type-B customers could decrease average waiting costs. Section 2.5 introduces the ROC curve approximation to model the imperfect information case and shed light on the optimal sensitivity and specificity trade-off. Section 2.6 lets the ROC curve take a more simplistic form and discusses how the system parameters affect the EVPI. Section 2.7 compares our model to the M/G/1 queuing system with type-dependent service rates (see Argon and Ziya [2009]) and emphasizes the resulting differences in the optimal policy. Section 2.8 allows waiting costs to be nonlinear in time and discusses how nonlinearity changes the results.

2.2 Literature Review

Binary Classification: Many methods for threshold selection focus on balancing the trade-off between sensitivity and specificity. One of the earliest approaches in the threshold selection is to maximize the Youden index (see Youden [1950]), which measures the difference between the true positive rate and false positive rate, i.e., sensitivity + specificity - 1. This approach is commonly used in the literature and practice, particularly in clinical works (see Parikh and Philbrook [2011] and Martínez-Camblor and Pardo-Fernández [2019]). Another approach is to choose the point on the ROC curve that minimizes the Euclidean distance to the $(0, 1)$ point. A third approach is to choose the point that maximizes the product of sensitivity and specificity. The reader can refer to Unal [2017] for other methods to determine the optimal threshold value. There are more elaborate models (see Habibzadeh et al. [2016], Rücker and Schumacher [2010], and Li et al. [2018]) that take into account the misclassification costs and prevalence of classes. These methods are suitable when the classification outcome of an instance has no impact on other instances, allowing misclassification costs to be calculated without considering any externalities.

Priority Queues: For multiclass priority queues in various settings, the optimal policy is proved to be the $c\mu$ rule where the job with the highest $c_i\mu_i$ should be first served, where c_i is the cost of holding job i in the system and μ_i is the service rate of job i. Refer to Van Mieghem [1995] for a detailed review. The main assumption in these papers is that job

type is known. Our work assumes that the decision-maker does not know the job type but can predict the type likelihoods upon arrival.

There has been work on priority queues with imperfect information; however, their number is relatively limited. The first related work to our knowledge is by Van der Zee and Theil [1961]. This work analyzes a single server system with two customer types: type-1 and type-2. Type-1 customers require a shorter service time than type-2 customers. Customer types are unknown upon arrival, but each customer arrives with a signal that indicates the probability of being the prioritized type. The decision-maker can assign customers either type-1 or type-2 based on their signal, but the classification is subject to error. The decision-maker either serves customers in an FCFS manner ignoring the group signaling or prioritizes one group over the other and then serves according to the FCFS principle. First, they determine under what condition prioritization performs better than the FCFS policy to reduce the expected waiting time. Secondly, they recommend a three-way classification policy where the customer is labeled as type-1 (type-2) if her signal is "sufficiently" high (low). Otherwise, she is assigned to a mixed group prioritized over customers labeled as type-2 but served after customers labeled as type-1. For the three-way classification, they determine the optimal threshold values for prioritization. In the paper, the misclassification rates are fixed regardless of the threshold values used for classification. In reality, misclassification rates, the rate of falsely identified type-1 and type-2 customers depend on the threshold. Varying the threshold will increase the misclassified customers for one group while decreasing the misclassification for the other.

More recent work on priority assignment problems under imperfect information is by Argon and Ziya [2009]. They model the misclassification rates explicitly as a function of the classification process. They analyze an M/G/1 with two customer types where the customer arrives with a signal that indicates their true type. They propose a threshold policy where the customer is assigned to the high-priority class if their signal is above a certain threshold. The paper's main interest is to optimize the threshold to minimize the expected waiting cost. Our paper differs from Argon and Ziya [2009] in two ways: (i) they assume that service rate depends on the true customer type while we assume that the service rate depends on the classification; (ii) their optimal threshold is based on a signal distribution while we directly

work with the ROC curve, which can be built based on the signal distribution.

Dobson and Sainathan [2011] consider a service system with heterogeneous customers whose types are unknown upon arrival. There are sorters that categorize the customers into priority classes. The sorting process is not perfect and consumes both money and time. The authors evaluate under what conditions the sorting and prioritization are beneficial for reducing waiting and total costs.

Sun et al. [2018] study the dynamic triage and prioritization problem with two customer types where triage consumes time and resources are highly restricted. They characterize the optimal policy that decides when to triage and how to prioritize patients.

Alizamir et al. [2013] consider a service provider that identifies customer types by running imperfect diagnostic tests. An additional test improves the diagnostic accuracy but also increases congestion and causes delayed services for others. The authors study the problem of dynamic balancing this accuracy and congestion trade-off.

Saghafian et al. [2012] study the performance of patient streaming, segregation of ED beds, and care resources based on patient disposition predictions. They find that virtual streaming, where resources are shared across streams rather than physically separated, is more effective than the traditional pooling policy in situations where a high fraction of patients are admitted to an inpatient stay in the hospital. Their analytical results assume perfect prediction accuracy, but their numerical study takes the misclassification error into account and finds that patient streaming is preferred to the pooling if the misclassification rate is not high.

Saghafian et al. [2014] recommend a new triage system that takes patients' urgency and complexity information into account in patient routing decisions. A medical provider imperfectly classifies a patient's urgency and complexity type in their setting. They assume that misclassification rates are given and evaluate the performance of the complexity-augmented triage.

Singh et al. [2020] propose digital triage in healthcare where medical images are classified into priority queues (rather than clinical types such as diseases) based on their features. The digital triage aims to minimize the average waiting costs by capturing the interactions between the classification errors and queuing externalities.

2.3 Model

We study a non-preemptive M/G/1 queue with two priority types: $\{A, B\}$. Type-A customers represent a fraction α of the customer population and are more "important" than type-B customers for the decision-maker; thus, they receive priority. However, upon arrival, customer type is unknown; a binary classification algorithm, previously determined by the decision-maker, predicts the customer type. The classification algorithm calculates t, the likelihood of being type-A based on available information upon arrival. If the likelihood t is above a certain threshold $\bar{t}$, then the algorithm classifies the customer as type-A; otherwise, as type-B. The classification is not perfect, and the performance of the classification algorithm at various thresholds is illustrated by the ROC curve $(x, g(x))$, where x is the false positive rate and $g(x)$ is the true positive rate.

Once the threshold is chosen, the decision-maker categorizes customers into two classes: priority and non-priority. We denote the priority class by 1 and non-priority class by 2. Note that notation A, B represents the true customer type, and notation $1, 2$ represents the classification outcome. A priority, class-1 customer, can be a correctly classified type-A or a misclassified type-B customer. Similarly, a non-priority, class-2 customer can be a correctly classified type-B or misclassified type-A customer.

Customers incur waiting costs linear in time. We let γh (h) denote a type-A (a type-B) customer's waiting cost rate. Since a type-A customer is more important than a type-B customer, their waiting cost rate is higher, i.e., $\gamma > 1$. We want to find the optimal threshold to minimize the long-run average waiting cost. We can assume that $h = 1$ without loss of generality.

We assume that service rate is classification dependent, and thus we use the notation $1, 2$ to label the parameters related to the service time. The service time for priority customers (non-priority customers) has mean $\frac{1}{\beta\mu}$ $\left(\frac{1}{\mu}\right)$ and second moment e_1 (e_2). Priority customers are treated faster than non-priority customers, i.e., $\beta > 1$. Priority customers are also allowed to jump over non-priority customers in the queue. We assume $e_1 \leq e_2$, which implies that the variability in priority service time is not "too much" higher than the variability in non-priority service time. The variability in service distribution can be measured by the standard

coefficient of variation (SCV), a metric given by the ratio of variance to mean squared. Then we can express the second moment of any distribution in terms of the mean squared and the SCV by substituting the variance with the SCV times mean squared:

$$\begin{aligned} e_1 &= \frac{1 + \mathrm{SCV}_1}{\beta^2 \mu^2}, \\ e_2 &= \frac{1 + \mathrm{SCV}_2}{\mu^2}. \end{aligned} \tag{2.1}$$

By (2.1) and $\beta > 1$, it is clear that the inequality $e_1 \leq e_2$ allows that variability of the priority service time to be higher than that of the non-priority but not too much.

The customer interarrival time is exponentially distributed with mean λ^{-1}. The customer arrival rate is less than the non-priority service rate, i.e., $\lambda < \mu$, so the system is stable even if all customers are classified as non-priority. The customer's treatment is not interrupted by any customer arrival; i.e., preemption is not allowed.

The relationship between the threshold and the classification algorithm's performance is straightforward and illustrated by the ROC curve. Therefore, choosing the optimal threshold is equivalent to deciding the optimal trade-off between sensitivity and specificity, i.e., selecting the optimal point $(x^*, g(x^*))$ on the ROC curve. In the remainder of the paper, we will focus on the latter.

For a given threshold corresponding to the point $(x, g(x))$ on the ROC curve, the function $G(x) = (\alpha g(x) + (1 - \alpha)x)$ denotes the probability of classifying a random customer as priority, and notations $\lambda_1(x)$ and $\lambda_2(x)$ denote the arrival rate to the priority class and the non-priority class, respectively:

$$\begin{aligned} \lambda_1(x) &= \lambda G(x), \\ \lambda_2(x) &= \lambda\big(1 - G(x)\big). \end{aligned} \tag{2.2}$$

Let $\rho_1(x)$ and $\rho_2(x)$ denote the fraction of time allocated for customers classified as priority and non-priority for a given x, respectively:

$$\begin{aligned} \rho_1(x) &= \frac{\lambda_1(x)}{\beta\mu}, \\ \rho_2(x) &= \frac{\lambda_2(x)}{\mu}. \end{aligned} \tag{2.3}$$

We let $\mathbb{E}[W_1(x)]$ and $\mathbb{E}[W_2(x)]$ denote the average waiting time of the priority and the non-priority customer class, respectively. Using results on non-preemptive queues (Cobham [1954]), we have

$$\begin{aligned}\mathbb{E}[W_1(x)] &= \frac{\mathbb{E}[R(x)]}{1-\rho_1(x)},\\ \mathbb{E}[W_2(x)] &= \frac{\mathbb{E}[W_1(x)]}{(1-\rho_1(x)-\rho_2(x))},\end{aligned} \tag{2.4}$$

where $\mathbb{E}[R(x)]$ is the expected residual time:

$$\mathbb{E}[R(x)] = \frac{\lambda}{2}\left(G(x)(e_1-e_2)+e_2\right). \tag{2.5}$$

Let $W^a(x)$ and $W^b(x)$ denote type-A and type-B customers' average waiting time, respectively. Note that a true type-A customer is assigned to the priority class with probability $g(x)$ and to non-priority class with probability $(1-g(x))$. On the other hand, a true type-B customer is assigned to priority class with probability x and to non-priority class with probability $(1-x)$. Using equations given by (2.2-2.5), we have

$$W^a(x) = g(x)\mathbb{E}[W_1(x)] + (1-g(x))\,\mathbb{E}[W_2(x)] \tag{2.6a}$$

$$= \frac{\lambda}{2}\left(G(x)(e_1-e_2)+e_2\right)\frac{\beta\mu}{\beta\mu-\lambda G(x)}\left(g(x)+\frac{\beta\mu(1-g(x))}{\beta\mu-\lambda\beta+\lambda\left(\beta-1\right)G(x)}\right); \tag{2.6b}$$

$$W^b(x) = x\mathbb{E}[W_1(x)] + (1-x)\,\mathbb{E}[W_2(x)] \tag{2.6c}$$

$$= \frac{\lambda}{2}\left(G(x)(e_1-e_2)+e_2\right)\frac{\beta\mu}{\beta\mu-\lambda G(x)}\left(x+\frac{\beta\mu(1-x)}{\beta\mu-\lambda\beta+\lambda\left(\beta-1\right)G(x)}\right). \tag{2.6d}$$

We let $W(x)$ denotes the average expected cost per customer when the threshold corresponding to the point $(x, g(x))$ on the ROC curve is chosen:

$$W(x) = \gamma\alpha W^a(x) + (1-\alpha)W^b(x). \tag{2.7}$$

Plugging (2.6b, 2.6d) into (2.7), we have

$$W(x) = \frac{\lambda\beta\mu}{2}\frac{\left(G(x)(e_1-e_2)+e_2\right)}{\beta\mu-\lambda G(x)}\left(\gamma\alpha g(x)+(1-\alpha)x+\frac{\beta\mu\left(\gamma\alpha(1-g(x))+(1-\alpha)(1-x)\right)}{\beta\mu-\lambda\beta+\lambda\left(\beta-1\right)G(x)}\right). \tag{2.8}$$

Our objective is minimize the average waiting cost function $W(x)$ with respect to false positive rate $x \in [0,1]$. We evaluate how $W(x)$ looks for both perfect and imperfect information

scenarios. The main difference between the models in these scenarios is the shape of the ROC curve $g(x)$. For the perfect information case $g(x)$ is a horizontal line that takes a value of 1 for all $x \in [0,1]$. For imperfect information case, the ROC curve $g(x)$ can be approximated by an increasing, concave, and piecewise linear function.

2.4 Service Rate Differentiation under Perfect Information

Under perfect information, all type-A customers must be assigned to the priority class because type-A customers have higher delay sensitivity than type-B customers, and priority service is superior to the non-priority service. On the other hand, misclassification of some type-B customers, i.e., assigning them to the priority class, may reduce average waiting costs even if true customer types are known. In this section, we shed light on conditions under which it is optimal to misclassify type-B customers so that the average waiting cost is less than it would be if all customers were perfectly classified. All type-A customers are classified as priority under perfect information, i.e., $g(x) = 1 \ \forall x \in [0,1]$. By plugging this equality into (2.6b), (2.6d) and (2.8) we can write $W^a(x)$, $W^b(x)$, and $W(x)$ as follows

$$W^a(x) = \frac{\lambda}{2}\left(G(x)(e_1 - e_2) + e_2\right)\frac{\beta\mu}{\beta\mu - \lambda G(x)}, \tag{2.9a}$$

$$W^b(x) = \frac{\lambda}{2}\left(G(x)(e_1 - e_2) + e_2\right)\frac{\beta\mu}{\beta\mu - \lambda G(x)}\left(x + \frac{\beta\mu(1-x)}{\beta\mu - \lambda\beta + \lambda\left(\beta - 1\right)G(x)}\right), \tag{2.9b}$$

$$W(x) = \frac{\lambda}{2}\left(G(x)(e_1 - e_2) + e_2\right)\frac{\beta\mu}{\beta\mu - \lambda G(x)}\left(\gamma\alpha + (1-\alpha)x + \frac{\beta\mu(1-\alpha)(1-x)}{\beta\mu - \lambda\beta + \lambda\left(\beta - 1\right)G(x)}\right). \tag{2.9c}$$

where $G(x) = \alpha + (1-\alpha)x \ \forall x \in [0,1]$. By setting $z = G(x)$, i.e., the fraction of customers assigned to priority class, we can rewrite (2.9a-2.9c) as

$$W^a(z) = \frac{\lambda}{2}\left(z(e_1 - e_2) + e_2\right)\left(\frac{\beta\mu}{\beta\mu - \lambda z}\right), \tag{2.10a}$$

$$W^b(z) = \frac{\lambda\beta\mu}{2}\left(z(e_1 - e_2) + e_2\right)\frac{\beta\mu}{\beta\mu - \lambda z}\left(\frac{z-\alpha}{1-\alpha} + \frac{1-z}{1-\alpha}\frac{\beta\mu}{\beta\mu - \lambda\beta + \lambda\left(\beta - 1\right)z}\right), \tag{2.10b}$$

$$W(z) = \frac{\lambda}{2}\left(z(e_1 - e_2) + e_2\right)\frac{\beta\mu}{\beta\mu - \lambda z}\left(\gamma\alpha + (z-\alpha) + \frac{\beta\mu(1-z)}{\beta\mu - \lambda\beta + \lambda\left(\beta - 1\right)z}\right). \tag{2.10c}$$

Under a non-preemptive policy, a customer's service is not interrupted even when a priority customer arrives. Therefore, the expected residual time, i.e., remaining time in service,

impacts all customers' waiting time. The expected residual time depends on the second moments of priority and non-priority service time distributions:

$$\mathbb{E}[R(z)] = \frac{\lambda}{2}\left(z(e_1 - e_2) + e_2\right).$$

Depending on other system parameters, the second moments can change how the average waiting cost function behaves in the interval $z \in [\alpha, 1]$. The second moments are related to the spread of the distribution, i.e., variability. For a fixed value of the first moment, a higher second moment implies higher variability. Proposition 1 shows that if the variability in priority service time relative to the variability in non-priority service time is not *too much higher*, i.e., $\frac{e_1}{e_2}$ is under a certain threshold determined by other systems parameters, then the reduction in the expected residual time is *significant enough* that misclassification of some misclassification of type-B customers can decrease type-A customers' expected waiting time.

Proposition 1. *Function $W^b(z)$ is a decreasing function of z on interval $[\alpha, 1]$. Suppose that $e_1\beta\mu \leq e_2\left(\beta\mu - \lambda\right)$. Function $W^a(z)$ is a nonincreasing function of z on interval $[\alpha, 1]$ (if the inequality is strict, then $W^a(z)$ is decreasing).*

When there is no misclassification, all type-A customers are assigned to the priority queue, and all type-B customers are assigned to the non-priority queue. When there is some level of misclassification, there is some representation of type-B customers in the priority queue. Since the priority queue is served first and faster than the non-priority queue, it makes sense that increasing misclassification, i.e., increasing the representation of type-B customers in the priority queue, decreases type-B customers' average waiting time. However, it is not straightforward to see that misclassification of all type-B customers decreases the waiting time of type-A customers under the assumption that $e_1\beta\mu \leq e_2\left(\beta\mu - \lambda\right)$. To better understand this occurrence, let us write this inequality as:

$$\frac{e_1}{e_2} \leq 1 - \frac{\lambda}{\beta\mu} \implies \frac{1 + SCV_1}{1 + SCV_2}\frac{1}{\beta^2} \leq 1 - \frac{\lambda}{\beta\mu} \implies \frac{1 + SCV_1}{1 + SCV_2} \leq \beta^2 - \beta\frac{\lambda}{\mu}.$$

Assigning type-B customers in the priority queue has both positive externality and negative externality on type-A customers' average waiting time by decreasing the expected residual

time and simultaneously creating additional congestion in the priority queue. Suppose the relative variability in the priority service compared to the non-priority service is small enough. In that case, the expected residual time may decrease significantly when all customers are assigned to the priority service. Suppose the priority service is significantly faster than the non-priority service. In that case, type-B customers can be served so fast that the additional congestion in the priority queue may be small enough not to significantly increase the waiting time (excluding the expected residual time component) of type-A customers. The inequality $e_1\beta\mu \leq e_2(\beta\mu - \lambda)$ characterizes the set of system parameters where the positive externality can compensate for the resulting negative externality on type-A customers' waiting time at all misclassification levels.

Proposition 2. *Suppose that $e_1\beta\mu \geq e_2(\beta\mu - \lambda)$ then $W(z)$ is a strictly convex function of z on interval $[\alpha, 1]$.*

Proposition 2 shows that if the priority service is not superior enough in terms of speed and variability, misclassification of all type-B customers might increase the expected waiting time of type-A customers. In that case, some misclassification can still decrease the average waiting costs up to a point. After that, the negative externality on type-A customers will exceed the positive externality on type-B customers, and thus average waiting costs will increase.

Next, we want to explore how the misclassification incentive depends on the system parameters. Let $\mathcal{M}_B$ denote the optimal misclassification rate among type-B customers for a given set of system parameters:

$$\mathcal{M}_B = \frac{z^*(\alpha) - \alpha}{1 - \alpha}.$$

Proposition 3. *As γ and α increases, $\mathcal{M}_B$ stays the same or decreases.*

Suppose the cost of holding a type-A customer is not significantly high. In that case, it makes sense to misclassify some fraction of type-B customers since the reduction in their expected waiting time costs can compensate for the possible increase in the type-A customers' waiting costs. If type-B customers are the majority of the population, then mislabelling

some fraction of those can decrease overall waiting time costs. Fewer type-A customers are experiencing higher waiting times, while a larger number of type-B customers benefit from lower waiting times.

We have shown that misclassification can help decrease the average waiting costs when the service rate depends on the classification. This statement does not hold when the service rate depends on the true type.

Proposition 4. *Assume that the service rate and the holding cost is type-dependent and priority is determined by the cμ rule, i.e., setting in Argon and Ziya [2009]. Then there is no incentive to misclassify type-B customers as a priority class under perfect information.*

In our setting, misclassification of type-B customers as a priority can decrease the average waiting costs in two ways:(i) by decreasing the expected residual time and (ii) by providing faster service to more customers, i.e., decreasing the overall workload in the system. When service time requirements are type-dependent, both the expected residual time and overall workload in the system stay the same because type-dependent service rates are not controllable to the decision-maker.

In the following sections, we assume that the population mix is always heterogeneous, i.e., $\alpha \in (0, 1)$. Suppose that all customers are either type-A or type-B. Then the optimal decision is to assign all customers to the priority class. In practice, the customer types are not perfectly known and can be predicted with some classification error. The next section will explore the optimal classification strategy under the imperfect information.

2.5 Service Rate Differentiation under Imperfect Information

Under imperfect information, customer types are not known immediately upon arrival. The decision-maker assigns customers to priority classes using a machine learning algorithm with an already established ROC curve. On the ROC curve, the vertical axis, $g(x)$, represents the sensitivity, and the horizontal axis, x, represents (1-specificity). It is not possible to maximize sensitivity and specificity simultaneously. This section sheds light on the optimal sensitivity-specificity trade-off that minimizes the average waiting costs. We approximate the ROC curve by an increasing, concave, and piecewise linear function for mathematical

tractability:

$$g(x) = \begin{cases} \kappa_1 x \quad \text{if } x \in I_1, \\ \kappa_1 x_1 + \kappa_2(x - x_1) \quad \text{if } x \in I_2, \\ \kappa_1 x_1 + \kappa_2(x_2 - x_1) + \kappa_3(x - x_2) \quad \text{if } x \in I_3, \\ \dots \\ \sum_{i=1}^{n-1} \kappa_i(x_i - x_{i-1}) + \kappa_n(x - x_{n-1}) \quad \text{if } x \in I_n, \end{cases} \tag{2.11}$$

where

$$\begin{aligned} I_1 &= [0, x_1]\,, \\ I_n &= (x_{n-1}, x_n] \ \ \forall n > 1. \end{aligned}$$

Since the ROC curve is an increasing and concave function, we assume that $\kappa_1 > \kappa_2 > \cdots > \kappa_n \geq 0$. On any interval I_n, ROC curve can be written in the form of $g(x) = f + \kappa x$:

$$\begin{aligned} f &= \left(\sum_{i=1}^{n-1} \kappa_i(x_i - x_{i-1}) \right) - x_{n-1}\kappa_n, \\ \kappa &= \kappa_n \end{aligned} \tag{2.12}$$

where $x_0 = 0$. The slope parameter κ_n informs us about the trade-off between the sensitivity and specificity on the interval I_n.

Proposition 5. *Suppose that $e_1 \leq e_2$ and the ROC curve is given by (2.11). $W^a(x)$ and $W(x)$ are decreasing functions of x on interval $[0, x_{\bar{n}}]$, where $\bar{n} = \max\{n : \kappa_n \geq 1\}$.*

Proposition 5 shows that it is optimal to trade off a lower loss in the specificity for a higher gain in the sensitivity to minimize $W(x)$ and $W^a(x)$.

To have a better understanding of why the average waiting cost of type-A customers decreases on the interval $[0, x_{\bar{n}}]$, consider the following example. Suppose that $x = 0$, i.e., all customers are classified as non-priority. Then there is a single queue that represents the overall customer mix. Suppose that $x = x_{\bar{n}}$. Then there are both priority and non-priority queues. Since the marginal increase in true positive rate is at least as much as the increase in the false positive rate on the interval, $[0, x_{\bar{n}}]$, true type-A customers are over-represented in the priority queue and under-represented in the non-priority queue. Therefore, the average waiting time of a type A customer is less than what it would be if the false positive rate

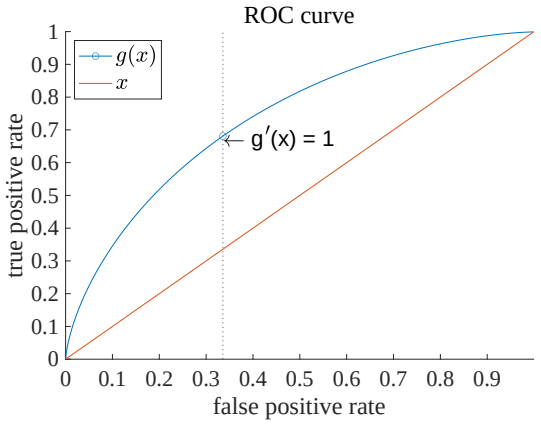
were set to $x = 0$ since the priority line is served first and at a faster rate.

On the other hand, we cannot even determine if the average waiting cost of type-B customers is monotone on the interval $[0, x_{\bar{n}}]$. Suppose that $x = x_{\bar{n}}$. Then we know there are both priority and non-priority queues. We also know that true type-B customers are under-represented in the priority queue and over-represented in the non-priority queue. Suppose that priority and non-priority service time are drawn from the same distribution. The priority line is served first, and the over-representation of type-B customers in the non-priority line hurts their waiting time. The average waiting time of a type-B customer at $x = x_{\bar{n}}$ is more than what it would be at $x = 0$. Now, retrieve our main assumption on the service time distributions, i.e., the priority service time is shorter and not so much more variable than non-priority service time. Then some representation of type-B customers in the priority reduces the waiting time of other type-B customers. The sensitivity-specificity trade-off on interval $[0, x_{\bar{n}}]$ has both negative and positive impact on type-B customers' waiting time. Figure 2.1 shows as β increases; the positive impact tends to exceed the negative impact. Higher β means shorter priority service and lower waiting time for patients in the non-priority queue and thus higher positive impact.

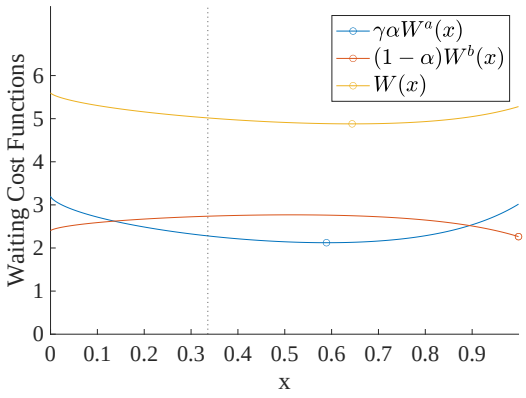
Now, we can give an intuitive explanation why average waiting cost function decreases on the interval $[0, x_{\bar{n}}]$. Suppose that type-B customers' waiting costs decrease on some subset of $[0, x_{\bar{n}}]$. Then it is not surprising that overall waiting costs go down. Suppose that type-B customers waiting costs stays the same or increases on some subset of $[0, x_{\bar{n}}]$. We know that type-B customers are negatively impacted because they are underrepresented in the priority queue and got jumped by some type-A customers who would be served afterward at $x = 0$. Since type-A customers have higher waiting costs, this negative impact on type-B customers waiting costs is immediately canceled by the positive impact on type-A customers waiting costs. Proposition 5 allows us to limit our search for the optimal x^* to interval $[x_{\bar{n}}, 1]$ where the slope is less than 1, i.e. a higher gain of sensitivity is traded for a lower loss of specificity.

Theorem 1. *Average waiting cost function $W(x)$ is unimodular. There exists a unique global minimum $x^* \in [x_{\bar{n}}, 1]$.*

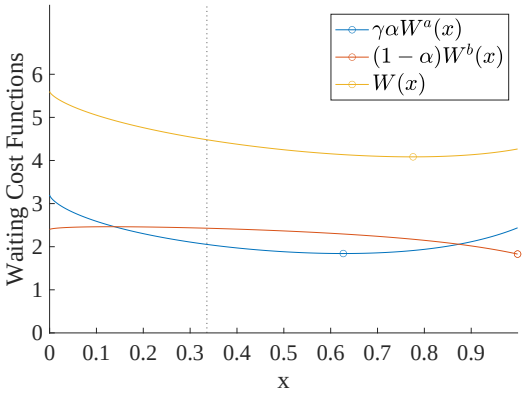
As x increases on the interval $(x_{\bar{n}}, 1]$, true type-B customers' representation in the priority



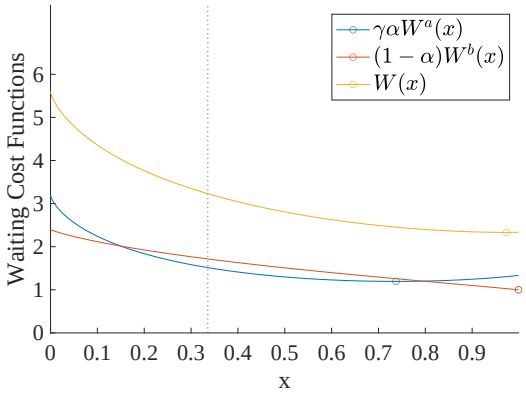
(a) ROC curve



(b) $\beta = 1.02$



(c) $\beta = 1.05$



(d) $\beta = 1.2$

Figure 2.1: An example of how β affects behavior of waiting cost functions on interval $[x_{\bar{n}}, 1]$ where $\alpha = 0.4$, $\gamma = 2$, $\mu = 1$, $\lambda = 0.8$, $\text{SCV}_1 = 1$, $\text{SCV}_2 = 1$

queue also increases because the marginal increase in false positive rate is at least as much as the increase in the true positive rate on the interval. On the other hand, the type-A representation in the priority queue decreases on interval $(x_{\bar{n}}, 1]$. This representation change would cause increases in average waiting costs if priority and non-priority service times were drawn from the same distribution because type-A customers are more costly to hold in the queue. However, priority service is faster and decreases the overall workload in the system without increasing the variability too much. Therefore, assigning more customers to the priority class can decrease both type-A and type-B customers' waiting time. Theorem 1 tells us that the residual time and system workload decrease induced by a higher priority service utilization can compensate for the advantage loss of type-A customers in the service order up to some degree, if not at all. Average waiting costs go down as x approaches x^* and then go up.

2.6 The Expected Value of Perfect Information

Throughout our analysis, we have assumed that the classification algorithm is fixed, and thus the ROC curve is also fixed. The decision-maker can improve the classification algorithm performance by investing resources and time. Depending on the application, performance improvements might be too costly. In emergency departments, utilization of ED laboratory tests can improve the classification accuracy at the expense of creating congestion for important resources and delaying the classification process itself. Therefore, it might not be ideal to order more tests to improve the classification accuracy. We would like to understand under what settings the investment in classification accuracy can significantly decrease average waiting costs to justify the investment cost. In other words, we want to analyze the EVPI. To have analytical tractability and interpretability, we approximate the ROC curve by

$$g(x) = \begin{cases} \kappa x & \text{if } x \in \left[0, \frac{1}{\kappa}\right], \\ 1 & \text{if } x \in \left(\frac{1}{\kappa}, 1\right]. \end{cases} \tag{2.13}$$

where $\kappa \geq 1$. The slope parameter κ measures the performance of the classification algorithm. Higher κ is associated with higher accuracy. Suppose that κ is very large, then $g(x)$ takes the value of 1 very quickly, which resembles the perfect information case.

Corollary 1. *Suppose that $e_1 \leq e_2$ and ROC curve is given by (2.13). Then $W^a(x)$ and $W(x)$ is a decreasing function of x on interval $\left[0, \frac{1}{\kappa}\right]$.*

Corollary 1 immediately follows from Proposition 1. Using this result and analysis of the perfect information case, we analyze the imperfect information case with a simple ROC curve (2.13). By Corollary 1, $W(x)$ is a decreasing function of x on interval $\left[0, \frac{1}{\kappa}\right]$. Then the optimal false positive rate $x^* \in \left[\frac{1}{\kappa}, 1\right]$ on where $g(x^*) = 1$, i.e., it is optimal to misclassify at least $\frac{1}{\kappa}$ fraction of the type-B customers. We restrict our search for optimal x^* to the interval $\left[\frac{1}{\kappa}, 1\right]$. By the construction of the ROC curve, $g(x) = 1$ for all $x \in \left[\frac{1}{\kappa}, 1\right]$, i.e., any $x \in \left[\frac{1}{\kappa}, 1\right]$ allows all type-A customers to be classified as priority. The optimal x^* in on interval $\left[\frac{1}{\kappa}, 1\right]$ and the waiting cost function on this interval behaves similar to the waiting cost function under perfect information only differing in the following. The optimal misclassification rate of type-A customers can take any value from 0 to 1 in the perfect information case, while we know that the optimal misclassification rate is at least $\frac{1}{\kappa}$ under the imperfect information case with a simple ROC curve.

We define $z = G(x) = \alpha + (1-\alpha)x$ for any $x \in \left[\frac{1}{\kappa}, 1\right]$ and express the average cost function for all $z \in \left[\alpha + \frac{1-\alpha}{\kappa}, 1\right]$ as follows:

$$W(z) = \frac{\lambda}{2}\left(z(e_1 - e_2) + e_2\right)\frac{\beta\mu}{\beta\mu - \lambda z}\left(\gamma\alpha + (z-\alpha) + \frac{\beta\mu(1-z)}{\beta\mu - \lambda\beta + \lambda\left(\beta - 1\right)z}\right). \tag{2.14}$$

Observe that the average cost function with a simple ROC curve (2.14) is identical to the average cost function with perfect information (2.10c) on interval $\left[\alpha + \frac{1-\alpha}{\kappa}, 1\right]$.

Corollary 2. *Suppose that the ROC is given by 2.13). If $e_1\beta\mu \leq e_2\left(\beta\mu - \lambda\right)$, then $W(z)$ is decreasing function of z on interval $\left[\alpha + \frac{1-\alpha}{\kappa}, 1\right]$. If $e_1\beta\mu \geq e_2\left(\beta\mu - \lambda\right)$, then $W(z)$ is a strictly convex function of z on interval $\left[\alpha + \frac{1-\alpha}{\kappa}, 1\right]$ and there exists a unique $z^* \in \left[\alpha + \frac{1-\alpha}{\kappa}, 1\right]$.*

By Proposition 1, $W(z)$ is a decreasing function of z on interval $[\alpha, 1]$ under the assumption that $e_1\beta\mu \leq e_2\left(\beta\mu - \lambda\right)$. By Proposition 2, $W(z)$ is a strictly convex function of z on interval $[\alpha, 1]$ under the assumption $e_1\beta\mu \geq e_2\left(\beta\mu - \lambda\right)$. The corollary immediately follows.

Recall that the optimal z^*, that minimizes function $W(z)$ (e.q. 2.10c) under perfect information, is unique and on interval $[\alpha, 1]$. Then the perfect information problem is

$$\mathcal{P}_p = \{\min W(z) \text{ s.t. } z - 1 \leq 0,\ \alpha - z \leq 0\}.$$

The optimal solution z^* that minimizes the average waiting cost function under imperfect

information with a simple ROC curve is unique and on interval $\left[\alpha + \frac{1-\alpha}{\kappa}, 1\right]$. On this interval, the average waiting cost function is given by the same function as in the perfect information case. Then we minimization problem under imperfect information is

$$\mathcal{P}_i = \{\min W(z) \text{ s.t. } z - 1 \leq 0,\ \alpha + \frac{1-\alpha}{\kappa} - z \leq 0\}.$$

We define the EVPI as follows:

$$\text{EVPI} = W(z_i^*) - W(z_p^*). \tag{2.15}$$

Note that both problems have the same objective, and the feasible region of $\mathcal{P}_i$ is a subset of the feasible region of $\mathcal{P}_p$. Let z_p^*, z_i^* be the optimal solution to $\mathcal{P}_p$ and $\mathcal{P}_i$, respectively. Suppose that z_p^* is feasible to $\mathcal{P}_i$, then $z_i^* = z_p^*$ and the EVPI is zero. We examine the EVPI in two different cases: (i) $e_1\beta\mu \leq e_2\,(\beta\mu - \lambda)$ and (ii) $e_1\beta\mu > e_2\,(\beta\mu - \lambda)$.

Case I: Suppose that $e_1\beta\mu \leq e_2\,(\beta\mu - \lambda)$. By Proposition 1, $W(z)$ decreases as z increases on interval $[0, 1]$, and thus $z_p^* = 1$ and $z_i^* = 1$. Since both $\mathcal{P}_p$ and $\mathcal{P}_i$ have the same objective function, $W(z_i^*) = W(z_p^*)$ and the EVPI zero.

Case II: Suppose that $e_1\beta\mu > e_2\,(\beta\mu - \lambda)$. There are two subcases: either z_p^* is a feasible solution to $\mathcal{P}_i$ or not. We separately examine both subcases.

- Suppose that z_p^* is also a feasible solution to $\mathcal{P}_i$, i.e., $z^* \geq \alpha + \frac{1-\alpha}{\kappa}$. Then $W(z_i^*) = W(z_p^*)$ and the EVPI is zero.

- Suppose that z_p^* is not feasible to $\mathcal{P}_i$, i.e., $z_p^* < \alpha + \frac{1-\alpha}{\kappa}$. By Proposition 2, $W(z)$, the objective function of the minimization problem under perfect information, is strictly convex. Therefore, $W(z)$ increases for all $z \geq z_p^*$. Therefore, the optimal solution to $\mathcal{P}_i$ is $z_i^* = \alpha + \frac{1-\alpha}{\kappa}$ and

$$\text{EVPI} = W(z_i^*) - W(z_p^*) = W\left(\alpha + \frac{1-\alpha}{\kappa}\right) - W(z_p^*) = \int_{z_p^*}^{\alpha + \frac{1-\alpha}{\kappa}} dW(z). \tag{2.16}$$

Proposition 6. *The EVPI is nondecreasing with α and γ.*

As type-A customers' waiting cost rate or prevalence increases, it becomes more important

to classify patients correctly.

To understand how β affects the EVPI, consider the inequality $e_1\beta\mu \leq e_2(\beta\mu - \lambda)$, i.e.,

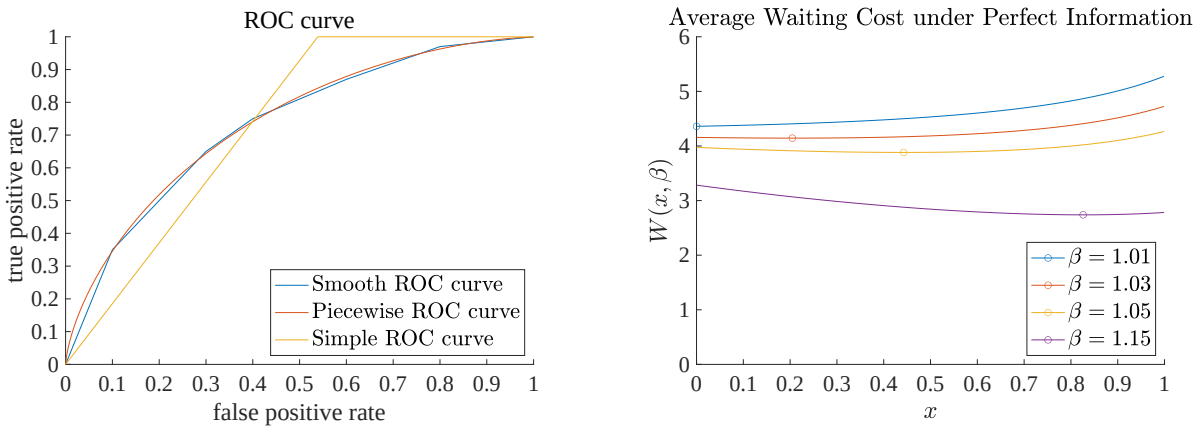
$$\frac{e_1}{e_2} - \beta^2 + \beta\frac{\lambda}{\mu} \leq 0,$$

holds, then the EVPI is zero. Note that this inequality holds if

$$1 - \beta^2 + \beta \leq 0,$$

because $\frac{\lambda}{\mu} < 1$ and $\frac{e_1}{e_2} \leq 1$. Then it is optimal to assign all type-B customers to the priority class and the EVPI is zero.

We ran numerical experiments to shed light on how the relative speed of the priority service affects the EVPI where $\beta < \frac{1+\sqrt{5}}{2}$ for various values of other system parameters. We observe that as β increases, the EVPI tends to decrease for a wide range of system parameters regardless of how we approximate the ROC curve. Figure 2.2a shows three different ROC curve approximations with similar Area Under Curve (AUC): a smooth ROC curve based on signal distribution Beta$(2, 3)$, a piecewise linear ROC curve with multiple breakpoints, and a simple ROC curve with a single breakpoint. Figure 2.2b shows an example of how average waiting cost function behaves and optimal solution changes at different values of β. We numerically calculate the average waiting cost function for three ROC curve scenarios shown in Figure 2.2a and then calculate the EVPI. Figure 2.3 shows that an increase in β decreases the absolute value of EVPI and the percentage value of EVPI, the ratio of the absolute difference between the minimum average cost under perfect and imperfect information to the minimum cost under imperfect information. This result is consistent with the fact that EVPI is zero if $e_1\beta\mu \leq e_2(\beta\mu - \lambda)$. Because the lowest value of the EVPI can get is zero and the inequality is more likely to hold for higher values of β given that other system parameters are fixed.



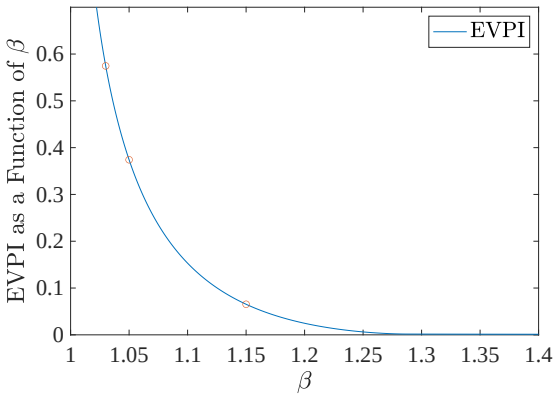
(a) ROC curve approximations with similar AUC (b) Average waiting cost functions

Figure 2.2: An example of average waiting cost functions at different values of β under perfect information where $\alpha = 0.4$, $\gamma = 2$, $\mu = 1$, $\lambda = 0.8$, $\text{SCV}_1 = 1$, $\text{SCV}_2 = 1$

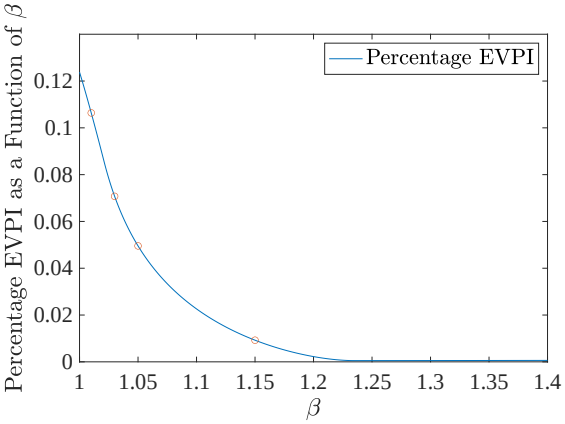
2.7 Classification when the Service Rate is Type-dependent

This section discusses similarities and differences between our work and Argon and Ziya [2009]. The authors consider an M/G/1 queuing model where each customer is either type-A with probability p_A or type-B with probability p_B. Arrivals follow a Poisson process with a rate λ. Service rates and delay sensitivities depend on the type. Service times of type $i \in \{A, B\}$ customers are i.i.d distributed with first and second moments a_i, e_i. The system load is $\rho = \lambda (p_A a_A + p_B a_B)$. The unit cost of keeping a type $i \in \{A, B\}$ customer is h_i. In the standard M/G/1 queueing system with two priority classes, the priority is given according to the $c\mu$ rule. Without loss of generality, $h_A/a_A > h_B/a_B$ and type-A customers should receive higher priority than type-B customers to minimize average waiting costs. Delay sensitivity depends on the true customer type in our setting, while the service rate depends on the classification. Type-A customers should receive priority in our setting because they have higher delay sensitivity than type-B customers, i.e., $h_A > h_B$. Service times of customers, who are classified as priority (non-priority), are i.i.d distributed with first and second moments a_1, e_1 (a_2, e_2).

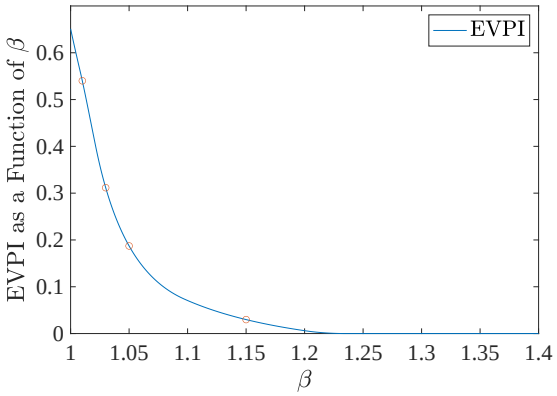
Argon and Ziya [2009] assume that customer type is unknown upon arrival, but each customer provides a signal representing the probability of being type A. The signal is an



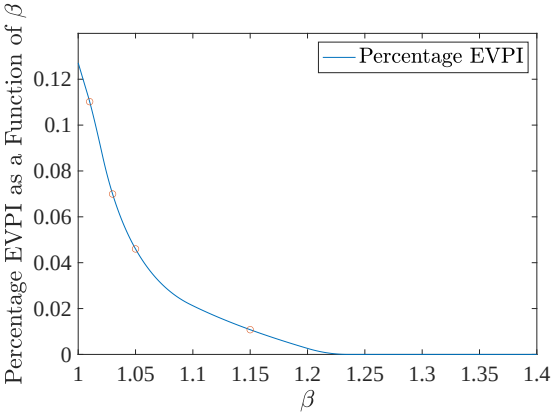
(a) EVPI with smooth ROC curve



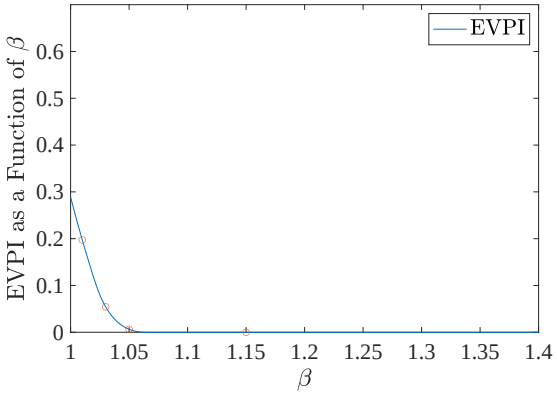
(b) Percentage EVPI with smooth ROC curve



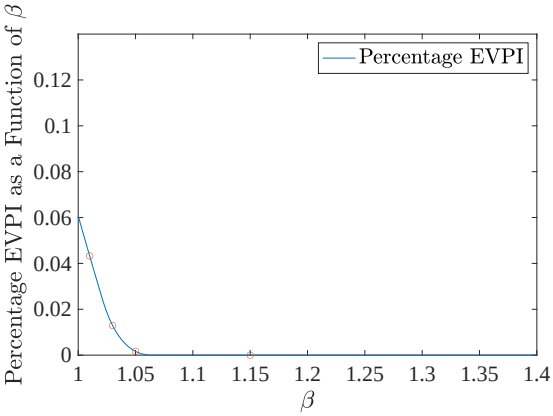
(c) EVPI with piecewise ROC curve



(d) Percentage EVPI with piecewise ROC curve



(e) EVPI with simple ROC curve



(f) Percentage EVPI with simple ROC curve

Figure 2.3: An example of how β affects the EVPI where $\alpha = 0.4$, $\gamma = 2$, $\mu = 1$, $\lambda = 0.8$, $\text{SCV}_1 = 1$, $\text{SCV}_2 = 1$

i.i.d. random variable with p.d.f. $b(\cdot)$ and c.d.f. $B(\cdot)$. By definition,

$$p_A = \int_0^1 xb(x)dx \text{ and } p_B = \int_0^1 (1-x)b(x)dx.$$

They study a single threshold policy that classifies those with signals t and above as class-1 and others as class-2. Their goal is find the optimal threshold policy that minimizes the average waiting costs.

We let α be the fraction of priority customers in the population, i.e. the probability of a customer being type A. We assume that patient mix $(\alpha, 1-\alpha)$ is given and capture the misclassification error in classification via ROC curve $\{(x, g(x)) : x \in [0,1]\}$ instead of directly working with the signal distribution. Recall that in the ROC plot, horizontal axis represents the false positive rate and vertical axis represents the true positive rate. By using the notation described in Argon and Ziya [2009], we have

$$\text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\int_t^1 xb(x)d(x)}{\int_0^1 xb(x)d(x)} = \frac{E(t)}{p_1}, \tag{2.17a}$$

$$\text{FPR} = \frac{\text{FP}}{\text{N}} = \frac{\int_t^1 (1-x)b(x)d(x)}{\int_0^1 (1-x)b(x)d(x)} = \frac{\bar{E}(t)}{p_2}. \tag{2.17b}$$

To make a meaningful comparison between our work and Argon and Ziya [2009], we let $\alpha = p_1$ and define the ROC curve as follows:

$$\text{ROC} = \left\{ \left(\frac{E(t)}{p_1}, \frac{\bar{E}(t)}{p_2} \right) : t \in [0,1] \right\}. \tag{2.18}$$

Figure 2.4 shows an example of the mapping given by equations (2.17-2.18).

We assume that the service time distribution of priority and non-priority classes in our model would be identical to the service time distribution (of type-A and type-B customers in the type-dependent service rate model, Argon and Ziya [2009], was used.) with the minimum mean and maximum mean, respectively. This ensures that prioritizing type-A customers as in Argon and Ziya [2009] does not violate the $c\mu$ rule.

We first show that the optimal threshold policy obtained for the model with class-dependent service rates is different from that the optimal threshold policy for the model

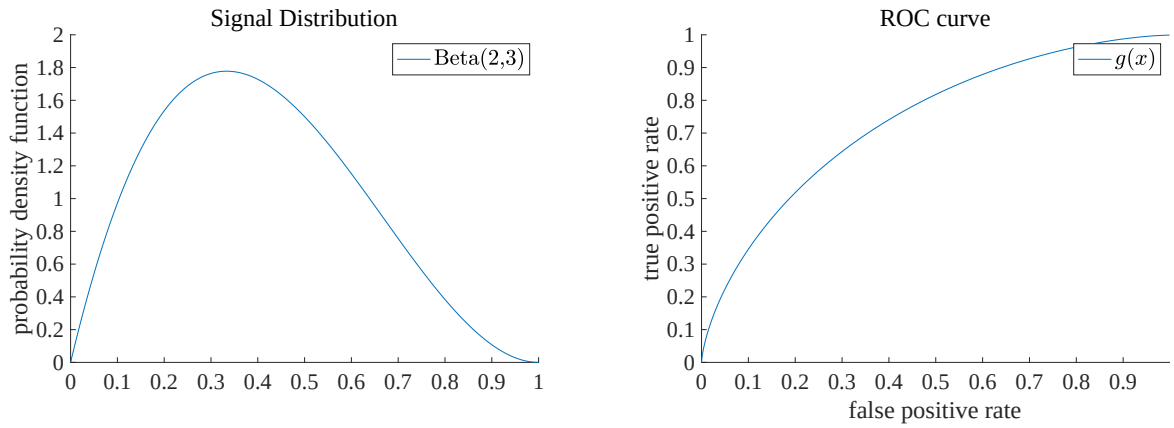


Figure 2.4: An example of a signal distribution and corresponding ROC curve.

with type-dependent service rates. We run a numerical example by setting $\alpha = 0.4$, $\lambda = 0.8$, $h_A = 2$, $h_B = 1$, $\beta = 1.1$, $\mu = \frac{1}{a_B} = 1$, $\text{SCV}_1 = \text{SCV}_A = \text{SCV}_2 = \text{SCV}_B = 1$. The signal distribution and corresponding ROC curve are given by Figure 2.4. Type A is the priority group in both settings. Average waiting costs at different thresholds are plotted in Figure 2.5.

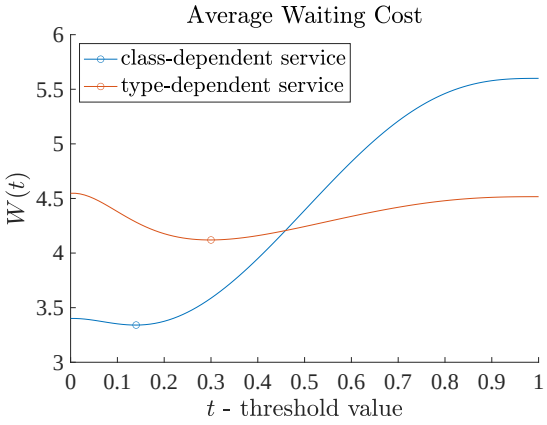


Figure 2.5: Comparison of average waiting costs at different thresholds.

Argon and Ziya [2009] show that there exists a unique threshold value that minimizes the waiting costs. This threshold depends on the signal distribution, system load and patient mix (p_1, p_2). Higher moments of service time distribution do not affect the optimal threshold. Note that higher moments of service time affect long-run average waiting costs by affecting expected residual time. Since service time requirements depend on the type, classifying a

customer as type-A or type-B does not change the expected residual time. If a customer is classified as a priority in our setting, their service time distribution has a lower mean and second moment than if they are classified as non-priority. Thus, classification affects long-run average waiting costs by changing the expected residual time, and the optimal threshold depends on the higher moments of the service time.

Argon and Ziya [2009] proves that the optimal threshold decreases as the system load increases; (ii) optimal threshold converges to p_1 as the system load approaches 0; (iii) optimal threshold converges to 0 as the system load approaches 1. Our numerical example shows that this statement does not hold when the service rate depends on the classification rather than the type. In Figure 2.6, average waiting costs are plotted by varying the arrival rate and keeping other problem parameters the same. This is equivalent to varying the system load. As shown in the figure, the threshold value at $\lambda = 0.8$, i.e., at a higher system load, the optimal threshold is larger than it would at $\lambda = 0.2$, i.e., at a lower system load. We observe that the threshold value at $\lambda = 0.95$, i.e., at a higher system load, the optimal threshold is smaller than what it would be at $\lambda = 0.8$, i.e., at a lower system load. Based on these two examples, we cannot argue that there is a monotonic change at optimal threshold values as the arrival rate varies.

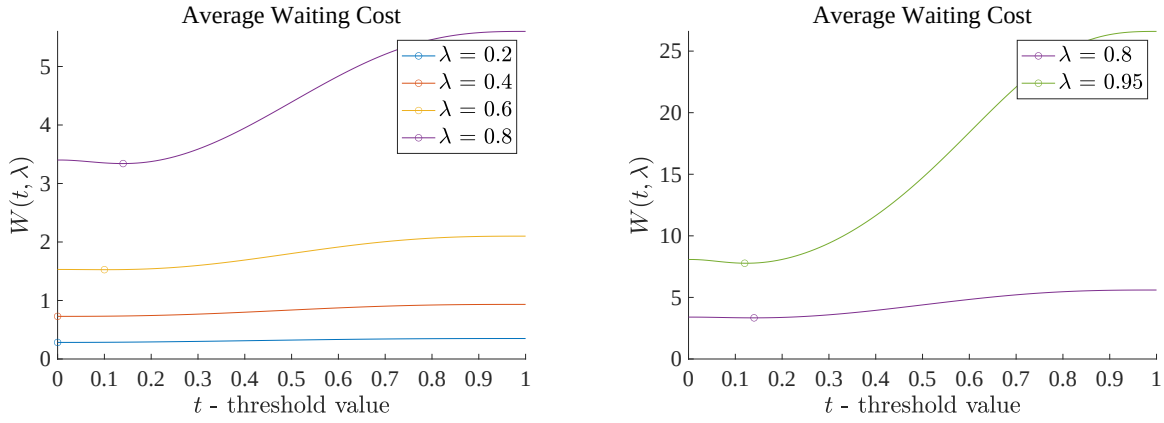


Figure 2.6: A comparison of optimal threshold values at different values of λ in system where the service rate is classification-dependent.

2.8 Nonlinear Waiting Costs

We have assumed that waiting costs are linear in time, which might not be realistic for some applications in practice. This section explores how our analysis changes if we relax the linear waiting cost assumption. Theoretical analyse of queuing systems with nonlinear waiting costs can be difficult. In the literature, increasing and convex functions such as quadratic functions are popular functional forms to study the queuing systems with nonlinear waiting costs.

Suppose that the waiting cost function is given by the following quadratic function hz^2 where h is a constant and z is the time spent in the queue. Using the results from Durr [1969], we can express average waiting cost function $W(x)$ in closed form:

$$W(x) = \mathbb{E}\left[W_1^2(x)\right](\gamma\alpha + (1-\alpha)x) + (1-x)\,\mathbb{E}\left[W_2^2(x)\right] \tag{2.19}$$

$$\begin{aligned}
\mathbb{E}\left[W_1^2(x)\right] &= \left(\frac{EU}{(1-\rho_1(x))^2} + \frac{ER\lambda_1(x)e_1}{(1-\rho_1(x))^3})\right)(1-\rho_1(x)) \\
\mathbb{E}\left[W_2^2(x)\right] &= \left(\frac{EU}{(1-\rho_1(x))\,(1-\rho_1(x)-\rho_2(x))^2} + \frac{ER\lambda_1(x)e_1}{(1-\rho_1(x))^2\,(1-\rho_1(x)-\rho_2(x))^2} + \right. \\
&\qquad \left. \frac{ER\,(\lambda_1(x)e_1+\lambda_2(x)e_2)}{(1-\rho_1(x))\,(1-\rho_1(x)-\rho_2(x))^3}\right)\frac{1-\rho_1(x)-\rho_2(x)}{1-\rho_1(x)},
\end{aligned} \tag{2.20}$$

where

$$\begin{aligned}
\lambda_1(x) &= \lambda G(x), \\
\lambda_2(x) &= \lambda\,(1-G(x))\,, \\
\rho_1(x) &= \frac{\lambda_1(x)}{\beta\mu}, \\
\rho_2(x) &= \frac{\lambda_2(x)}{\mu}, \\
ER &= \frac{\lambda}{2}\,(G(x)(e_1-e_2)+e_2)\,, \\
EU &= \frac{\lambda}{3}\,(G(x)(u_1-u_2)+e_2)\,.
\end{aligned} \tag{2.21}$$

Notation u_1 and u_2 refer to the third moment of service time distribution for priority and nonpriority classes, respectively.

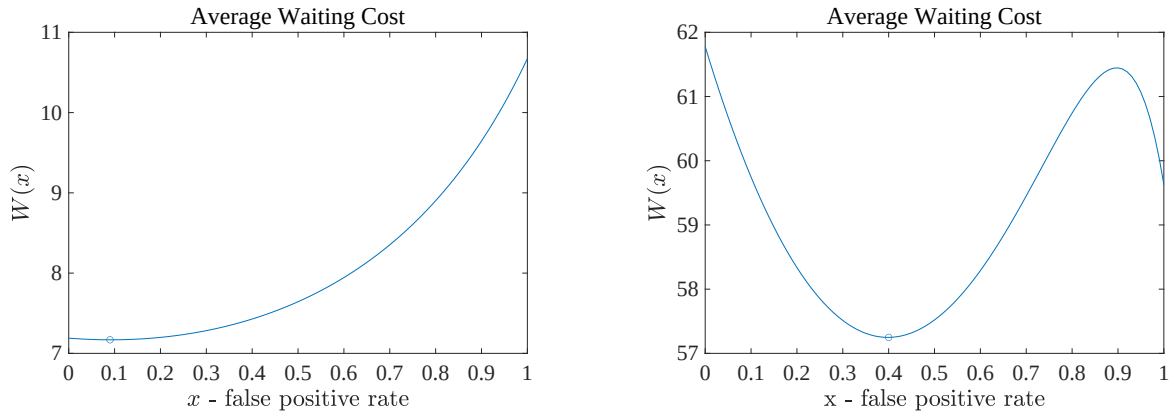
We run three numerical examples with quadratic waiting costs. The first and second

numerical examples are counterexamples for Proposition 2, and Proposition 5, respectively. The third numerical example is a counterexample for a lemma used in the proof of Theorem 1.

In section 2.4 where we assumed perfect information and linear waiting costs, we have proved Proposition 2, i.e., if the inequality $e_1\beta\mu \geq e_2\left(\beta\mu - \lambda\right)$ holds, then $W(x)$ is a strictly convex function of x on interval $[0, 1]$. This proposition allowed us prove the uniqueness of the optimal threshold and perform a theoretical analysis of the expected value of the perfect information.

We run the first numerical example by setting $\alpha = 0.6$, $\lambda = 0.8$, $h_A = 7$, $h_B = 1$, $\beta = 1.2$. In the example, we assume that priority and nonpriority service time are random variables drawn from Gamma(6,6) and Gamma(36,30), respectively where the first term is the shape parameter and the second term is the rate parameters: $\mu = 1/1.2$, $\beta = 1.2$, $e_1 \approx 1.16$, $u_1 \approx 1.55$, $e_2 = 1.48$, and $u_2 \approx 1.87$.

Figure 2.7 illustrates that the convexity property is no longer satisfied when the waiting cost is quadratic in time.



(a) waiting cost is linear in time

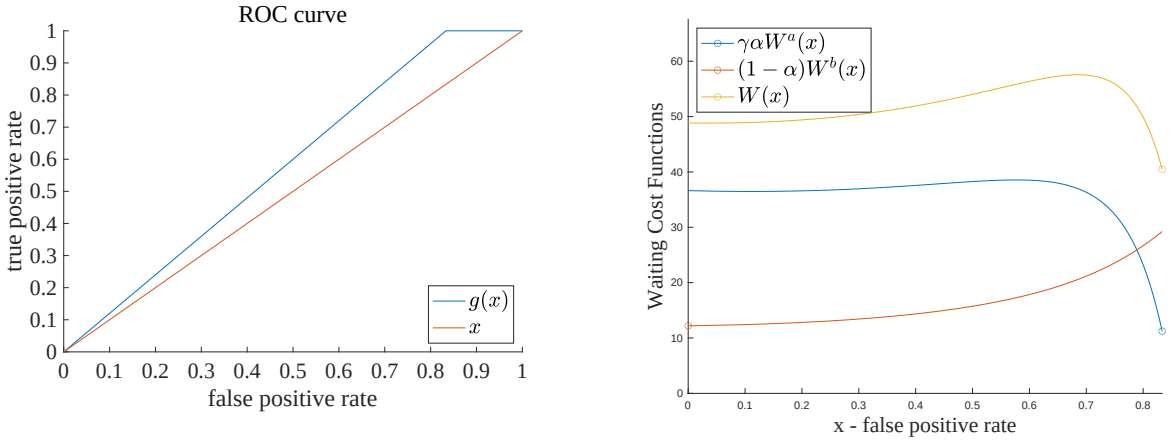
(b) waiting cost is quadratic in time

Figure 2.7: Comparison of average waiting cost function with linear and quadratic waiting costs under perfect information

In section 2.5, we proved that there is a unique solution to minimize $W(x)$. For that purpose, we divide the domain $[0, 1]$ into two region: region where $\kappa \geq 1$ and region where $\kappa < 1$. We have showed that $W^a(x)$ and $W(x)$ are monotonically decreasing functions of x on region where $\kappa \geq 1$ which allowed us to limit our search to the latter region.

In the second numerical example, we set $\alpha = 0.6$, $\lambda = 0.8$, $h_A = 7$, $h_B = 1$ and assume that priority and nonpriority service time are random variables drawn from Gamma(4,4) and Gamma(3.15,3), respectively. In other words, $\mu = 1/1.05$, $\beta = 1.05$, $e_1 = 1.25$, $u_1 \approx 1.875$, $e_2 \approx 1.4525$, and $u_2 \approx 2.49$. We use the curve shown in Figure 2.9a to model the imperfect information.

As shown in Figure 2.8, $W^a(x)$ and $W(x)$ are no longer monotonic decreasing functions of x on region where $\kappa \geq 1$ under the quadratic waiting costs assumption. The second part



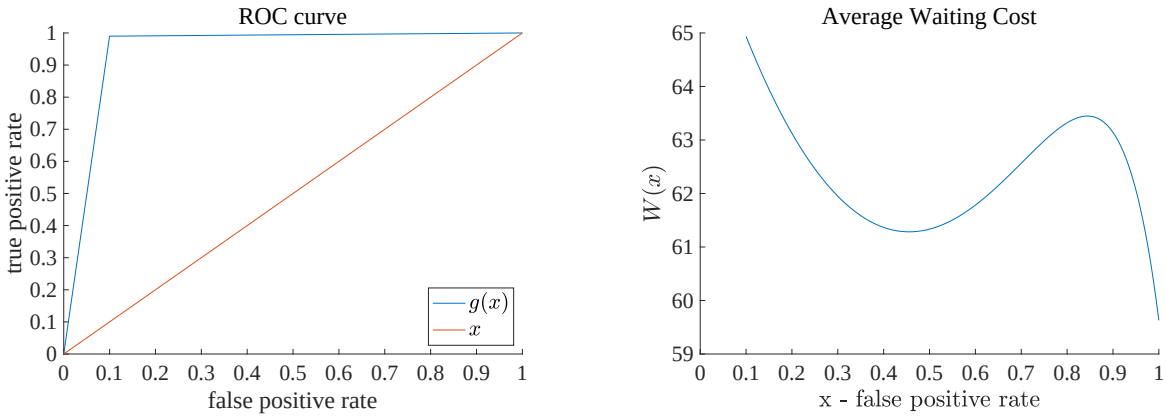
(a) ROC curve

(b) Close-up on region where $\kappa \geq 1$

Figure 2.8: Counter example to Proposition 5 when the waiting costs are quadratic

of the uniqueness proof also depends on the proof that $W(x)$ is a unimodular function of κ on that region where $\kappa < 1$. Please refer to the Appendix for details. We run the third numerical example by setting $\alpha = 0.6$, $\lambda = 0.8$, $h_A = 7$, $h_B = 1$. In the example, we assume that priority and nonpriority service time are random variables drawn from Gamma(6,6) and Gamma(30,48), respectively where the first term is the shape parameter and the second term is the rate parameters: $\mu = 1/1.2$, $\beta = 1.2$, $e_1 \approx 1.16$, $u_1 \approx 1.55$, $e_2 = 1.48$, and $u_2 \approx 1.87$. We use the ROC curve shown in Figure 2.9a to model the imperfect information.

As shown in Figure 2.9, $W(x)$ is no longer unimodular function of x on the region where $\kappa < 1$.



(a) ROC curve

(b) Close-up on region where $\kappa < 1$

Figure 2.9: Counter example to the unimodularity of $W(x)$ on region where $\kappa < 1$

CHAPTER 3
DECISION TREES FOR PATIENT PRIORITIZATION IN EMERGENCY DEPARTMENTS

3.1 Introduction

In the Emergency Department (ED), overcrowding occurs when "the identified need for emergency services exceeds available resources for patient care in the emergency department (ED), hospital, or both (ACEP [2019])". Due to this mismatch between resource capacity and the need for healthcare services, patients can experience negative consequences such as long wait times and poor care quality. When an ED is overcrowded, it approaches its full capacity; new patients are placed in a waiting room until a licensed ED bed becomes available or they get medical treatment in the waiting room/hallway bed. The quality of care in hallway beds and chairs is inferior compared to the treatment on a standard and licensed ED bed (Richards et al. [2014]). Given the limited availability of standard ED beds, patients are prioritized to be placed in a standard ED bed based on their urgency and resource needs.

All patients are "triaged", i.e., sorted by acuity upon their arrival. In the United States, the five-level ESI (Emergency Severity Index) triage system is widely used in emergency departments (EDs), with ESI-1 being the most urgent and ESI-5 being the least urgent. The ESI score is determined by demographics, chief complaints, medical history, vital signs, and estimated resource needs (Gilboy et al. [2020]). The ESI triage system provides triage nurses with guidelines on assigning acuity scores to patients, which can be seen as a proxy for how urgent the patient needs medical treatment. The decision-makers consider the time-sensitive nature of the patient's health status and aim to provide medical treatment to patients quickly and effectively. Since ESI scores are a proxy for how long a patient can wait without significant deterioration in the health status, the ESI score is an important factor in the patient prioritization for ED beds.

Other factors that play a role in patient prioritization might be related to various aspects of the expected care that cannot be explicitly captured in the data. For example, predicted disposition outcomes can impact the assignment decision. Li et al. [2021] study the patient

prioritization for physician assessment and empirically show that if the ED blocking is sufficiently low, Admit patients are prioritized over Discharge within the high acuity group. On the other hand, if the ED blocking is sufficiently high, then Discharge patients are prioritized over Admit patients.

Not all patients in the ED have life-threatening emergencies. Patients who show up with minor complaints are usually assigned to the lower acuity class. Although treating patients with minor complaints takes a shorter time, they could wait more because they are de-prioritized over patients with more complicated complaints. To improve patient flow and decrease the waiting time, ED departments have established a fast-track area dedicated to patients with less urgent medical conditions (Welch [2009]). Patients with higher acuity scores are prioritized for the primary service area, while patients with lower acuity scores are prioritized for the fast-track area.

The Emergency Department at the University of Chicago Emergency Medicine (UCM) had thirty-six treatment rooms between November 1, 2016, and December 15, 2017. Thirty-one of those were sufficiently equipped to treat patients with mid-to-high acuity. Five ED rooms were better suited to treat patients with non-emergency situations and utilized similarly to the fast-track area. In literature, most studies focus on the primary service area by excluding the fast-track area during the peak load hours. At UCM, a choice incident, where multiple patients are waiting to be placed in a bed, happens at any time during the day, not just during a specific time window. Figure 3.1 shows that the time window between 12 pm and 6 am on the next day includes the highest number of choice incidents while 17.88% of the incidents occur between 6 am and 12 pm, during which non-urgency rooms are not available. Figure 3.2 shows that the 33.36% of patients who are assigned to an urgency room between 6 am and 12 pm are by low-priority patients while this ratio goes down to 1.49% between 6 pm and midnight. In other words, urgency rooms are not only used for mid-high acuity patients but also for low acuity patients. The utilization percentage by each acuity group changes whether the fast tract is open and the time of the day.

We model the patient prioritization problem using the discrete choice framework, particularly the conditional logit model. A patient assigned to a bed represents a decision epoch. All patients waiting to be assigned to a bed in the ED comprises a choice set; the current

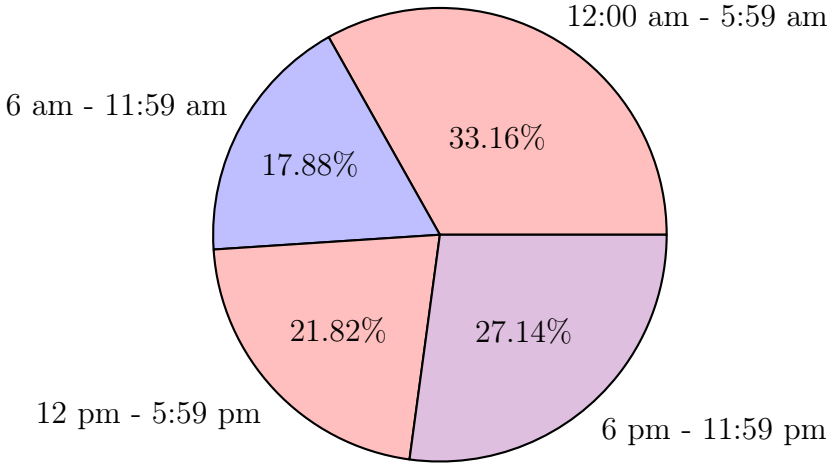


Figure 3.1: Percentage of choice incidents for urgency room by time of the day.

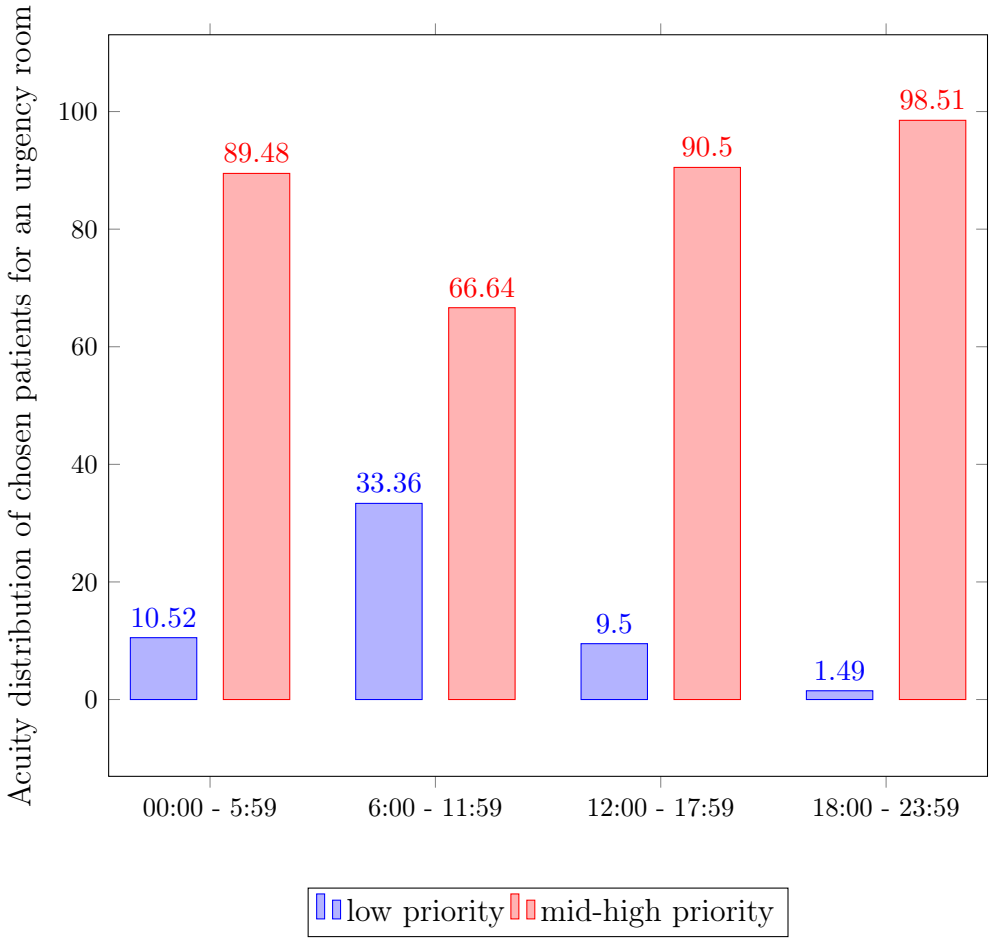


Figure 3.2: Acuity distribution of chosen patients in urgency room choice incidents by the time of the day.

status of the ED represents the decision-maker. The decision-maker selects the patient who brings the most utility to the present ED status. The empirical evidence from our data analysis and the literature suggests that the utility of each alternative depends not only on its attributes but also on interactions between its attributes and the decision-maker attributes. For instance, the likelihood of selecting a low acuity patient instead of a high acuity patient is higher between 6 am and 12 pm than during the rest of the day at UCM ED. At the Canadian hospital, ED studied by Li et al. [2021], the likelihood of selecting a Discharge patient rather than an Admit patient increases as the ED blocking increases.

Theory-driven models, e.g., logit models, assume an additive and linear utility structure rather than learning it from the data. This inflexibility can be restricting to model the observed choice behavior in complex environments such as an ED, where system attributes impact the choice outcome. The interaction terms between system and patient attributes can be added to capture the system effects on the choice outcome. This method requires either an a priori hypothesis formed by the domain knowledge or the addition of all pairwise interaction terms. Although the former can result in more interpretable estimation results, the interaction effects unknown to the domain expert can be missing. The latter can capture the most relevant interactions, but it also results in estimation results that are harder to interpret. Both methods capture the system and patient interactions in a very specific form, multiplication, and thus might not be flexible enough for a more complicated form of interactions.

Machine learning offers techniques and practices that could overcome the limitations of the current theory-driven models in the choice modeling field (van Cranenburgh et al. [2021]). For example, the decision tree approach, one of the most popular supervised learning algorithms in machine learning, is flexible enough to capture the interaction effects without explicit modeling and is still easy to interpret and visualize. We utilize decision trees and logit models to discover the system and patient interaction terms and study the impact of ED status on the patient prioritization rule.

Aouad et al. [2019] recommend a tree-based segmentation algorithm, *Marketing Segmentation Trees* (MST). The MST algorithm is a modified version of the CART algorithm, the standard decision tree generation approach proposed by Breiman et al. [2017], and performs

successive decision tree splits into the consumer attributes. It fits a choice model at each leaf node. We first evaluate the performance of the MST algorithm in an artificial choice dataset generated via simulating a simple ED system where all patient routing decisions are based on an optimal deterministic policy. The rationale behind this step is that any plausible algorithm should accurately predict the choice outcomes in this setting. We observe that the performance of the MST algorithm is comparable to the conditional logit model with all pairwise system-patient interactions, and both approaches can predict the choice outcomes with almost perfect accuracy.

The rest of this chapter is organized as follows. Section 3.2 first reviews relevant papers on patient routing in emergency departments and then provides a brief literature review on customer segmentation. In Section 3.3, we model the patient prioritization problem using the discrete choice framework and introduce the MST algorithm. In Section 3.4, we simulate the patient routing decisions at UCM and generate an artificial choice dataset on which we run the MST algorithm.

3.2 Literature Review

Choice Modelling in ED Operations: We first summarize three recently published papers that model patient routing as a discrete choice problem.

Ding et al. [2019] analyze patient prioritization decisions in an emergency department that uses the Canadian Triage and Acuity Scale (CTAS). CTAS proposes a fractile response objective for each urgency class but does not provide clear guidelines on routing patients within the same urgency level. The authors model the ED as a multiclass queuing system and observe that the routing decisions generally follow the "$c\mu$" rule proposed by Van Mieghem [1995]. They estimate the ED patient marginal cost structure perceived by decision-makers using the conditional logit model proposed by McFadden et al. [1973]. Their setting focuses on the primary service area, where rooms are similar when the system is critically loaded during a busy period. Thus, they justify that decision-makers are homogeneous. Our model focuses on primary (better suited to treat patients with high urgency) and secondary service areas when a choice incident appears without restricting to a specific time window. The

study data shows that not all room types are available all day, so an hour of the day also matters in routing decisions.

Li et al. [2021] analyze the average waiting time for Admit and Discharge patients with middle-to-low urgency levels at a large Canadian hospital. They observe that Discharge patients have a shorter waiting time than Admit patients within the same triage level. The authors use a mixed logit model to model patient routing decisions for the primary service area during peak hours. They add the related interaction terms to the model to capture how the expected disposition outcome affects the prioritization decision across triage and ED blocking levels.

The intensive care units (ICUs) generally include ED patients, transfers from other ICUs (planned and unplanned), and scheduled patients. Shen et al. [2020] observe that the ICU admission rate from ED varies as ED and ICU occupancy varies but is not sufficient to explain the heterogeneity in the admission rates. The authors hypothesize that forward-thinking behavior can impact admission decisions and find confirming evidence in the data. They deploy a multinomial logit model to estimate ICU admission decisions. They divide the time into 2-hour intervals. There are three choices for each patient: admission to ICU, admission to the non-ICU unit, or continuing to wait in ED. Both system and patient-level characteristics are used as covariates in the MNL model.

Customer Segmentation: The discrete choice framework allows us to redefine the patient prioritization problem in terms of marketing terminology. The ED status at each decision epoch represents the customer, and the patients in the waiting area represent the assortment. In marketing literature, customers are segmented based on their shared attributes to better understand their choice behaviors. Similar to customer segmentation, ED attributes can be segmented so that a similar decision rule is observed in ED environments belonging to the same segment.

A popular approach in the industry to perform customer segmentation is to cluster patients based on their attributes before any choice model is fit. An unsupervised algorithm, K-clustering, is commonly used to form K clusters so that the customers who belong to the same cluster are as homogeneous as possible in their attributes.

A more recent approach in customer segmentation is decision trees, a supervised learning

algorithm. Decision trees are heavily studied for classification and regression problems in machine learning literature (Breiman et al. [2017]). Decision trees are easy to interpret and capture complex non-linear relationships; however, they could be unstable due to their high variance nature. Combining decision trees with stable models such as linear and logistic regression could help mitigate the model instability and improve performance. Quinlan et al. (Quinlan et al. [1992]) introduce the "model tree" algorithm based on the idea of decision trees with leaves that contain linear regression functions. Landwehr et al. (Landwehr et al. [2005]) extend this idea for classification problems by replacing linear regression with logistic regression. The splitting criterion for Quinlan et al. and Landwehr et al. is based on class purity. Zeileis et al. [2008] focus on the decision trees' interpretability, proposes a stability metric, and splits the nodes based on parameter stability.

To our knowledge, Mišić [2016] is the first to use model trees for the customer segmentation process. The author uses a CART-based decision tree algorithm to partition the customers based on their attributes and fits a multinomial logit model (MNL) at every leaf. The objective is to minimize the log-likelihood to improve the predictive accuracy. Aouad et al. [2019] propose a similar decision tree generation process where the MNL is replaced with the conditional logit model (CLM). Replacing CLM with MNL offers a more general framework that could be applied to settings where choice alternatives are not explicitly categorized. The authors test the algorithm on synthetic datasets and a publicly available Expedia dataset. The synthetic datasets include the choice alternatives comprised of 2 to 5 options. In the Expedia dataset, each hotel represents the choice alternative. In both datasets, alternatives appear in more than one choice incident, and an alternative appears only once in a given choice incident. For example, the same hotel can appear on the display page for different Expedia users. The total number of hotels on the Expedia website is countably finite. In contrast, there is no natural categorization of patients in the ED dataset; thus, the set of all choice alternatives - patients - is not countable finite. Our work is the first to deploy the model tree in the ED setting with infinite choice alternatives to the best of our knowledge.

3.3 Methodology

We model the patient prioritization problem in ED as a discrete choice problem. Patients waiting to be roomed represent the choice set, while the current status of the ED system represents the decision-maker. The patient and system-level covariates determine the utility of choosing a patient at a given ED status. The triage nurse chooses the patient with the highest utility for the next bed assignment among all patients in the waiting room.

We will first explore the traditional logit models to model the prioritization rule in ED. Logit models are derived from random utility maximization theory and assume that utility function is a linear weighted function of observable covariates and an error term. They are widely used for discrete choice modeling where the choice set has two or more discrete alternatives. The choice probabilities are expressed in an interpretable closed form, making logit models attractive for researchers.

The standard logit models, multinomial logit, and conditional logit model are suitable for discrete choice modeling when the Independence of Irrelevant Alternative (IIA) holds. This property implies that the relative probability of choosing an alternative does not depend on other alternatives. Suppose that a patient is deemed more important than another patient. Even if we add more patients similar to the former, they will still be more important than the latter. Therefore, we assume that IIA property holds in the ED setting.

In the multinomial logit model, the utility of alternatives depends solely on the decision-maker attribute values. Each alternative has a different coefficient vector which measures how the decision-maker attribute value impacts the utility of that alternative. On the other hand, in the conditional logit model, the utility of the alternative depends on its attributes. The coefficient vector is allowed to be the same across all alternatives and measures the marginal utility of an alternative attribute.

Conditional logit is better suited than the multinomial logit to model patient routing decisions. To use the MNL framework, we need to cluster patients so that patients belonging to the same cluster will have the same or nearly the same attribute covariates. Even if we could cluster patients successfully, there is no guarantee that only one patient from each category will appear at any choice incident. Therefore, we eliminate the multinomial choice

model as a modeling approach. The simplest conditional logit model assumes that the utility function is a weighted linear function of alternative characteristics. Weights, i.e., coefficients, are estimated via maximum likelihood estimation with no need to cluster patients. It is possible to add interaction terms between the decision-maker and patients to capture the impact of decision-maker covariates on the choice probabilities.

Now, we explain how we constructed the dataset to apply the CLM model for patient prioritization. In our terminology, incident refers to the decision epoch. If two patients are in the system at an incident, two rows are associated with the incident in the dataset. Each row stores the information regarding system status and patient covariates at the time of that incident. Since both patients are experiencing the same system at the same decision epoch, columns related to system status have the same values. The columns related to patients' covariates store the information about the patient at that incident.

At each incident $t \in \mathbb{T} = \{1, \ldots, T\}$, the system (ED) is represented by the vector $\mathbf{z}_t = \langle z_{t1}, \ldots, z_{tM^s} \rangle$, and each choice alternative (patient) i in choice set C_t is represented by vector $\mathbf{x}_i = \langle x_{i1}, \ldots, x_{iM^a} \rangle$, where M^s and M^a are the number of system and patient level covariates, respectively.

Patients stay in the choice set until they get roomed in ED or depart from the ED. Therefore, the same patient can appear multiple times in the choice dataset. However, a subset of the patient attributes could change across these decision epochs. For example, a patient's health can worsen or improve, resulting in a change in the ESI score. The waiting time increases as the patient stays in the waiting room and transfers to the next choice set. Therefore, each patient-incident pair represents an alternative that appears only once in the dataset. We let $\mathbb{I}$ to be the set of all alternatives in the dataset, i.e, $\mathbb{I} = C_1 \cup C_2 \cup \cdots \cup C_T$.

The notation V_{ti} denotes the utility obtained by choosing alternative i at incident t. The decision-maker chooses the alternative with the highest utility. Let $c_t \in C_t$ be the actual choice outcome at decision epoch t, and $y_{ti} \in \{0, 1\}$ denote whether the alternative i is chosen at at incident t:

$$y_{ti} = \begin{cases} 1, & \text{if decision-maker chooses alternative } i \text{ at incident } t, \text{ i.e., } c_t = i \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

The conditional logit model assumes that utility is a linear function of covariates. Since all patients experience the same ED at the same incident, we cannot add ED-level covariates in the utility function unless we interact those covariates with patient-level covariates. The utility V_{ti} in the conditional logit model with no interaction terms is given by:

$$V_{ti} = \sum_{k=1}^{M^a} \beta_k x_{i,k} + \epsilon_{ti}. \tag{3.2}$$

The utility V_{ti} in the conditional logit model with all pairwise system-patient interaction terms is given by:

$$V_{ti} = \sum_{k=1}^{M^a} \beta_i x_{i,k} + \sum_{k=1}^{M^a} \sum_{m=1}^{M^s} \beta_{k,m} x_{i,k} z_{t,m} + \epsilon_{ti}. \tag{3.3}$$

The noise terms ϵ_{ti} are Gumbel random variables i.i.d across incidents and alternatives. The probability that decision-maker chooses alternative i at incident i is given by

$$\mathbb{P}\left(y_{ti} = 1\right) = \frac{e^{V_{ti}}}{\sum_{i' \in C_t} e^{V_{ti'}}} \quad \forall i \in C_t. \tag{3.4}$$

We use the “Pylogit" package developed by Brathwaite and Walker [2018] on Python to implement the conditional logit model.

Instead of adding all pairwise system-patient interaction terms in the form of multiplication, we use decision trees to capture the interactions. We split training observations using a decision tree and fit a conditional logit model at each leaf node. In this tree, splits are based on the system-level ED attributes. Therefore all observations belonging to the same incident fall into the same leaf. Leaf nodes represent the system status. System status is similar for the incidents in the same leaf; thus, we expect a similar decision rule. As the system status changes, decision rules can change. Room type, time of the day, and waiting room census can change the prioritization rule; accordingly, it is likely to run the CLM model at each leaf with the same covariates but obtain different coefficient estimates and thus different decision rules.

In the next two subsections, we give an overview of the decision trees and then introduce the MST algorithm recommended by Aouad et al. [2019].

3.3.1 Decision Trees Overview

Decision trees are created using a top-down, greedy approach that splits the predictor space into non-overlapping regions. A split, a branch node, is defined by a predictor and cut point. A decision tree splits the predictor space into two regions at the top of the tree and iteratively splits the previously defined regions until a stopping criterion is reached. At each step of the branching process, the algorithm picks the best split that minimizes an error function (e.g., mean squared prediction error for regression, classification error rate, the Gini index, or entropy for classification) within the resulting tree. The predicted value of an observation is determined by its leaf membership. The predicted value is the mean of all training observations that fall into that leaf for regression problems. For classification problems, the predicted value is the most commonly occurring class among the training observations that fall into that leaf (James et al. [2014]).

Decision trees are likely to overfit the training data due to the complexity of the resulting tree. To mitigate this problem, we can only continue to branch the tree if there is a split that decreases the error function by at least a certain predefined amount. Although this method can decrease tree complexity, it is too myopic since powerful splits can be hidden behind weaker splits. This issue can be avoided by a two-step process that includes growing a very deep tree and pruning it back to a subtree by penalizing the tree's complexity (James et al. [2014]).

3.3.2 The MST algorithm

The MST algorithm is a modified version of the Classification and Regression Trees (CART) algorithm presented by Breiman et al. [2017], which involves two main steps: tree growing and tree pruning. The MST algorithm performs successive decision tree splits into the consumer attributes and fits a choice model at each leaf node.

The algorithm starts with the root node, which includes all consumers, and considers a split encoded by a splitting variable and split point to partition the data. The split values are limited to a small set for categorical variables such as ESI score and abnormal vital sign indicators. All possible values of a categorical variable are considered as a split

point. Searching for the best split over all possible values can be computationally expensive for continuous variables. Therefore, the observed values of the continuous variable in the training data are sorted, and various quantile values are considered a split point. Once the data is split into two leaf nodes, a choice model is fit to model the consumers' response at each leaf. The choice model is trained to maximize the log-likelihood function. The performance of the split is measured as the summation of the log-likelihood functions of the best choice models fit at newborn leaves. The algorithm selects the best split and applies this branching procedure recursively until a stopping criterion is reached.

Once the algorithm completes the tree growing stage, it prunes back the fully grown tree to a subtree using the CART pruning technique to prevent overfitting. This technique evaluates the performance of the tree on the held-out validation set and prunes it if the marginal benefit of the split is below a certain threshold. The reader can refer to Aouad et al. [2019] for an in-depth explanation of the MST algorithm.

3.4 ED Simulation

We model the ED environment as a Markov Decision Process and find the optimal policy that minimizes the expected infinite horizon discounted waiting costs via value iteration. Given the system state, the optimal decision is deterministic. Once we have the system parameters and optimal policy, we can simulate the ED and decision process and create an artificial choice set. We first evaluate the performance of algorithms on the artificial choice set. Since there is no randomness in the choice decisions in the controlled environment, we expect that any candidate algorithm should closely mimic the optimal deterministic policy and have perfect or nearly perfect accuracy.

Let us describe the simulated ED system in more detail. There are three patient types with different urgency levels: urgent (1), moderately urgent (2), and non-urgent (3). Patient arrival and treatment are time-homogeneous Poisson processes with the rate λ_b and μ_b, respectively, where b denotes the urgency level of the patient. Parameter h_b denotes the waiting cost of a patient with urgency level b. Higher urgency patients have higher waiting costs than lower urgency patients.

There are two different types of rooms: urgency rooms and non-urgency rooms. There are N^u urgency rooms and N^n non-urgency rooms. Urgent patients can be treated only in urgency rooms. Moderately urgent patients can be treated in either room, but if a moderately urgent patient is treated in a non-urgent room, a linear cost with w is incurred. Non-urgent patients can be treated in either room at no additional cost. Patients will leave the system only after their treatment is completed, i.e., there is no abandonment.

While waiting, urgent and non-urgent patients do not switch to a different urgency class. However, moderately urgent patients can switch to higher or lower urgency classes while waiting. They can become urgent after an exponentially distributed waiting time with a mean of $1/\theta^u$ and become non-urgent after an exponentially distributed waiting time with a mean of $1/\theta^n$. Figure 3.3 illustrates the simulated ED environment.

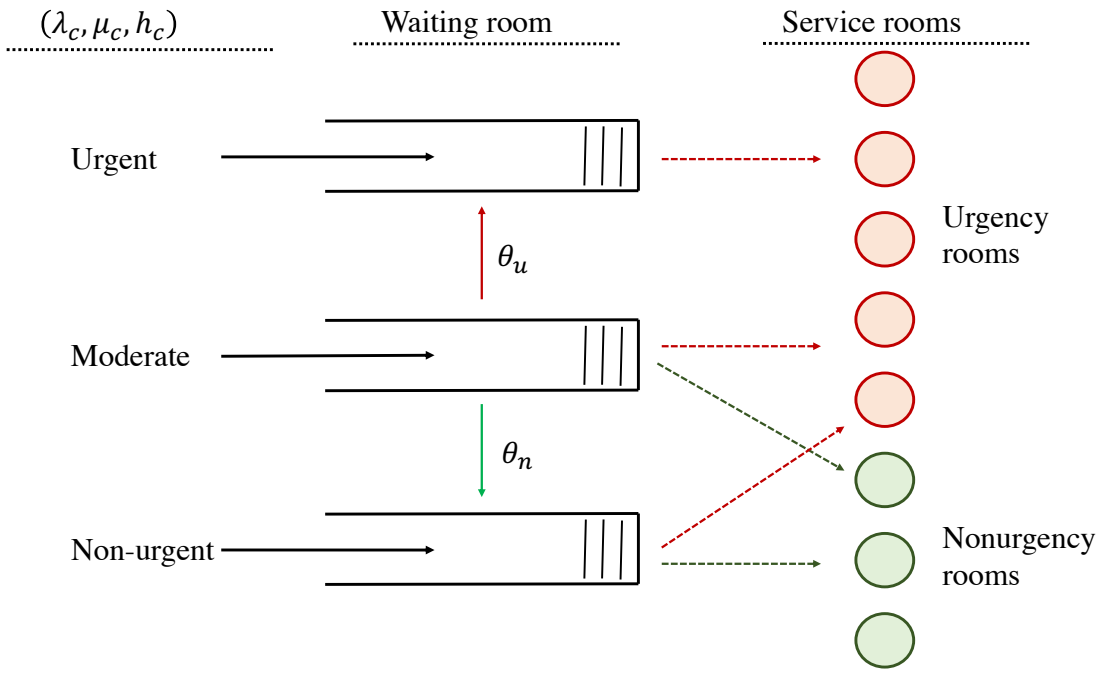


Figure 3.3: Illustration of ED simulation

The system's control times are when a patient arrives, treatment of a patient is completed, and a moderately urgent patient switches to a different urgency class. The state $s_t = \langle x_{1,t}, x_{2,t}, x_{3,t}, N_t^u, N_t^n, y_t \rangle$ where $x_{b,t}$ denotes the number of patients with urgency level $b \in \{1, 2, 3\}$ waiting for a room at decision epoch t; N_t^u (N_t^n) is the number of available urgency and (non-urgency rooms) at decision epoch t; and y_t is the number of moderately

urgency patients in non-urgency beds. The action is to decide how many patients with level b to send urgency rooms and non-urgency rooms. Let $a^u_{b,t}$ be the number of patients with level b to send urgency rooms where $b \in \{1,2,3\}$; and $a^n_{b,t}$ be the number of patients with level b to send non-urgency rooms where $b \in \{2,3\}$ at decision epoch t. The action space for state s_t is:

$$\begin{aligned} A(s_t) = \Big\{ \langle a^u_{1,t}, a^u_{2,t}, a^u_{3,t}, a^n_{2,t}, a^n_{3,t} \rangle : & \; 0 \le a^u_{1,t} \le x_{1,t}, \\ & \; 0 \le a^u_{2,t} + a^n_{2,t} \le x_{2,t}, \\ & \; 0 \le a^u_{3,t} + a^n_{3,t} \le x_{3,t}, \\ & \; a^u_{1,t} \le N^u_t, \\ & \; a^u_{1,t} + a^u_{2,t} + a^u_{3,t} \le N^u_t, \\ & \; a^n_{2,t} + a^n_{3,t} \le N^n_t, \\ & \; a^u_{1,t} + a^u_{2,t} + a^u_{3,t} + a^n_{2,t} + a^n_{3,t} \le 1 \Big\}. \end{aligned} \tag{3.5}$$

The last inequality ensures that only a single patient is routed at each decision epoch. This condition naturally holds when the number of occupied rooms is close to its capacity. We enforce this constraint, not for optimization purposes but to create a well-defined choice set. To limit the state space size, we will assume that if the total number of patients in the waiting room and servers is M, any patient arrival is rejected. Once the decision a is made, the state $s_t = \langle x_{1,t}, x_{2,t}, x_{3,t}, N^u_t, N^n_t, y_t \rangle$ immediately transitions to an intermediate state:

$$\begin{aligned} T(s_t, a_t) = \big\langle & x_{1,t} - a^u_{1,t}, \\ & x_{2,t} - a^u_{2,t} - a^n_{2,t}, \\ & x_{3,t} - a^u_{3,t} - a^n_{3,t}, \\ & N^u_t - a^u_{1,t} - a^u_{2,t} - a^u_{3,t}, \\ & N^n_t - a^n_{2,t} - a^n_{3,t}, \\ & y_t + a^n_{2,t} \big\rangle. \end{aligned} \tag{3.6}$$

The cost rate function for state $s_t = \langle x_{1,t}, x_{2,t}, x_{3,t}, N_t^u, N_t^n, y_t \rangle$ is given by

$$C(\langle x_{1,t}, x_{2,t}, x_{3,t}, N_t^u, N_t^n, y_t \rangle) = \left(\sum_{b=1}^{3} h_b x_{b,t} \right) + w y_t.$$

Let $V(s)^\pi$ denote the expected expected infinite-horizon discounted costs under the policy π where the initial state is $s_0 = s$:

$$V(s) = \mathbb{E}_{s^\pi} \left[\int_{t=0}^{\infty} e^{-\alpha t} C(s_t^\pi) dt \right],$$

where s_t^π is the random variable that denotes the state at time t under policy π.

We use the uniformization method to approximate the described continuous MDP by a discrete MDP. We find the optimal policy by using value iteration. Since optimal policy does not distinguish between patients with the same urgency level, we assume that First-Come-First-Served (FCFS) is applied at each urgency level.

3.4.1 Numerical Example

Once we have the model and optimal policy, we simulate the ED and collect choice data. Table 3.1 is an example of a choice dataset formatted in the long format for illustration purposes. We run 4 algorithms on the simulated data:

- CLM-1: Conditional Logit Model with patient-level covariates: urgency (encoded as a dummy variable), wait time.
- CLM-2: Conditional Logit Model with patient-level covariates and interaction terms between urgency and urgency room indicator.
- CLM-3: Conditional Logit Model with patient-level covariates and all pairwise interactions between patient and system-level covariates.
- The MST Algorithm splits the decision tree based on system-level attributes and fits a CLM model at each leaf by using only patient-level attributes.

Table 3.1: Simulated choice dataset in the long format.

incident	1	1	...	N	N	N
altkey	1	2	...	3	4	5
urgency room indicator	1	1	...	0	0	0
# urgent patients in waiting room	0	0	...	1	1	1
# moderately urgent patients in waiting room	1	1	...	2	2	2
# non-urgent patients in waiting room	1	1	...	2	2	2
# urgent patients in urgency rooms	2	2	...	2	2	2
# moderately urgent patients in urgency rooms	1	1	...	2	2	2
# non-urgent patients in urgency rooms	0	0	...	1	1	1
# urgent patients in non-urgency rooms	0	0	...	0	0	0
# moderately urgent patients in non-urgency rooms	1	1	...	0	0	0
# non-urgent patients in non-urgency rooms	1	1	...	1	1	1
# available urgency rooms	2	2	...	0	0	0
# available non-urgency rooms	1	1	...	2	2	2
urgency	2	3	...	1	2	3
wait time	1	3	...	1	2	3
choice	1	0	...	0	0	1

We evaluate the algorithm performance by measuring the percentage of the times that the patient chosen by the algorithm is the same as the patient chosen at the time of the incident. Based on the patient routing restrictions, adding interaction terms between the room type and urgency level to the CLM-1 should significantly improve the prediction accuracy. Furthermore, there might be system-patient interactions that are not as obvious but can significantly improve prediction accuracy; thus, CLM-3 should perform better than the CLM-2 in a sufficiently large dataset so that overfitting issues can be avoided.

Table 3.2 indicates that most of the variation in the choice outcomes that cannot be explained solely on patient attributes could be explained by the interaction between the urgency level and urgency room indicator. The other pairwise system-patient interactions could almost capture the remaining variation. Table 3.2 shows that the MST algorithm performs better than CLM-1 and CLM-2 and is comparable to CLM-3. Since CLM-3 includes all pairwise system-patient interactions, it would be harder to interpret. On the other hand, the decision tree generated by the MST algorithm illustrates how the patient prioritization rule can change based on the system state in a clear way and makes it easier to interpret the coefficient estimates of CLM models fit at each leaf node. Furthermore, a systematic and

sound comparison of coefficient estimates helps us to discover not-so-obvious system-patient interactions. Later, we can use the newly discovered interactions in the CLM model and assess their significance in explaining the variation in choice outcomes. The next chapter uses the MST algorithm and the described approach to gain insights into the patient prioritization decisions at UCM ED.

Table 3.2: Prediction accuracy of conditional logit models and the MST algorithm at different waiting room census levels on the artificial dataset.

no. waiting	CLM-1	CLM-2	CLM-3	MST	no. incidents
2	75.49%	93.75%	100.0%	100.0%	1681
3	75.75%	92.88%	98.97%	99.06%	1068
4	77.12%	94.31%	99.33%	99.11%	896
5	76.65%	94.49%	99.34%	99.21%	762
6	73.13%	92.82%	98.78%	98.78%	655
7	73.64%	94.89%	99.36%	99.04%	626
8	75.52%	95.66%	99.31%	98.44%	576
9	69.59%	92.65%	98.98%	98.37%	490
10	67.07%	89.27%	99.27%	98.54%	410
11	67.08%	85.64%	99.01%	99.50%	404
12	66.00%	83.25%	99.25%	98.50%	400
13	54.84%	67.74%	100.0%	100.0%	31
14	80.00%	80.00%	100.0%	100.0%	5

CHAPTER 4
APPLICATIONS TO THE UNIVERSITY OF CHICAGO MEDICINE EMERGENCY DEPARTMENT

4.1 Data Description

We study a dataset that includes all adult patients with emergency encounters from November 1, 2016, to December 15, 2017, at the UCM. There are 31 rooms in the primary service area, which we call urgency rooms, that are sufficiently equipped to treat patients with medium-to-high urgency levels. There are five rooms, which we call nonurgency rooms, preferable for patients with low urgency levels. The urgency rooms are available all day, while the nonurgency rooms are available from 12 pm to 6 am the next day.

Each patient encounter includes clinical information and operational event timestamps. Clinical information includes the patient's ESI score, vital signs (pulse, respiratory rate, blood oxygen level, blood pressure, temperature), age, gender, arrival mode, chief complaint, pain score, and level of consciousness. Main operational event timestamps include:

- Arrival time to the ED.
- Triage start and completion time.
- Initial provider contact time.
- Room-time (time at which patient roomed in ED).
- Disposition time (when the provider decides whether to admit or discharge the patient).
- Bed request time for Admit patients.
- Departure (the time when the patient leaves the ED).

The choice set comprises patients in the waiting area to be roomed. A patient enters the waiting area if their triage is completed. If the patient has a missing triage completion time, we assume the triage completion date is the same as the arrival date. Patients depart from the waiting area if they are roomed or leave without being seen/roomed. Once we determine

which patients are in the waiting area at the time of the incident, we determine the most recent clinical and operational information available to the decision-maker. For instance, patients' health might get worse or better, resulting in a change in the assigned ESI score while waiting. We assume the decision-maker decides which patient to select based on the most recent information.

4.1.1 Patient-Level Data Description

We assume that the decision-maker selects which patient to room based on the following patient-level attributes.

- ESI score. Based on clinical information and resource needs, patients are categorized into five urgency classes, with ESI-1 being the most urgent and ESI-5 being the least urgent. We use the ESI score as a categorical variable in our analysis.

- Wait. Wait time is calculated as the time elapsed from triage completion until the time of the incident.

- Age. We use age as a categorical variable by grouping patients into six categories: (18-45 years), (46-55 years), (56-65 years), (66-75 years), (76-85 years), and (86 years and over).

- Vital signs. There are four primary vital signs: pulse, respiratory rate, blood pressure, and temperature. We categorize the vital signs based on whether they fall into an abnormal range. An adult can have a pulse of 60-100 beats per minute (bpm). If the pulse rate exceeds this range, we encode it as an abnormal pulse. The normal respiratory rate for adults is 12-20 breaths per minute. A respiratory rate below 12 or over 20 breaths per minute indicates a serious health problem and is thus encoded as an abnormal respiratory rate. The normal adult body temperature can range from 97.8 °F to 99 °F. An abnormal body temperature refers to either hypothermia or fever. Hypothermia occurs when the body temperature drops below 95 °F (of Rochester Medical Center [2020]). Fever is not usually a concern if the body temperature reaches 103 °F or higher (Clinic [2020]). We categorize the body temperature as abnormal if it

drops below 95 °F or reaches 103 °F or higher. If the systolic blood pressure is over 180 or diastolic blood pressure is over 120, this is encoded as a hypertensive crisis. If the systolic blood pressure is between 130 and 179 or diastolic blood pressure is between 80 and 119, this is encoded as hypertension (Association [a]). Pulse oximetry (SpO2) is considered the "fifth vital sign". SpO2 measures the oxygen saturation in the blood. A drop in SpO2 below 92% is considered as dangerous (Association [b]); and thus we encode it as *abnormal SpO2*.

- Disposition outcome. Inspired by Li et al. [2021], we use the estimated probability of being admitted to an inpatient stay rather than the realized disposition outcome to avoid potential endogeneity issues. We train the prediction algorithm on the dataset that includes patient encounters from January 1, 2018, to October 31, 2019. This dataset has approximately 55000 encounters. We use age, gender, arrival mode, ESI score, vital signs, pain score, and level of consciousness as explanatory variables. We assign 80% of the sample to the training set and the remaining 20% to the test set. We implement several prediction models: random forests, gradient boosted trees, and logistic regression to predict the disposition outcome. The ROC AUC (area under the ROC curve) is around 0.82 for the prediction models.

- Special conditions. The dataset includes if a patient has the following special conditions: immunocompromised, at-risk, infection concern.

- Arrival by ambulance.

- Last ED visit date. Using operational data, we determine when the patient's last ED visit was before their current visit. If the last visit is within 30 days (90 days), the patient is encoded as 30-day readmission (90-day readmission).

- Seen by a provider. While the patient is still waiting to be roomed, a provider might initiate the first contact by ordering diagnostic tests. If the patient is contacted before the incident, this covariate takes the value of 1; 0 otherwise.

- Chief complaint. In our dataset, there are more than a thousand chief complaints; and most of these chief complaints occur for a very small number of patients. For

that reason, we categorize all chief complaints based on natural language and clinical information. Please refer to the Appendix A.2 for details.

Figure 4.1 shows the correlation between the patient-level attributes. Expected disposition outcome is strongly and positively correlated with patient urgency and age. Infection concern and immuno-compromised are moderately and positively correlated with each other. Correlation between other patient-level attributes seems to be small or none.

Correlation Matrix of Patient (Incident-Patient) Level Attributes

	esi	admit probability	ambulance	seen by provider	male	30 day readmission	90 day readmission	age	abnormal resp	abnormal SpO2	abnormal pulse	abnormal temp	hypertension	hypertensive crisis	immuno compromised	at_risk	infection concern	wait
esi	1.000																	
admit probability	-0.921	1.000																
ambulance	-0.076	0.151	1.000															
seen by provider	-0.081	0.075	0.006	1.000														
male	0.011	0.124	0.032	-0.009	1.000													
30 day readmission	-0.058	0.063	0.037	-0.008	0.016	1.000												
90 day readmission	-0.056	0.062	0.013	0.007	-0.014	-0.168	1.000											
age	-0.291	0.567	0.126	0.036	0.098	0.014	0.038	1.000										
abnormal resp	-0.135	0.139	0.031	0.008	0.007	0.019	0.020	0.059	1.000									
abnormal SpO2	-0.073	0.075	0.018	0.003	0.008	0.008	0.012	0.039	0.130	1.000								
abnormal pulse	-0.117	0.221	0.009	-0.003	0.006	0.026	0.017	-0.020	0.111	0.043	1.000							
abnormal temp	-0.049	0.054	-0.011	-0.008	0.017	-0.007	-0.006	-0.021	0.034	0.027	0.116	1.000						
hypertension	-0.007	0.004	0.005	0.006	0.067	-0.021	-0.012	0.189	0.021	-0.004	0.022	-0.020	1.000					
hypertensive crisis	-0.100	0.117	0.003	0.004	0.008	0.009	0.001	0.131	0.035	0.010	0.021	-0.005	-0.239	1.000				
immuno compromised	-0.123	0.126	0.000	0.023	0.043	0.006	0.028	0.063	0.018	0.015	0.043	0.029	-0.012	-0.007	1.000			
at_risk	-0.013	0.019	0.023	0.021	-0.003	-0.004	0.005	0.019	0.022	0.000	0.009	-0.001	0.004	-0.001	0.001	1.000		
infection concern	-0.099	0.109	0.002	0.017	0.036	0.030	0.020	0.045	0.025	0.023	0.063	0.127	-0.015	0.000	0.362	-0.001	1.000	
wait	0.043	-0.035	-0.003	0.056	-0.030	-0.008	-0.001	-0.004	-0.057	-0.033	-0.034	-0.012	-0.003	-0.014	-0.033	-0.001	-0.020	1.000

Figure 4.1: Correlation Matrix of Patient (Incident-Patient) Level Attributes

4.1.2 System-Level Data Description

We define the emergency department by using the following system-level attributes:

- Type of the room to which patient is routed at the time of the incident.
- Waiting room census.
- Clean room availability. All ED rooms must be cleaned after every patient encounter to maintain a sterile environment and reduce infection risk. We do not have the information on when the cleaning process starts and ends. We first estimate the average

time to clean a room based on the data and then calculate the estimated number of clean ED rooms at each decision epoch.

- Number of boarders. A boarding patient is defined as a patient who is admitted to an inpatient stay but has not been transferred to an inpatient bed because no beds are available. According to Li et al. [2021], the boarding starts when an inpatient bed is requested for the patient at the hospital and ends once the patient has departed the ED and transferred to the inpatient stay. We use their definition to calculate the number of boarders at each decision epoch.

- The time of the day. We divide the day into six-hour periods: from midnight to 6 am, from 6 am to noon, from noon to 6 pm, and from 6 pm to midnight.

- Whether or not the day associated with the incident is a holiday.

To estimate the average cleaning time, we focus on the busy period that many patients are waiting to be roomed. For each room, we find the median length of the time that starts with the existing patient's departure from the room and ends with a new patient's arrival in the room. During this time, the room's number of occupants will be zero. We will call this period a zero-state period. It is important to limit our analysis to the busy period. If a room is available during a busy period, we infer that the room is being disinfected and thus not utilized. During a non-busy period in that not many patients are waiting to be seen, a room will not necessarily be given to a patient immediately after cleaning. We observe that median zero-state time decreases as the number of patients in the waiting room increases. Figure 4.2 shows that the median zero-state time stabilizes around 0.3 hours. Figure 4.3 shows the correlation between system-level attributes. The waiting room census is positively correlated with the number of boarders but negatively correlated with the number of clean urgency and nonurgency rooms. The number of clean urgency and nonurgency rooms is positively correlated with each other while negatively correlated with the number of boarders. This is not surprising because as the ED gets busier, waiting room census and the number of boarders increases, and clean room availability decreases. Figure 4.4 illustrated how system attributes at UCM ED change throughout the day.

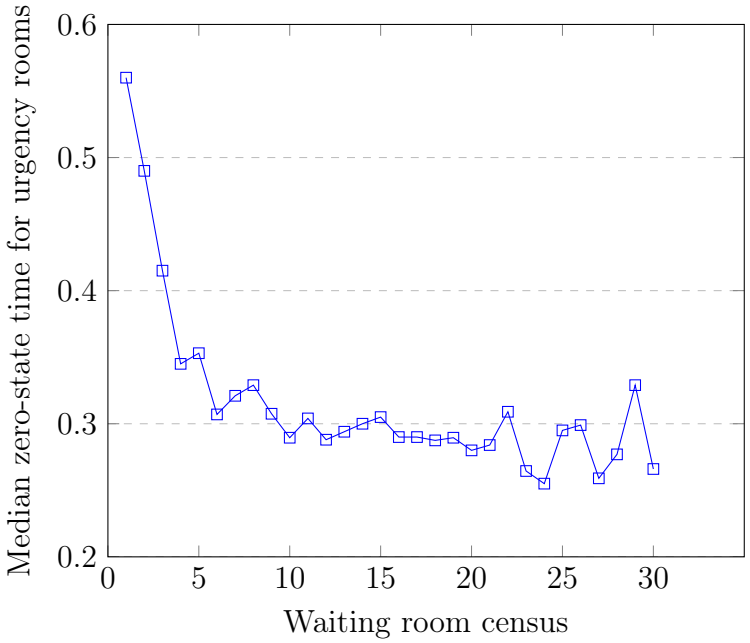


Figure 4.2: Median time from an outflux event to an influx event for an urgency room (excluding 6 am -12 pm)

4.2 Discrete Choice Model

At incident t, the decision-maker chooses to room a patient from choice set C_t. The choice set C_t is dynamic and changes over time as patients depart from and arrive in the waiting room. The choice set C_{t+1} comprised all patients in the choice set C_t minus patients, who were selected at incident t or left the waiting room due to other reasons, plus patients who arrived between incident t and $t+1$. Once we have the choice set data, we perform the following steps to make the discrete choice model better reflect the patient prioritization decisions.

1. If only a single patient is waiting at incident t, i.e., $|C_t| = 1$, we drop that incident from the choice data.

2. If there is a patient with a missing ESI score at incident t, we drop the incident t from the choice data.

3. We exclude incidents that include an ESI-1 patient (0.7% of encounters). Because ESI-1 patients have unstable medical conditions and require immediate care; thus, the

Correlation Matrix of System Level Attributes (from 12 pm to 6 am)				
	num waiting	clean avail urgent	clean avail nonurgent	num boarding
num waiting	1.000			
clean avail urgent	-0.455	1.000		
clean avail nonurgent	-0.341	0.212	1.000	
num boarding	0.542	-0.388	-0.242	1.000

Correlation Matrix of System Level Attributes (from 6 am to 12 pm)			
	num waiting	clean avail urgent	num boarding
num waiting	1.000		
clean avail urgent	-0.385	1.000	
num boarding	0.292	-0.257	1.000

Figure 4.3: Correlation Matrix of Patient System Level Attributes

prioritization rule is always to choose the ESI-1 patient.

4. If there is a patient at incident t with missing and unfilled entries, we exclude that patient from the choice set C_t. If the updated choice set size is zero or one, we drop the incident t from the choice data.

5. If the chosen patient at incident t has missing and unfilled entries, we drop the incident t from the choice data.

6. If there is a "not" chosen patient in C_t who gets roomed within X minutes of incident t, we exclude those patients from the choice set C_t. If the number of patients in the waiting room after the exclusion is zero or one, we drop the incident t from the choice data.

The rationale behind the last step is that room-time information serves as a proxy for the time that a prioritization decision is made. Suppose two patients, A and B, are triaged and waiting to be roomed. Suppose there are more than a couple of rooms available, and the triage nurses simultaneously assign these two patients to the ED beds. The most likely scenario is that these two patients are not roomed simultaneously but within a couple of minutes of each other. Therefore, we cannot always claim that one patient is prioritized over another, even if they roomed earlier than the other patient. To mitigate this problem, we should estimate the time it takes to transfer a patient from the waiting room to a room in the ED.

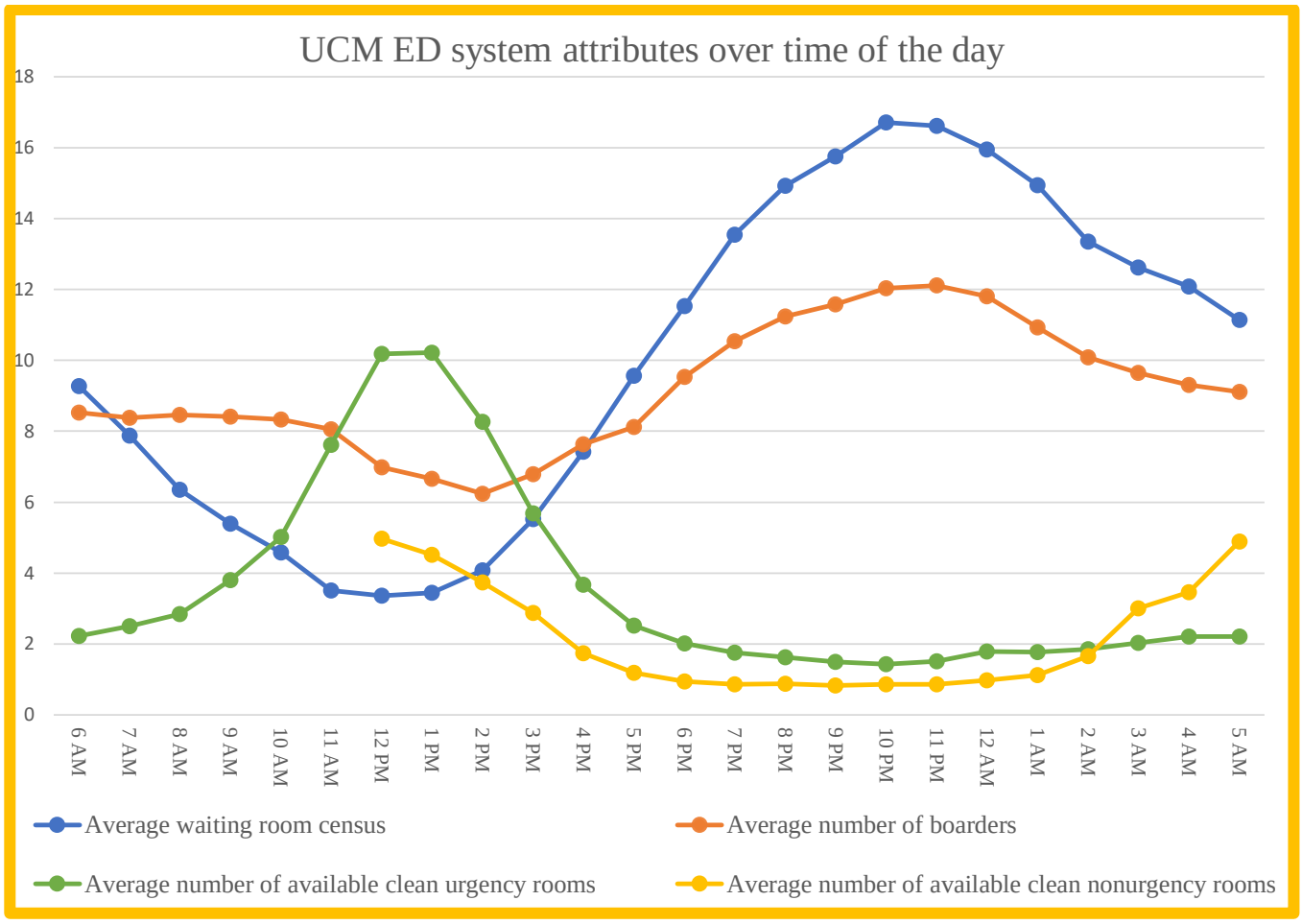


Figure 4.4: UCM ED system attributes over time of the day

The time from triage completion to room time includes the transfer time and the time spent waiting for patients deemed a higher priority to be roomed in the ED. If the patient is the first person to be roomed after they are triaged, we conclude that there are no patients deemed a higher priority than them. That is why the waiting time should only include the transfer time. We first identify the patients who never see another patient to be roomed during their stay in the waiting room. Taking the median of the wait times across these patients, we estimate the transfer time as 5 minutes.

Once we update the choice set C_t per steps described above, we apply the conditional logit model. The decision-maker chooses alternative i at incident i with probability

$$\mathbb{P}\left(y_{ti}=1\right)=\frac{e^{V_{ti}}}{\sum_{i'\in C_t}e^{V_{ti'}}}\quad\forall i\in C_t.$$

Recall that alternative i refers to the patient-incident pair rather than the patient itself. In other words, the same patient can appear in different choice incidents, and we use a different label each time we refer to the patient. We first apply the CLM model with no interaction terms:

$$V_{ti} = \sum_{k=2}^{4} \beta_{\text{esi},k} x_i^{\text{esi},k} + \sum_{k=2}^{5} x_i^{\text{esi},k} \left(\beta_{\text{wait},k} \text{Wait}_i + \beta_{\text{wait}^2,k} \text{Wait}_i^2 + \beta_{\text{admit},k} x_i^{\text{admit}} \right) + \boldsymbol{\beta}_p^T \mathbf{x}_i^p. \tag{4.1}$$

Notation $x_i^{\text{esi},k} = 1$ if alternative i has an ESI score of k at time of the incident for all $k \in \{2, 3, 4\}$. The notation Wait_i is the time elapsed from alternative i's triage completion to the time of incident t. We add quadratic term Wait_i^2 to capture the nonlinear effect of waiting time on the decision maker's utility. Notation x_i^{admit} is the estimated probability that alternative i will be admitted into inpatient-stay. The vector $\mathbf{x}_i^p$ contains information on alternative i's arrival mode, gender, previous ED visit, age, vital signs, the existence of a special condition (such as substance abuse, altered mental status, etc.), chief complaint and whether a provider has seen the alternative.

Table 4.1: Estimation results for urgency and non-urgency rooms on UCM ED dataset.

	Urgency room		Nonurgency room	
No. observations	26606		6346	
McFadden pseudo R^2	0.259		0.510	
McFadden pseudo R^2_{adj}	0.256		0.499	
Variables	Coef	Std Err	Coef	Std Err
ESI = 2				
ESI	2.9587^{***}	(0.223)	0.5682	(0.983)
ESI × Wait	1.2039^{***}	(0.026)	0.7531^{*}	(0.352)
ESI × Wait^2	-0.1123^{***}	(0.005)	−0.0556	(0.072)
ESI × admit prediction	0.4078	(0.228)	-7.0589^{***}	(1.380)
ESI = 3				
ESI	1.1838^{***}	(0.181)	0.2414	(0.286)
ESI × Wait	1.1322^{***}	(0.020)	0.9212^{***}	(0.092)
ESI × Wait^2	-0.0750^{***}	(0.003)	-0.1400^{***}	(0.019)
ESI × admit prediction	−0.1731	(0.178)	-6.1436^{***}	(0.536)

Continued on next page

Table 4.1 – continued from previous page

Variables	Urgency room Coef	Std Err	Nonurgency room Coef	Std Err
ESI = 4				
ESI	−0.3160	(0.184)	−0.0645	(0.241)
ESI × Wait	1.0706***	(0.037)	2.1304***	(0.069)
ESI × Wait^2	−0.0630***	(0.005)	−0.2768***	(0.015)
ESI × admit prediction	−1.4070	(0.778)	−2.3128	(1.369)
ESI = 5				
ESI × Wait	0.6251***	(0.079)	2.2921***	(0.256)
ESI × Wait^2	−0.0333***	(0.008)	−0.4219***	(0.072)
ESI × admit prediction	−8.5222*	(3.903)	−8.0470	(5.767)
ambulance	0.1674***	(0.031)	−0.2402*	(0.118)
seen by provider	−0.8899***	(0.036)	−0.5798***	(0.070)
gender(base = female)	0.0613**	(0.021)	0.1916***	(0.050)
30-day ED readmission	−0.0912***	(0.021)	−0.1995**	(0.060)
90-day ED readmission	−0.0551**	(0.023)	−0.1681**	(0.060)
age (base = 18-45 years)				
46-55 years	0.0187	(0.031)	−0.0398	(0.075)
56-65 years	0.0950*	(0.038)	−0.0390	(0.099)
66-75 years	0.1569**	(0.047)	−0.0855	(0.140)
76-85 years	0.2868***	(0.058)	−0.2310	(0.198)
86+ years	0.4757***	(0.074)	−0.0337	(0.299)
abnormal Resp	0.4516***	(0.039)	0.0501	(0.248)
abnormal SpO2	0.3111***	(0.073)	0.4212	(0.359)
abnormal Pulse	0.1229***	(0.024)	0.0591	(0.060)
abnormal Temp	0.1173*	(0.055)	0.2694	(0.162)
hypertension	−0.0009	(0.018)	−0.0394	(0.041)
hypertensive crisis	0.0782*	(0.038)	−0.3804**	(0.144)
special conditions				
immunocompromised	0.3576***	(0.054)	−0.1275	(0.280)
at risk	2.1143***	(0.266)	1.0586	(1.662)
infection concern	0.1753***	(0.040)	0.1485	(0.119)
chief complaints (base = other)[a]				
abdominal pain	−0.1775***	(0.034)	−0.7322***	(0.093)
abnormal labs	0.2297***	(0.066)	−1.1251*	(0.489)

Continued on next page

Table 4.1 – continued from previous page

	Urgency room		Nonurgency room	
Variables	Coef	Std Err	Coef	Std Err
alcohol intoxication	0.3962^{**}	(0.136)	$-M$	M
allergic reaction	0.3426^{**}	(0.107)	0.5322^{*}	(0.237)
altered mental status	0.7704^{***}	(0.091)	$-M$	M
back pain	-0.1580^{**}	(0.059)	-0.0705	(0.093)
chest pain	-0.2249^{***}	(0.036)	-0.9584^{***}	(0.138)
drug abuse	0.7817^{***}	(0.267)	$-M$	M
extremity pain	-0.4096^{***}	(0.097)	0.0955	(0.163)
headache	-0.1650^{**}	(0.049)	-0.2217	(0.119)
leg pain	-0.2708^{***}	(0.066)	-0.0382	(0.118)
lower extremity edema	-0.2957^{**}	(0.067)	-1.0147^{**}	(0.341)
motor vehicle crash	-0.0239	(0.079)	-0.1455	(0.117)
possible stroke	1.1230^{***}	(0.163)	$-M$	M
psyc evaluation	0.5874^{***}	(0.156)	$-M$	M
respiratory distress	0.8523^{**}	(0.309)	$-M$	M
seizure	0.4696^{***}	(0.101)	-1.3500	(1.014)
shortness of breath	-0.1456^{***}	(0.040)	-0.9639^{***}	(0.195)
sickle cell pain	-0.6090^{***}	(0.062)	-3.5870^{***}	(1.007)
sore throat	-0.1573	(0.089)	0.2525^{*}	(0.107)
suicidal ideation	2.0898^{***}	(0.120)	$-M$	M
urinary problem	-0.3746^{***}	(0.095)	-0.3240^{*}	(0.137)
vaginal bleeding	-0.1743^{**}	(0.061)	-0.6295^{***}	(0.168)

[a] A subset of chief complaints are shown; $^{*}\, p < 0.05$; $^{**}\, p < 0.01$; $^{***}\, p < 0.001$; M is a varying large number

The MST algorithm discovers that the room type is the most important system-level attribute impacting patient prioritization. Therefore, we present estimation results separately for urgency and nonurgency rooms. Table 4.1 shows that the patient prioritization rule indeed differs based on the room type. Estimation results suggest that the decision-maker prioritizes low acuity patients by putting more weight on their waiting time costs for nonurgency rooms and prioritizes mid/high acuity patients over low acuity patients for urgency rooms. We have the following observations based on the estimation results on urgency rooms.

- The decision-maker prioritizes high acuity patients over mid/low acuity patients; prioritizes mid acuity patients over low acuity patients. Within the same acuity class, the First-Come-First-Served principle is followed.

- The coefficient of Wait and Wait^2 are significant at the level 0.001% across all ESI scores. Based on the coefficient estimates, we infer that wait time has an increasing and concave impact on the utility function.

- Patients with abnormal respiratory rate, pulse, SpO2, old age, and special conditions are more likely to be selected than those with normal vital signs and young/middle age.

- If a provider sees the patient, it is likely that the provider has ordered diagnostic tests and is waiting for results to initiate the treatment process, thus prioritizing other patients over patients with in-progress diagnostics.

- If the patient has an ED encounter within 90 days before their current visit, they might be considered a frequent patient who uses the ED as primary care and thus be deprioritized.

- Patients with chief complaints such as altered mental status, allergic reaction, abnormal labs, drug abuse, alcohol intoxication, possible stroke signs, psychiatric evaluation, respiratory distress, seizure, and suicidal evaluation receive priority to be placed in urgency rooms. On the other hand, patients with the most common chief complaints such as abdominal pain, chest pain, shortness of breath, back pain, and headache are deprioritized over patients with less common chief complaints, given that all other attributes are the same.

Next, we present the decision tree generated by the MST algorithm on the study data by controlling the room type as urgency rooms. The resulting tree is given by Figure 4.5. Afterward, we present the estimation results for urgency rooms at different waiting room census, time of the day, and clean nonurgency room availability and compare the estimation results to understand how the system status impacts the decision rule.

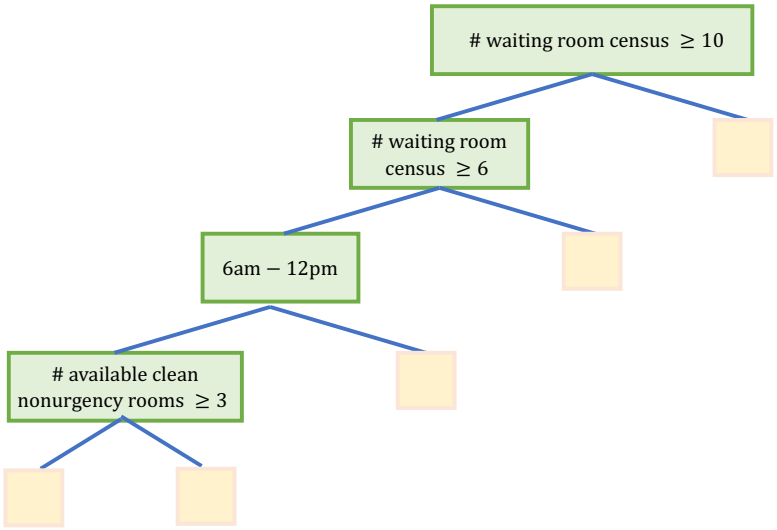


Figure 4.5: Decision tree on emergency department attributes controlling the room type = urgency room.

Table 4.2: Estimation results for urgency rooms at different waiting room census level I

	num-waiting $\in [2, 10)$		num-waiting ≥ 10	
No. observations	11681		14925	
McFadden pseudo R^2	0.250		0.288	
McFadden pseudo R^2_{adj}	0.241		0.284	
Variables	Coef	Std Err	Coef	Std Err
ESI = 2				
ESI	2.2831^{***}	(0.269)	5.5917^{***}	(0.241)
ESI × Wait	1.3927^{***}	(0.090)	1.0235^{***}	(0.029)
ESI × Wait^2	-0.2248^{***}	(0.025)	-0.0822^{***}	(0.005)
ESI × admit prediction	0.6537	(0.419)	0.3695	(0.270)
ESI = 3				
ESI	1.1187^{***}	(0.106)	2.8281^{***}	(0.193)
ESI × Wait	1.5860^{***}	(0.045)	1.2873^{***}	(0.031)
ESI × Wait^2	-0.1885^{***}	(0.008)	-0.0752^{***}	(0.004)
ESI × admit prediction	0.1288	(0.298)	−0.1646	(0.223)
ESI = 4,5				

Continued on next page

Table 4.2 – continued from previous page

	num-waiting < 10		num-waiting ≥ 10	
Variables	Coef	Std Err	Coef	Std Err
ESI × Wait	1.2227^{***}	(0.048)	1.6013^{***}	(0.074)
ESI × Wait^2	-0.1084^{***}	(0.008)	-0.0952^{***}	(0.008)
ESI × admit prediction	-2.3939^{*}	(1.139)	0.0920	(1.219)
ambulance	0.2188^{***}	(0.054)	0.1084^{**}	(0.038)
seen by provider	-1.1861^{***}	(0.074)	-0.8361^{***}	(0.041)
gender(base = female)	0.0770^{*}	(0.035)	0.0473	(0.026)
30-day ED readmission	-0.1918^{***}	(0.035)	-0.0496	(0.025)
90-day ED readmission	-0.0439	(0.039)	-0.0556^{*}	(0.028)
age (base = 18-45 years)				
46-55 years	-0.0283	(0.051)	0.0142	(0.038)
56-65 years	-0.0088	(0.065)	0.1068^{*}	(0.047)
66-75 years	0.1096	(0.083)	0.1430^{*}	(0.057)
76-85 years	0.2019	(0.104)	0.2748^{***}	(0.069)
86+ years	0.3721^{**}	(0.139)	0.4768^{***}	(0.087)

A subset of covariates removed for the sake of space; $^{*}\, p < 0.05$; $^{**}\, p < 0.01$; $^{***}\, p < 0.001$

Table 4.2 hints that the decision-maker follows the urgency-based prioritization rule more frequently if the waiting room is moderately and highly crowded. Suppose that all patient characteristics, excluding ESI score, are the same. If there are equal to or more than ten patients in the waiting room, the selection of an ESI-2 patient generates 2.7696 higher utility than the selection of an ESI-3 patient; the selection of an ESI-3 patient generates 2.8221 higher utility than the selection of a low acuity patient. If there are less than ten patients in the waiting room, the selection of an ESI-2 patient generates 1.1664 higher utility than an ESI-3 patient; the selection of an ESI-3 patient generates 1.1187 higher utility than the selection of a low acuity patient.

As the waiting room becomes moderately or highly crowded, the estimated coefficient of Wait^2 increases but is still negative, i.e., the impact of waiting time on the utility function becomes more linear but still concave. Suppose two patients are waiting to be roomed with the same covariates except waiting time. Assume, for instance, that the first patient is waiting for 30 minutes and the second patient is waiting for 60 minutes. The likelihood of selecting the first patient is higher than what it would be if the former and latter were

waiting for 120 minutes and 150 minutes, respectively. Suppose the waiting room is moderately/highly crowded. In that case, the probability of selecting the first patient in both situations is closer to each other than if the waiting room was less crowded.

Table 4.3: Estimation results for urgency rooms at different waiting room census levels II

	num-waiting $\in [2, 6)$		$\in [6, 10)$	
No. observations	6272		5409	
McFadden pseudo R^2	0.229		0.295	
McFadden pseudo R^2_{adj}	0.208		0.281	
Variables	Coef	Std Err	Coef	Std Err
ESI = 2				
ESI	2.0468^{***}	(0.405)	3.1295^{***}	(0.370)
ESI × Wait	0.8948^{***}	(0.149)	1.4969^{***}	(0.116)
ESI × Wait^2	-0.1069^{***}	(0.029)	-0.2728^{***}	(0.033)
ESI × admit prediction	−0.2241	(0.631)	1.0033	(0.566)
ESI = 3				
ESI	0.9379^{***}	(0.147)	1.6887^{***}	(0.172)
ESI × Wait	1.7008^{***}	(0.092)	1.7273^{***}	(0.058)
ESI × Wait^2	-0.2853^{***}	(0.023)	-0.1804^{***}	(0.010)
ESI × admit prediction	−0.0191	(0.444)	0.3058	(0.409)
ESI = 4 or 5				
ESI × Wait	1.2826^{***}	(0.074)	1.5231^{***}	(0.071)
ESI × Wait^2	-0.1598^{***}	(0.013)	-0.1133^{***}	(0.010)
ESI × admit prediction	-3.3034^{*}	(1.619)	−1.6740	(1.601)
ambulance	0.3216^{***}	(0.084)	0.1357	(0.073)
seen by provider	-1.4869^{***}	(0.118)	-0.9547^{***}	(0.096)
gender(base = female)	0.0091	(0.052)	0.1251^{**}	(0.048)
30-day ED readmission	-0.1162^{*}	(0.051)	-0.2477^{***}	(0.049)
90-day ED readmission	−0.0621	(0.059)	−0.0444	(0.053)
age (base = 18-45 years)				
46-55 years	0.0096	(0.075)	−0.0680	(0.071)
56-65 years	−0.0403	(0.095)	0.0108	(0.091)
66-75 years	0.0834	(0.124)	0.1410	(0.114)
76-85 years	0.2982	(0.160)	0.1261	(0.140)
86+ years	0.3156	(0.206)	0.3616	(0.188)

A subset of covariates removed for the sake of space; $^{*}\, p < 0.05$; $^{**}\, p < 0.01$; $^{***}\, p < 0.001$

Table 4.3 also hints that adherence to urgency-based prioritization rules increases as the waiting room census increases, even if the waiting room is less crowded than the busy hours. If less than six patients are waiting, the ambulance arrival significantly increases the utility of the patient. In comparison, if there are more than six patients, this impact either decreases

or vanishes.

Table 4.4: Estimation results for urgency rooms at different time of the day by controlling num-waiting $\in [2, 6)$

	time $\notin$ [6am,12pm)		time $\in$ [6am,12pm)	
No. observations	4532		1740	
McFadden pseudo R^2	0.314		0.147	
McFadden pseudo R^2_{adj}	0.284		0.069	
Variables	Coef	Std Err	Coef	Std Err
ESI = 2				
ESI	2.3597^{***}	(0.492)	0.3397	(0.900)
ESI × Wait	2.5632^{***}	(0.267)	0.3142	(0.209)
ESI × Wait^2	-0.6540^{***}	(0.112)	−0.0306	(0.020)
ESI × admit prediction	0.1349	(0.762)	0.8200	(1.347)
ESI = 3				
ESI	1.0888^{***}	(0.193)	0.6147^{*}	(0.284)
ESI × Wait	2.4405^{***}	(0.146)	1.2751^{***}	(0.131)
ESI × Wait^2	-0.4158^{***}	(0.048)	-0.2075^{***}	(0.026)
ESI × admit prediction	0.4926^{*}	(0.548)	-1.9977^{*}	(0.872)
ESI = 4 or 5				
ESI × Wait	1.4551^{***}	(0.137)	0.7685^{***}	(0.093)
ESI × Wait^2	-0.1478^{***}	(0.028)	-0.0993^{***}	(0.015)
ESI × admit prediction	-1.5284^{***}	(2.084)	-7.4387^{*}	(2.875)
ambulance	0.3431^{**}	(0.110)	0.3375^{*}	(0.145)
seen by provider	-1.8521^{***}	(0.139)	−0.5686	(0.312)
gender(base = female)	−0.0254	(0.066)	0.1067^{***}	(0.095)
30-day ED readmission	−0.0094	(0.065)	-0.3050^{**}	(0.091)
90-day ED readmission	0.0560	(0.073)	-0.3356^{**}	(0.116)
age (base = 18-45 years)				
46-55 years	0.0423	(0.094)	−0.0311	(0.138)
56-65 years	0.0729	(0.119)	−0.2110	(0.175)
66-75 years	0.1128	(0.154)	0.0589	(0.244)
76-85 years	0.1995^{**}	(0.192)	1.0063^{**}	(0.361)
86+ years	0.3928^{*}	(0.252)	0.5959	(0.411)

A subset of covariates removed for the sake of space; $^{*}\, p < 0.05$; $^{**}\, p < 0.01$; $^{***}\, p < 0.001$

Table 4.4 hints that the ESI scores are not as important as between 6 am and 12 pm compared to the rest of the day when there are not many patients in the waiting room. A plausible explanation for this observation is that the fast track is closed during this period.

Therefore, low acuity patients also need to be treated in the urgency rooms. Furthermore, not many patients are waiting to be roomed, and thus, prioritization becomes relatively less important.

Table 4.5: Estimation results for urgency rooms at different clean non-urgency room availability by controlling num-waiting $\in [2, 6)$ and time $\notin$ [6am,12pm)

	# available clean non-urgency rooms < 3		# available clean non-urgency rooms ≥ 3	
No. observations	2139		2393	
McFadden pseudo R^2	0.413		0.292	
McFadden pseudo R^2_{adj}	0.353		0.233	
Variables	Coef	Std Err	Coef	Std Err
ESI = 2				
ESI	3.6634^{***}	(0.742)	1.5804^{***}	(0.720)
ESI × Wait	3.2816^{***}	(0.466)	2.8510^{***}	(0.406)
ESI × Wait^2	-1.1646^{***}	(0.263)	-0.6079^{***}	(0.131)
ESI × admit prediction	1.0238	(1.157)	0.0296	(1.114)
ESI = 3				
ESI	2.5150^{***}	(0.286)	0.6051^{*}	(0.256)
ESI × Wait	2.5120^{***}	(0.215)	2.9108^{***}	(0.228)
ESI × Wait^2	-0.4454^{***}	(0.073)	-0.4771^{***}	(0.068)
ESI × admit prediction	1.4789	(0.858)	−1.0090	(0.770)
ESI = 4 or 5				
ESI × Wait	3.1174^{***}	(0.383)	1.4519^{***}	(0.162)
ESI × Wait^2	-0.6794^{***}	(0.144)	-0.1274^{***}	(0.032)
ESI × admit prediction	3.0830	(3.353)	−4.2825	(2.755)
ambulance	0.2761	(0.174)	0.3822^{*}	(0.153)
seen by provider	-1.7494^{***}	(0.189)	-2.3505^{***}	(0.245)
gender(base = female)	−0.0439	(0.106)	−0.0265	(0.092)
30-day ED readmission	−0.0192	(0.102)	−0.0097	(0.090)
90-day ED readmission	0.0204	(0.113)	0.0515	(0.102)
age (base = 18-45 years)				
46-55 years	-0.1303^{*}	(0.149)	0.2073	(0.130)
56-65 years	0.0265	(0.187)	0.0955	(0.167)
66-75 years	−0.2268	(0.241)	0.3114	(0.216)
76-85 years	−0.0756	(0.292)	0.3083	(0.275)
86+ years	0.4342	(0.388)	0.1962^{*}	(0.360)

A subset of covariates removed for the sake of space; $^{*}\, p < 0.05$; $^{**}\, p < 0.01$; $^{***}\, p < 0.001$

Table 4.5 suggests that the resource availability impacts the prioritization rule. If nonurgency room availability is limited, the likelihood of selecting a low acuity patient instead of

a mid/high patient with all other attributes being the same is less than what it would be if there were a large number of nonurgency rooms. These estimation results can be seen as counter-intuitive at first. One can expect that if there are not many available nonurgency rooms, the selection likelihood of low acuity patients should increase as in the case of Table 4.4 interpretation. However, there is no urgency room availability between 6 am and 12 pm. If a low acuity patient arrives at 10 am, they cannot be placed in an urgency bed until noon. When less than six patients are waiting to be roomed while nonurgency rooms are currently and soon to be available, the need to room a low-acuity patient in an urgency room is weak. Therefore, we interpret nonurgency room availability as resource availability since they are positively correlated.

We use our observations based on coefficient estimates to form hypotheses about how ED system status impacts patient prioritization. The most important factor that splits the tree for urgency room prioritization is whether ED is highly crowded. If the waiting room census is ten or more, i.e., the ED is highly crowded, then the algorithm finds no further segmentation. Therefore, we conclude that the prioritization rule is stable and tends to comply ESI-based FCFS principle during the busy period. The algorithm finds further segmentation if the waiting room census is less than ten patients. The impact of the ESI scores on the utility function changes as the time of the day, waiting room census, and resource availability, approximated by the number of available clean urgency rooms, changes. We estimate the aforementioned interaction effects by adding the following variables: ESI $\times \mathbb{1}_{[12\text{am},6\text{am})}$, ESI $\times \mathbb{1}_{[12\text{pm},6\text{pm})}$, ESI $\times \mathbb{1}_{[6\text{pm},12\text{am})}$, ESI $\times$ clean-avail-nonurgent, and ESI $\times$ num-waiting for all ESI $\in \{2, 3\}$. Furthermore, Table 4.2 and Table 4.3 hints that waiting room census impacts how the utility function behaves in waiting time. We estimate the interaction effect between the wait time and waiting room census by adding the following variables: ESI $\times$ Wait $\times$ num-waiting and ESI $\times$ Wait^2 $\times$ num-waiting for all ESI $\in \{2, 3, 4-5\}$. Based on Table 4.3, we also add the interaction impact between the arrival by ambulance and waiting room census.

Table 4.6: Estimation results for urgency rooms at different waiting room census level with interaction terms

	num-waiting $\in [2, 10)$		num-waiting $\in [10, -)$	
No. observations	11681		14925	
McFadden pseudo R^2	0.280		0.297	
McFadden pseudo R^2_{adj}	0.270		0.293	
Variables	Coef	Std Err	Coef	Std Err
ESI = 2				
ESI	0.1943	0.297	5.1436^{***}	0.331
ESI $\times$ $\mathbb{1}_{[12\text{am},6\text{am})}$	1.4146^{***}	0.172	0.6415^{**}	0.191
ESI $\times$ $\mathbb{1}_{[12\text{pm},6\text{pm})}$	2.6221^{***}	0.178	0.0656	0.323
ESI $\times$ $\mathbb{1}_{[6\text{pm},12\text{am})}$	2.3353^{***}	0.203	0.3925	0.203
ESI $\times$ num-waiting	0.4020^{***}	0.035	0.0299	0.054
ESI $\times$ clean-avail-nonurgent	-0.3462^{***}	0.034	-0.0727	0.037
ESI $\times$ Wait	0.6396^{**}	0.202	1.3844^{***}	0.056
ESI $\times$ Wait^2	-0.0451	0.044	-0.1726^{***}	0.011
ESI $\times$ Wait $\times$ num-waiting	0.1502^{**}	0.042	-0.0521^{***}	0.006
ESI $\times$ Wait^2 $\times$ num-waiting	-0.0383^{**}	0.010	0.0109^{***}	0.001
ESI $\times$ admit prediction	0.6886	0.420	0.3620	0.269
ESI = 3				
ESI	-0.2720	0.150	2.4370^{***}	0.296
ESI $\times$ $\mathbb{1}_{[12\text{am},6\text{am})}$	1.2978^{***}	0.150	0.3355	0.182
ESI $\times$ $\mathbb{1}_{[12\text{am},6\text{am})}$	2.2705^{***}	0.161	0.6558^{*}	0.315
ESI $\times$ $\mathbb{1}_{[6\text{pm},12\text{am})}$	1.9596^{***}	0.189	0.7112^{***}	0.196
ESI $\times$ num-waiting	0.2035^{***}	0.032	-0.0597	0.056
ESI $\times$ clean-avail-nonurgent	-0.2708^{***}	0.031	-0.0316	0.037
ESI $\times$ Wait	1.6919^{***}	0.123	1.7435^{***}	0.054
ESI $\times$ Wait^2	-0.3307^{***}	0.030	-0.1408^{***}	0.007
ESI $\times$ Wait $\times$ num-waiting	0.0279	0.024	-0.0464^{***}	0.008
ESI $\times$ Wait^2 $\times$ num-waiting	0.0219^{***}	0.005	0.0082^{***}	0.001
ESI $\times$ admit prediction	0.2317	0.301	-0.1024	0.224
ESI = 4 or 5				
ESI $\times$ Wait	0.7762^{***}	0.096	1.9171^{***}	0.119
ESI $\times$ Wait^2	-0.1286^{***}	0.016	-0.1402^{***}	0.014
ESI $\times$ Wait $\times$ num-waiting	0.0984^{***}	0.021	-0.0782^{**}	0.024

Continued on next page

Table 4.6 – continued from previous page

	num-waiting $\in [2, 10)$		num-waiting $\in [10, -)$	
Variables	Coef	Std Err	Coef	Std Err
ESI $\times$ Wait^2 $\times$ num-waiting	0.0052	0.003	0.0105***	0.002
ESI $\times$ admit prediction	−2.5095*	1.147	0.1972	1.287
ambulance	0.2433*	0.107	0.0232	0.061
ambulance $\times$ num-waiting	−0.0025	0.023	0.0129	0.007
seen by provider	−1.1499***	0.076	−0.8694***	0.041
gender(base = female)	0.0601	0.035	0.0430	0.026
30-day ED readmission	−0.1810***	0.035	−0.0443	0.025
90-day ED readmission	−0.0408	0.039	−0.0542	0.028
age (base = 18-45 years)				
46-55 years	−0.0325	0.051	−0.0006	0.038
56-65 years	−0.0247	0.066	0.0949*	0.047
66-75 years	0.0756	0.084	0.1353*	0.057
76-85 years	0.1615	0.105	0.2553***	0.069
86+ years	0.3557*	0.140	0.4591***	0.087

A subset of covariates removed for the sake of space; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Bases on the estimation results shown on Table 4.6, we make the following observations:

- When less than ten patients are waiting, it is more likely to select a mid/high acuity patient than low acuity patient if
  - time $\in [12pm, 6am\ next\ day)$
  - more patients are waiting to be roomed
  - there are fewer available nonurgency rooms.
- The positive waiting room census effect on the likelihood of selecting mid/high acuity patients rather than low acuity patients vanishes after the waiting room census reaches a certain level. Suppose that there are less than ten patients in the waiting room. Then an ESI-2 patient and ESI-3 patient gains additional utility of 0.4020 and 0.2035, respectively, for each patient in excess of two in the waiting room. This implies that as the waiting room census increases, ESI-2 patient brings more utility to the decision-maker than ESI-3 patient. Similarly, an ESI-2 or ESI-3 patient brings more utility to

the decision-maker than an ESI-4 or ESI-5 patient. Suppose that ten or more patients are waiting, then the coefficients ESI $\times$ num-waiting for all ESI $\in \{2, 3\}$ estimates the additional utility gain of an ESI-2 and ESI-3 relative to low acuity patient for each patient in excess of ten in the waiting room. Our estimation results show that these interaction terms are not statistically significant. Therefore, we infer that the interaction effect between the ESI score and waiting room census decreases once the waiting room census reaches a certain level.

- Marginal waiting cost rate behaves more "linear" as the waiting room census increases.
- Table 4.3 suggests that as the waiting room census increases, the positive impact of the arrival by ambulance on the utility function starts to decrease. In Table 4.6, the coefficient estimate of the interaction between ambulance and waiting room census is negative as expected, but not significant. This might be due to insufficient sample size or the strong nonlinear impact of waiting room census on the additional utility obtained if the arrival mode is an ambulance.

4.3 Performance Measure

We evaluate the algorithm's performance based on the number of patients in the waiting room. If the choice set comprises only two patients, a random algorithm predicts the true chosen patient correctly with a 50% chance. If twenty patients are in the choice set, the likelihood of accurately identifying the true chosen patient goes down to the 5%. Therefore, an algorithm with a predictive accuracy of 50% can be considered bad if two patients are in the choice set but good if twenty patients are in the choice set. Let w denote the choice set size and N_w denote the set of possible levels of choice set sizes. We let n_w be the number of incidents with a choice set size of w. At each $w \in N_w$, we measure the algorithm's performance using $\text{alg}_w(N)$ which we define as the percentage of the times that the actual chosen patient is ranked in the top N by the prediction algorithm. We take a weighted

average of $\text{alg}_w(N)$ to measure the overall algorithm performance.

$$\frac{1}{\sum_{w \in N_w} n_w} \sum_{w \in N_w} n_w \text{alg}_w(N). \tag{4.2}$$

We observe and compare the performance of the MST algorithm to ESI-FCFS policy and CLM with the utility function given by equation 4.1. The ESI-FCFS policy provides strict priority for patients with the most urgent ESI score regardless of how long the patient with a less urgent ESI score waits. If multiple patients have the highest ESI score, the algorithm picks the patient with the longest waiting among those.

Table 4.7 shows that the ESI-FCFS policy performs the best at metric $\text{alg}_w(1)$, i.e., predicting the actual chosen patient for urgency rooms. As indicated by Table 4.7 and Table 4.8, the CLM model and the MST algorithm perform similar to or better than the ESI-FCFS policy at metrics $\text{alg}_w(N)$ where $N > 2$. Figure 4.6 supports this claim by illustrating how the overall performance of ESI-FCFS, CLM, and the MST algorithm (excluding trivial cases) changes based on the selected measure $\text{alg}_w(N)$ as a function of N. Based on the empirical evidence, we conclude that the CLM model and the MST algorithm cannot predict the chosen patient as well as ESI-FCFS policy but provides a set of patients who are more likely to be chosen than those who are highly ranked by ESI-FCFS policy in practice.

Table 4.8 suggest that performance of the MST algorithm over the CLM model tends to increase as we use larger n to define the metric $\text{alg}_w(N)$.

Furthermore, the MST algorithm uses the brier score to measure the performance and generate the segmentation tree. The Brier score of an algorithm is calculated by taking the mean of squared error terms between the realized outcome (either 0 or 1) and the assigned probability by the algorithm. See Aouad et al. [2019] for a better understanding of the Brier score calculation. The Brier score of the CLM model is 0.6917, while the Brier score of the MST algorithm is 0.6784 on the test data. The MST algorithm provides a 1.93% improvement in the Brier score over the CLM model. Furthermore, the MST algorithm provides useful pedagogical insights into how the ED system impacts patient prioritization, which can be further used to develop a better decision rule that improves patient outcomes.

Table 4.7: Comparison of the MST algorithm, ESI-FCFS and CLM model on UChicago ED dataset I.

w	n_w	$\text{alg}_w(1)$			$\text{alg}_w(2)$			$\text{alg}_w(3)$		
		ESI FCFS	CLM	MST	ESI FCFS	CLM	MST	ESI FCFS	CLM	MST
2	720	75.83%	68.89%	71.81%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
3	562	65.84%	61.21%	65.30%	91.10%	88.43%	88.43%	100.00%	100.00%	100.00%
4	439	61.73%	61.28%	56.72%	85.88%	84.74%	83.14%	95.22%	95.44%	94.76%
5	363	60.06%	58.68%	56.20%	80.99%	82.92%	80.17%	90.91%	91.74%	90.63%
6	358	50.00%	49.72%	50.00%	72.07%	74.86%	74.86%	86.59%	89.11%	87.71%
7	363	46.56%	43.53%	42.70%	71.35%	74.10%	71.63%	84.57%	86.23%	86.23%
8	383	45.17%	45.95%	42.56%	68.41%	71.54%	67.62%	80.94%	87.21%	83.55%
9	323	43.96%	47.99%	46.75%	66.56%	69.97%	69.35%	80.80%	81.11%	81.11%
10	336	45.83%	44.05%	45.83%	67.86%	68.45%	70.24%	77.98%	81.55%	79.46%
11	326	38.65%	40.49%	40.49%	61.96%	64.11%	63.19%	78.22%	77.30%	77.61%
12	319	38.87%	37.30%	37.62%	54.86%	59.25%	57.99%	69.91%	74.29%	73.98%
13	316	37.34%	39.56%	37.66%	62.97%	60.44%	60.13%	73.10%	73.73%	74.05%
14	250	37.20%	34.00%	34.00%	58.80%	57.20%	56.00%	71.20%	75.20%	73.60%
15	274	40.15%	32.48%	33.58%	57.30%	55.47%	54.74%	68.98%	70.44%	70.80%
16	262	35.88%	30.15%	31.68%	52.29%	53.82%	55.73%	66.79%	67.94%	68.70%
17	224	34.82%	30.80%	33.04%	52.68%	52.23%	53.57%	64.29%	63.84%	66.07%
18	203	30.54%	33.00%	32.51%	51.72%	50.74%	52.22%	60.10%	63.05%	64.53%
19	146	29.45%	28.08%	29.45%	48.63%	47.26%	47.95%	58.22%	57.53%	58.90%
20	128	21.88%	23.44%	25.00%	42.19%	41.41%	39.84%	53.91%	51.56%	50.00%
21+ (23.37)	412	32.52%	27.43%	27.43%	48.06%	43.93%	45.14%	58.49%	58.98%	58.74%
overall	6707	48.19%	46.01%	46.19%	69.90%	70.15%	69.63%	80.39%	81.72%	81.33%
overall ($w > 2$)	5987				66.28%	66.56%	65.98%			
overall ($w > 3$)	5425							75.76%	77.40%	76.92%

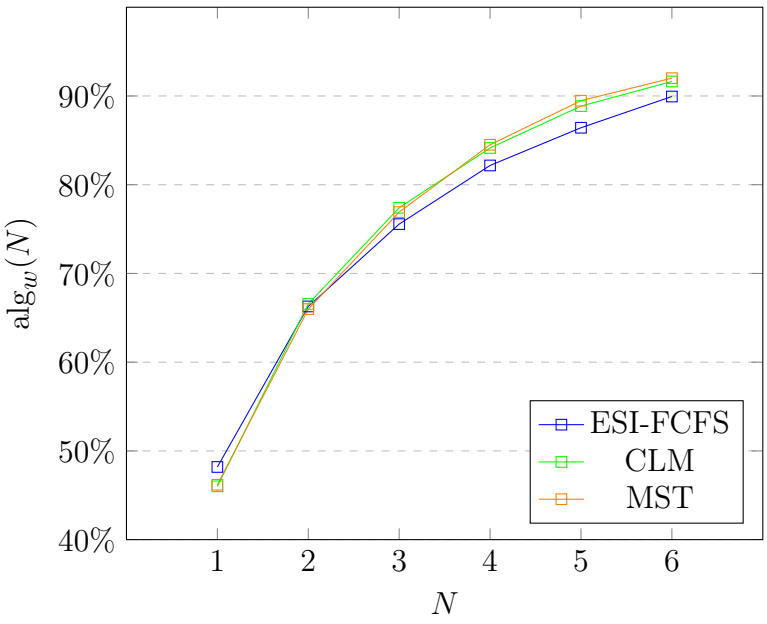


Figure 4.6: Comparison of the MST algorithm, ESI-FCFS and CLM model on UCM ED dataset III.

Table 4.8: Comparison of the MST algorithm, ESI-FCFS and CLM model on UCM ED dataset II.

w	n_w	$\text{alg}_w(4)$			$\text{alg}_w(5)$			$\text{alg}_w(6)$		
		ESI FCFS	CLM	MST	ESI FCFS	CLM	MST	ESI FCFS	CLM	MST
2	720	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
3	562	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
4	439	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
5	363	95.87%	97.52%	96.69%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
6	358	93.58%	95.53%	94.13%	97.49%	97.49%	97.49%	100.00%	100.00%	100.00%
7	363	92.56%	93.94%	94.49%	97.52%	98.07%	97.52%	99.17%	99.45%	99.45%
8	383	87.99%	92.43%	89.82%	93.21%	95.82%	94.52%	97.91%	98.17%	96.87%
9	323	87.93%	90.09%	91.33%	93.19%	96.28%	96.28%	97.83%	97.83%	97.83%
10	336	85.42%	89.58%	89.29%	91.37%	94.35%	95.54%	95.24%	98.51%	97.92%
11	326	87.73%	88.34%	89.26%	91.10%	93.87%	93.56%	93.25%	96.32%	96.63%
12	319	81.19%	85.27%	85.89%	88.40%	91.54%	92.48%	91.22%	95.30%	94.67%
13	316	80.70%	81.96%	82.91%	85.76%	86.39%	87.66%	90.51%	90.82%	90.82%
14	250	80.00%	82.80%	83.60%	86.40%	91.20%	92.80%	93.20%	95.20%	95.60%
15	274	78.10%	81.75%	81.39%	84.67%	90.15%	90.51%	88.32%	92.34%	93.43%
16	262	75.95%	77.48%	79.77%	80.53%	83.97%	87.02%	85.11%	88.93%	90.46%
17	224	76.34%	74.11%	75.00%	83.48%	84.82%	83.93%	89.73%	87.50%	89.29%
18	203	66.50%	72.41%	75.86%	73.89%	78.82%	79.80%	80.79%	82.27%	82.27%
19	146	65.75%	67.12%	66.44%	73.29%	78.08%	79.45%	82.19%	83.56%	84.93%
20	128	67.97%	60.94%	64.84%	73.44%	67.19%	71.88%	79.69%	71.88%	77.34%
21+ (23.37)	412	65.05%	65.54%	66.26%	0.6796	70.87%	71.84%	72.57%	77.18%	77.91%
overall	6707	86.75%	88.21%	88.47%	90.64%	92.32%	92.74%	93.60%	94.68%	94.92%
overall ($w > 4$)	4986	82.17%	84.14%	84.50%						
overall ($w > 5$)	4623				86.42%	88.86%	89.47%			
overall ($w > 6$)	4265							89.94%	91.63%	92.01%

CHAPTER 5
ALTERNATIVE-SPECIFIC DECISION TREES

5.1 Introduction

The Emergency Department choice set comprises patients with different health conditions and complaints. ESI scores could be used for the segmentation, but there is room for improvement. First, there is empirical evidence that a prioritization rule solely based on ESI scores is not strictly followed. Figure 5.1 and Figure 5.2 show that the prioritization is not only based on ESI scores at UCM, even for the primary service area dedicated to triage levels 1,2 and 3 patients. Furthermore, a study conducted at a large Canadian hospital shows that patients with medium urgency are chosen over patients with high urgency for the primary service area 57.1% when at least one of each is present in the waiting room (Ding et al. [2019]). Secondly, there are no clear guidelines on prioritizing patients within the same triage level, which evokes the need for greater segmentation. We develop an algorithm that simultaneously performs the alternative (e.g., patient) segmentation and prediction process so that the alternative segmentation is (nearly) optimized for the prediction process. This tree-based algorithm can discover the underlying alternative segmentation process of the decision-maker when there is no explicit categorization for alternatives as patients at ED. We evaluate the performance of our algorithm and conditional logit model (CLM) on artificially created datasets for different settings with different utility function forms. We observe that the tree-based algorithm does not overfit the data where the interaction effects are weak, or there is no alternative segmentation. When the interaction effects are stronger, or there is alternative segmentation, the tree-based algorithm performs better than the CLM on the artificially generated datasets.

5.2 Literature Review

In choice models, alternative specific constants reflect the additional utility of choosing that alternative instead of the reference alternative. The inherent utility of the alternatives could differ across different segments. We summarize the relevant papers that allow customer

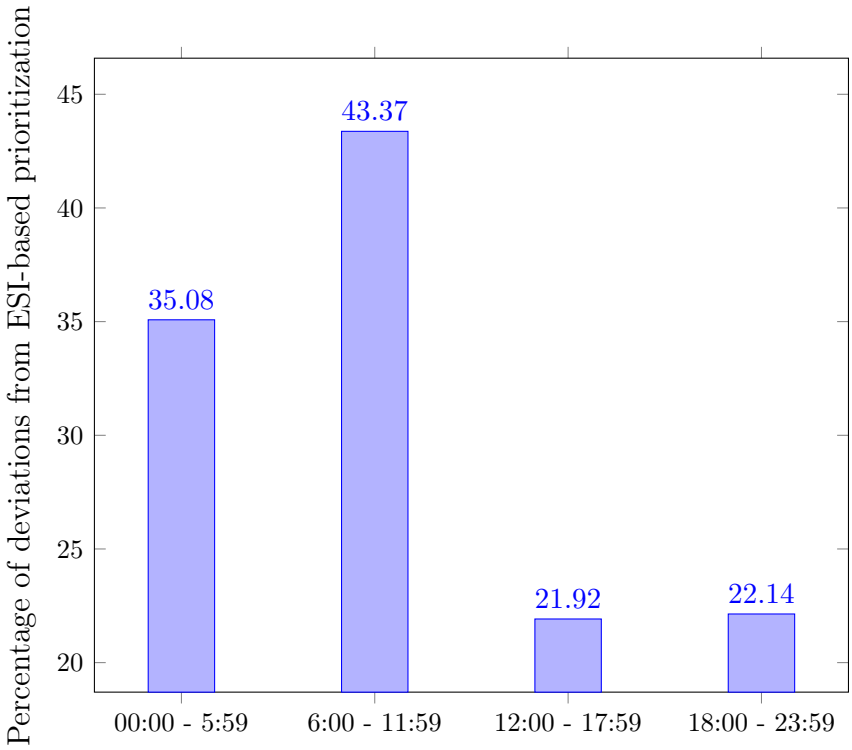


Figure 5.1: Percentage of deviations from ESI-based prioritization by the time of the day.

segmentation using alternative specific constants.

Arentze and Timmermans [2007] propose a parametric action decision tree algorithm (PADT) that includes alternative-specific constants in the parametric model. The PADT algorithm first uses a CHAID-based decision tree method to partition the population-based on its categorical attributes. Then it adds alternative specific constants that reflect the leaf membership in the multinomial logit model. The information on continuous attributes is used in the final model to improve the predictive accuracy.

Kim [2009] proposes a two-stage logistic regression to capture the model's interaction effects. A decision tree algorithm is used in the first stage to partition the population into clusters. The optimal split criterion is based on the chi-squared test. In the second stage, dummy variables are added to the model to indicate the observation's cluster – leaf membership–.

Similar to the previous paper, Kim and Kim [2011] proposes a two-stage MNL model (TMLM) to capture the interaction effects of the attributes in a choice model with multi-category responses. In contrast to PADT, in the TMLM model, all attributes, including both

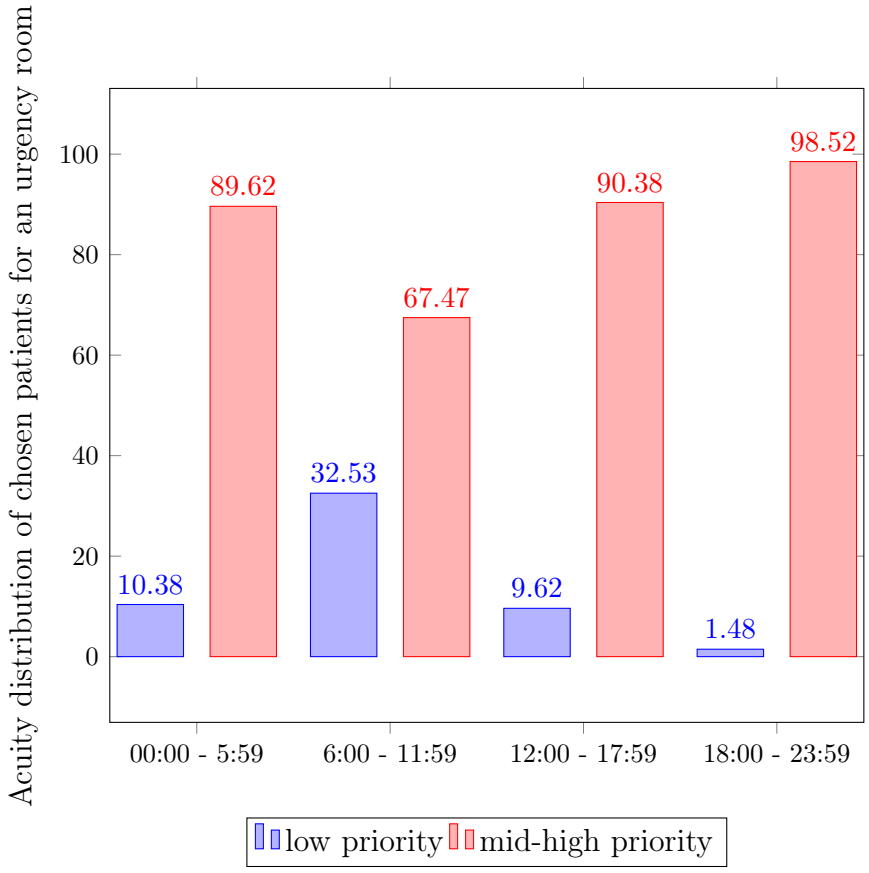


Figure 5.2: Acuity distribution of chosen patients in urgency room choice incidents when there is at least one patient from each priority group by the time of the day.

continuous and categorical attributes, belong to the decision-maker and can be used to construct the decision tree. In the first stage, a decision tree algorithm is used for clustering. In the second stage, alternative-specific constants are added to the model, similar to the PADT. The main difference between the two models is whether additional information is available on the alternatives. PADT is used for choice models in that alternatives have information that affects the outcome. For example, PADT uses the information on continuous attributes of the alternatives in the second stage. The main application behind PADT is to model travel mode choice where each transport mode can have specific information on attributes such as cost and duration. On the other hand, TMLM is used to model multi-categorized outcomes rather than choice outcomes. For example, the authors used TMLM to predict the degree of the disability. The outcome is solely based on the available information on the decision-maker.

Our model is most similar to PADT. In both models, leaf nodes of the resulting decision tree represent the group memberships that could be integrated into the model in the form of alternative specific constants. Both models result in a single logit model that could leverage the information on alternatives. The PADT algorithm model is used for the choice models in which the set of alternatives is finite. It does not capture the interactions between the alternative attributes because decision tree generation in the PADT model is solely based on the decision-maker attributes. Our model uses the MST algorithm to partition the decision-makers, which already captures the decision-maker interactions and partitioning. Our decision tree generation process on the alternative attributes at each partition allows us to capture the interaction terms. Secondly, the PADT algorithm first performs the clustering and then fits a multinomial logit model. In our model, clustering is based on the response behavior, i.e., how much it improves the predictive accuracy of the resulting logit model. In other words, the clustering step cannot be isolated from the model fit step, while the clustering step is independent of the model fit step in the PADT algorithm.

The second part of the algorithm is similar to the MST algorithm. The decision tree is based on response behavior and integrated with the model fit process in both models. In the MST algorithm, leaf nodes represent a disjoint group of decision-makers, and a separate logit model is used for each segment. In our model, leaf nodes represent the high-level

priority segmentation of alternatives and are integrated into a single logit model in the form of alternative-specific constants.

5.3 The Algorithm Description

We create a decision tree based on alternative-specific covariates. Branch nodes split the set of alternatives $\mathbb{I}$ into smaller partitions presented by leaf nodes. We let $\mathbb{L}$ be the set of leaf nodes, and A_ℓ denote the set of alternatives that fall into leaf node ℓ.

$$\mathbb{I} = \bigcup_{\ell \in \mathbb{L}} A_\ell,$$

where $A_\ell \cap A_{\ell'} = \emptyset \ \forall \ell \neq \ell'$. We expect alternatives with similar covariates to fall into the same leaf node and provide similar utility to the decision-maker. We model leaf-node membership in the form of alternative specific constant in the CLM. In other words, leaf node membership represents the alternative category and captures the interactions and nonlinear relationships between alternative attributes.

We formulate the leaf-node membership by using indicator variables $\mathbb{1}_{i \in A_\ell}$:

$$\mathbb{1}_{i \in A_\ell} = \begin{cases} 1, & \text{if alternative } i \text{ falls into leaf node } \ell \\ 0, & \text{otherwise} \quad \forall \ell \in \mathbb{L}, i \in \mathbb{I}. \end{cases} \tag{5.1}$$

We let α_m be the utility constant of alternative category m then we can write the utility of patient i at decision epoch t as

$$\begin{aligned} V_{ti} &= \sum_{m \in \mathbb{M}} \mathbb{1}_{i \in \ell_m} \alpha_k + \sum_{k=1}^{M^a} \beta_k x_{i,k} + \epsilon_{ti}, \\ &= \boldsymbol{\alpha}^T \mathbf{e}_i + \boldsymbol{\beta}^T \mathbf{x}_i, \end{aligned} \tag{5.2}$$

Let $\mathbf{e}_i$ is a $1 \times |\mathbb{L}|$ dimensional unit vector that takes value of 1 at the position that corresponds to leaf node into which the alternative i fall and takes value of 0 at anywhere else. Let $\boldsymbol{\beta}$ be $1 \times M^a$ dimensional vector such that $\boldsymbol{\beta} = \langle \beta_1, \ldots, \beta_{M^a} \rangle$ and $\boldsymbol{\alpha}$ be $1 \times |\mathbb{L}|$ $\boldsymbol{\alpha} = \langle \alpha_\ell : \ell \in \mathbb{L} \rangle$,

Without loss of generality, we assume that a leaf node with higher ℓ value is placed after the leaf node with a smaller ℓ value in the vector $\boldsymbol{\alpha}$.

After assigning each patient to a category, we run a single CLM model with alternative specific constants on the training data. In the decision tree generation process for standard regression and classification problems, it is sufficient the evaluate the performance of a split only on a subset of the training data that fall into that node that just got split because observations only impact the prediction outcome of the observations that fall into that same node. In our case, patients who belong to the same choice incident can fall into different leaf nodes. Changing the leaf node membership of one observation will impact the prediction outcome of observations associated with the same incident because only one patient in a choice incident can have the "chosen" outcome.

We split the dataset into training, validation, and test set. We use maximum likelihood estimation to estimate the problem parameters $\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle$ on the training set. The objective is to minimize the log-likelihood function:

$$\sum_{t \in \mathbb{T}} \sum_{i \in C_t} -\log \left(\boldsymbol{\alpha}^T \mathbf{e}_i + \boldsymbol{\beta}^T \mathbf{x}_i \right)$$

We evaluate the performance of the CLM with the estimated parameters on the validation set. Let us explain the first iteration in the decision tree process in detail to better understand the algorithm. We start with the root node 1 at the first iteration, i.e., assume that all alternatives belong to a single category, i.e., the initial tree $\mathbb{T}_1$ is a single node. Then, the utility function is given by

$$V_{ti} = \alpha + \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_{ti}.$$

For each split, attribute and cut-off value $(k, \bar{x}_k)$, we partition the set of alternatives into two sets: $\mathbb{I} = C_2 \cup C_3$.

$$C_2\,(k, \bar{x}_k) = \{i \in \mathbb{I} : x_{ik} < \bar{x}_k\},$$
$$C_3\,(k, \bar{x}_k) = \{i \in \mathbb{I} : x_{ik} \geq \bar{x}_k\}.$$

Then, the utility function is given by

$$V_{it} = \boldsymbol{\alpha}^T \mathbf{e}_i + \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_{ti},$$

where $\boldsymbol{\alpha} = \langle \alpha_2, \alpha_3 \rangle$ and

$$\mathbf{e}_i = \begin{cases} \langle 1, 0 \rangle & \text{if } i \in C_2\left(k, \bar{x}_k\right), \\ \langle 0, 1 \rangle & \text{if } i \in C_3\left(k, \bar{x}_k\right). \end{cases}$$

The problem parameters $\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle$ is estimated via maximum likelihood estimation on the training set. If the size of the newborn leaf nodes satisfy the minimum observation requirement, then this is a feasible split. We perform the feasible split that results in the best performance on the validation set. Let $\left(k^*, \bar{x}_k^*\right)$ be the best feasible split at the root node and

$$C_2 = \{i \in \mathbb{I} : x_{ik} < \bar{x}_k^*\},$$
$$C_3 = \{i \in \mathbb{I} : x_{ik} \geq \bar{x}_k^*\}.$$

If there is no split at the parent node 1 such that both children C_2 and C_3 satisfy the minimum observation requirement, then the parent node is terminated and represents a leaf node in the final tree. We assume that the root node has at least one split that satisfies the minimum observation requirement. Suppose that no split satisfies this requirement but does not perform better than the previous CLM model on the validation set, then we do not perform the split.

At each iteration, we let $\mathbb{L}_h$ represent the set of leaf nodes in the $\mathbb{T}_h$. For example, the root node is the only node on that we can perform a feasible split at the first iteration: $\mathbb{L}_1 = \{1\}$. For all $\ell \in \mathbb{L}_h$, we iterate over all feasible splits and measure the performance of the split on the validation set. Suppose that $(k, \bar{x}_k^*)$ is the best feasible split that results in two newborn leaves 2ℓ and $2\ell + 1$:

$$C_{2\ell}\left(k, \bar{x}_k\right) = \{i \in C_\ell : x_{ik} < \bar{x}_k^*\},$$
$$C_{2\ell+1}\left(k, \bar{x}_k\right) = \{i \in C_\ell : x_{ik} \geq \bar{x}_k^*\}.$$

The resulting tree $\mathbb{T}_{h,\ell,k,\bar{x}_k}$ is the modified version of the $\mathbb{T}_h$ such that the ℓ is a branch node

with two newborn leaves 2ℓ and $2\ell+1$. The CLM model for this split is

$$\boldsymbol{\alpha} = \langle \alpha_{\ell'} : \ell' \in \mathbb{L}_h \setminus \ell \rangle + \langle \alpha_{2\ell}, \alpha_{2\ell+1} \rangle,$$

where plus operator concatenates two vectors in a way the leaf node with higher m value gets positioned at a higher index. We update the vector $\mathbf{e}_i$ accordingly for all $i \in \mathbb{I}$. We estimate $\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle$ via maximum likelihood estimation on the training set. We evaluate the performance of the new CLM model on the validation set. We let accuracy $\left(\mathbb{T}_{h,\ell,k,\bar{x}_k}, \mathbf{X}_1, \mathbf{y}_1, \mathbf{X}_2, \mathbf{y}_2\right)$ function returns the performance of the CLM model, induced by the tree $\mathbb{T}_{h,\ell,k,\bar{x}_k}$ by using the data $(\mathbf{X}_1, \mathbf{y}_1)$, on the data $(\mathbf{X}_2, \mathbf{y}_2)$. At each iteration our objective is to find the node (ℓ^*), attribute (k^*), cutoff value $(\bar{x}_k^*)$ such that

$$\ell^*, k^*, \bar{x}_k^* = \underset{\ell,k,\bar{x}_k}{\operatorname{argmax}} \operatorname{accuracy}\left(\mathbb{T}_{h,\ell,k,\bar{x}_k}, \mathbf{X}, \mathbf{y}, \mathbf{X}, \mathbf{y}\right). \tag{5.3}$$

Then,

$$\begin{aligned}
\mathbb{T}_{h+1} &= \mathbb{T}_{h,\ell^*,k^*,\bar{x}_k^*}, \\
\mathbb{L}_{h+1} &= \mathbb{L}_{h+1} \setminus \ell^* \cup \{2\ell^*, 2\ell^*+1\}
\end{aligned}$$

Algorithm 1 TreeGrowing

Input: Starting with root node $\mathbb{T}$; training data $(\mathbf{X}, \mathbf{y})$
metric $\leftarrow$ accuracy $(\mathbb{T}, \mathbf{X}, \mathbf{y}, \mathbf{X}, \mathbf{y})$
repeat
    $\mathbb{L} \leftarrow$ leaf-nodes$(\mathbb{T})$
    $\ell_{\text{best}} = -1$
    for all $\ell \in \mathbb{L}$ **do**
        $\mathbb{T}_\ell, \text{metric}_\ell \leftarrow$ BranchingProcess$(\mathbb{T}, \ell, \mathbf{X}, \mathbf{y})$
        if $\text{metric}_\ell > \text{metric}$ **then**
            $\ell_{\text{best}} = \ell$
            $\text{metric} \leftarrow \text{metric}_\ell$
        end if
    end for
    if $\ell_{\text{best}} \neq -1$ **then**
        $\mathbb{T}, \text{metric} \leftarrow \mathbb{T}_{\ell_{\text{best}}}, \text{metric}_{\ell_{\text{best}}}$
    end if
until $\ell_{\text{best}} = -1$
return $\mathbb{T}$

Algorithm 2 BranchingProcess

Input: $\mathbb{T}_{\text{m}}$, $\ell \in$ leaf-nodes($\mathbb{T}_{\text{m}}$), metric_{m} and training data $(\mathbf{X}_{\text{m}}, \mathbf{y})$ where $\mathbf{X}_{\text{m}}$ is the set of observations that fall into leaf node m and $\mathbf{y}$ is the outcome of all observations regardless of which leaf node they fall into.
$n, p \leftarrow$ size of $\mathbf{X}_{\text{m}}$
$\mathbb{T}_{\text{best}} \leftarrow \mathbb{T}_{\text{m}}$, $\text{metric}_{\text{best}} \leftarrow \text{metric}_{\text{m}}$
for all $j = 1, \ldots, p$ **do**
    values $\leftarrow$ candidate split values for j^{th} patient attribute
    sort values in ascending order
    for all $i = 1, \ldots, |\text{values}| - 1$ **do**
        $b \leftarrow \frac{1}{2}(\text{values}_i + \text{values}_{i+1})$
        $\mathbb{T} \leftarrow$ modified $\mathbb{T}_{\text{m}}$ s.t. ℓ is a branch with $\mathbf{e}_j^T \mathbf{x} < b$; two new leaves 2ℓ and $2\ell + 1$ are born.
        if $\min\{\text{size}(2\ell), \text{size}(2\ell + 1)\}) \geq N_{\min}$ **then**
            $\text{metric} \leftarrow \text{accuracy}(\mathbb{T}, \mathbf{X}, \mathbf{y}, \mathbf{X}, \mathbf{y})$
            if $\text{metric} > \text{metric}_{\text{best}}$ **then**
                $\text{metric}_{\text{best}} \leftarrow \text{metric}$
                $\mathbb{T}_{\text{best}} \leftarrow \mathbb{T}$
            end if
        end if
    end for
end for
return $\mathbb{T}_{\text{best}}, \text{metric}_{\text{best}}$

Algorithm 3 TreePruning

Input: $\mathbb{T}_{\mathrm{m}}$, $\ell \in$ leaf-nodes($\mathbb{T}_{\mathrm{m}}$), $\mathrm{metric}_{\mathrm{m}}$, training data $(\mathbf{X}_{\mathrm{m}}, \mathbf{y})$ and validation data $(\mathbf{X}_{\mathrm{m}}^{v}, \mathbf{y}^{v})$
Output:
$\mathrm{metric}_{\mathrm{best}} \leftarrow \mathrm{metric}_{\mathrm{m}}$
repeat
    $\mathbb{T}_{\mathrm{best}} \leftarrow \mathbb{T}_{\mathrm{m}}$, $\mathrm{metric}_{\mathrm{prev}} \leftarrow \mathrm{metric}_{\mathrm{best}}$
    for all $\ell \in$ leaf-nodes **do**
        if sibling(ℓ) $\in$ leaf-nodes **then**
            $\mathbb{T} \leftarrow$ modified $\mathbb{T}_{\mathrm{m}}$ s.t. ℓ and sibling(ℓ) are pruned and parent(ℓ) is a leaf.
            $\mathrm{metric} \leftarrow \mathrm{accuracy}(\mathbb{T}, \tilde{\mathbf{X}}, \mathbf{y}, \tilde{\mathbf{X}}^{v}, \mathbf{y}^{v})$
            if $\mathrm{metric} > \mathrm{metric}_{\mathrm{best}}$ **then**
                $\mathrm{metric}_{\mathrm{best}} \leftarrow \mathrm{metric}$
                $\mathbb{T}_{\mathrm{best}} \leftarrow \mathbb{T}$
            end if
        end if
    end for
    $\mathrm{metric}_{\mathrm{m}} \leftarrow \mathrm{metric}_{\mathrm{best}}$
    $\mathbb{T}_{\mathrm{m}} \leftarrow \mathbb{T}_{\mathrm{best}}$
until $|\text{leaf-nodes}(\mathbb{T}_{\mathrm{m}})| = 1$ or $\mathrm{metric}_{\mathrm{prev}} = \mathrm{metric}_{\mathrm{best}}$
return $\mathbb{T}_{\mathrm{m}}, \mathrm{metric}_{\mathrm{m}}$

5.4 Numerical Examples

We evaluate the performance of the tree-based first on the artificially generated datasets and then on the UCM ED dataset.

We first randomly generate ten datasets. Each dataset includes 6000 observations, i.e., incidents. At each observation, the decision-maker is offered a set of alternatives. The number of alternative options at each observation is randomly sampled from the set $\{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Each alternative (i) is encoded by five attributes $(\{x_{i,k}\}_{k=1,2,3,4,5})$

that are independently sampled from Uniform(0,1) distribution.The utility function is generated for each of the three ground truth models:

- Linear utility function. We assume that the utility function has the following form:

$$V_{ti} = \sum_{k=1}^{5} \beta_k x_{i,k} + \epsilon_{ti}.$$

  For each dataset, we generate parameter vector $\boldsymbol{\beta}$ by independently sampling element $\boldsymbol{\beta}_i$ from Uniform(-1,1) distribution for all $i \in \{1, 2, 3, 4, 5\}$. We sample each ϵ_{ti} independently from Gumbel(0,0.01).

- Linear utility function with interaction terms. We assume that the utility function has the following form:

$$V_{ti} = \sum_{k=1}^{5} \beta_k x_{i,k} + \beta_{int} \prod_{k \in K_{int}} x_{i,k} + \epsilon_{ti}.$$

  To generate set K_{int}, we randomly sample 2 attributes without replacement. We generate the parameter vector $\boldsymbol{\beta}$ and random shock as described in the previous bullet point. We let β_{int} denote the interaction coefficient term and repeat the same data generation process for each value of $\beta_{int} \in \{-1, -0.5, 0.5, 1\}$.

- Linear and tree-based utility function. We assume that alternatives are partitioned into segments based on a tree. We first create a set of tree structures where the tree depth is at most three. We randomly select a tree from this set. For each branch node of the sampled tree, we independently sample an attribute from the set $\{1, 2, 3, 4, 5\}$ and a split value from the Uniform(0,1) distribution with the constraint that each leaf node should contain at least N alternatives. For each leaf node (ℓ) of the sampled tree, we independently sample a leaf membership value ($U(\ell)$) from the Uniform(0,1) distribution. We let ℓ_i denote the leaf node that the alternative i falls into given the

constructed tree. We assume that the utility function has the following form:

$$V_{ti} = \sum_{k=1}^{5} \beta_k x_{i,k} + U(i_\ell) + \epsilon_{ti}.$$

We generate the parameter vector $\boldsymbol{\beta}$ and random shock as described in the first bullet point.

We compare the performance of the tree-based CLM with the standard CLM for each ground-truth model. We first make the performance comparison for the linear utility model on ten different randomly generated datasets. We observe that tree-based CLM prunes the tree to depth 0 root node at each dataset. This shows that the tree-based CLM does not overfit the attributes.

We then evaluate the performance of the tree-based CLM against the CLM model for the linear utility model with interaction terms on ten different randomly generated datasets. Figure 5.4 shows that the improvement provided by tree-based CLM over the CLM model increases as the value of the interaction coefficient term, β_{int}, increases. This result is expected due to the fact that decision trees are able to capture the interaction effects.

Finally, we compare the performance of the tree-based CLM against the CLM model on ten different randomly generated datasets where the alternatives provide utility determined by an unknown tree structure in addition to the linear utility. Figure 5.3 shows that the improvement provided by tree-based CLM is significantly higher compared to the previous utility model scenarios.

In the previous chapter, we applied the MST algorithm to the UCM ED data to find system segments where a similar prioritization is followed for patient routing decisions. We reported that the room type is the most important factor than changes in the patient prioritization rule. We also showed that waiting room census and then the time of the day are also important system-level attributes that impact the patient prioritization rule but do not significantly improve the prediction accuracy. In this chapter, we apply tree based logit model on the UCM ED data where an urgency room is utilized. We run the CLM model with the utility function given by equation 3.3. Recall that this utility function includes the linear and quadratic impact of the waiting time and ESI interactions with the wait time

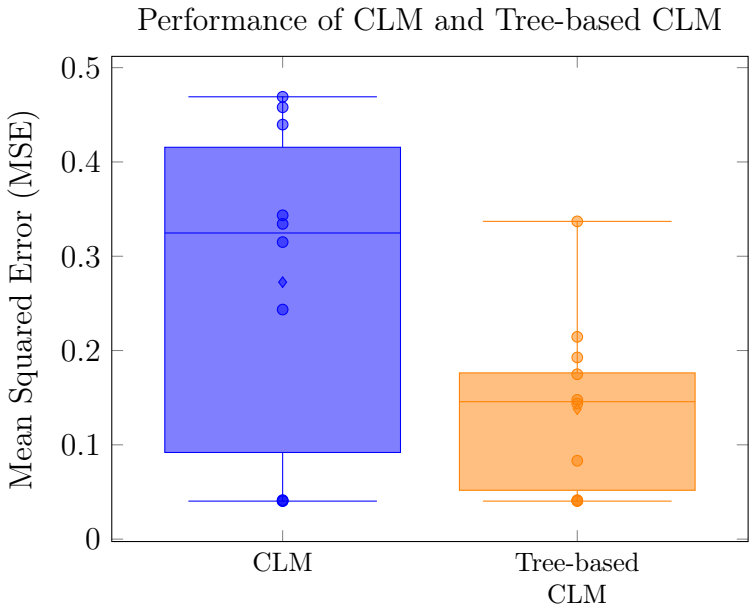


Figure 5.3: Performance of CLM and Tree-based CLM on synthetic datasets generated on linear and tree-based utility model.

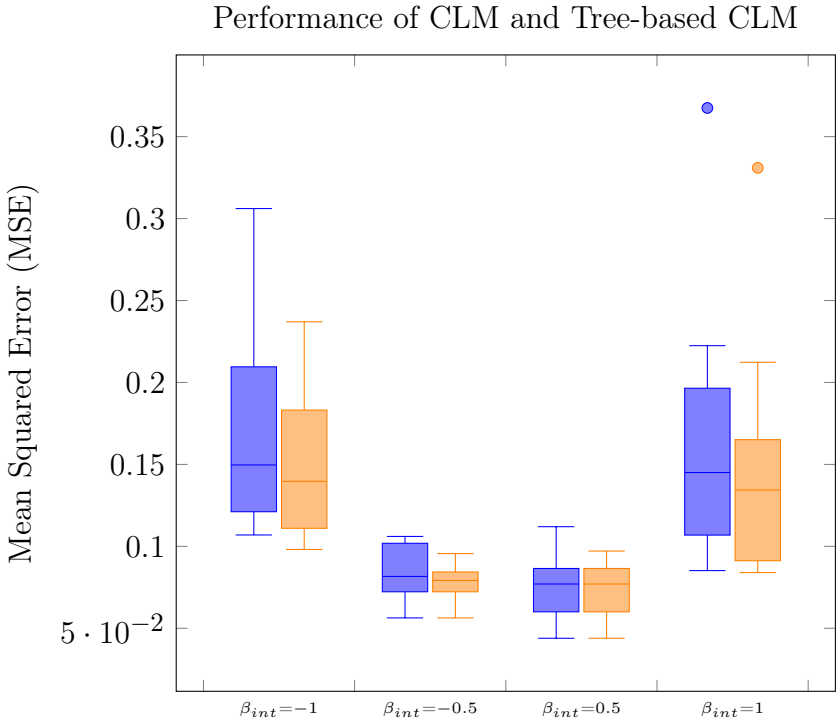


Figure 5.4: Performance of CLM and Tree-based CLM on synthetic datasets generated on linear utility with interactions model.

and disposition outcome. We observe that either there is no patient further segmentation or segmentation by wait time is the most important patient-level attribute. However, we do not find a significant improvement in prediction accuracy as we see in the application of the alternative segmentation algorithm on the synthetic datasets. Therefore, we infer that other nonlinear and interaction impacts of the patient-level attributes not already included in the utility function are relatively weak.

CHAPTER 6
CONCLUSION AND FURTHER RESEARCH

In this thesis, we studied machine learning for queue prioritization in theoretical and empirical settings. In the first part of the thesis, we have focused on analyzing the multiclass queuing system under imperfect information. In the second part of the thesis, we implemented a tree-based algorithm to predict the patient prioritization decisions at the UCM ED. We drove insights into how patient prioritization decisions are based on the system attributes. Finally, we have implemented another tree-based prediction algorithm that can capture patient-level attributes' nonlinear impact and interactions by creating patient clusters.

Multiclass queuing systems are extensively studied in the operations literature. While there is a vast literature on the perfect information case where all customer types are immediately known, relatively few papers focus on the imperfect information case. Many real-life service system providers segment their customers into different priority classes to efficiently allocate their limited resources. However, the service providers usually have limited knowledge on to which class the customer belongs upon arrival. With the growing popularity of artificial intelligence, service providers have started implementing machine learning models to accurately predict the customer class and improve their operations. However, these models tend not to consider the externalities customers impose on each due to shared resources. Multiclass queuing models can capture these externalities; thus are useful tools to improve operations. In this thesis, we have provided a queueing model integrated with machine learning concepts, particularly binary classification, and ROC curve. Our model also captures the service differentiation based on the customer classification outcome. Although the traditional queueing literature assumes that service rate depends on the true customer type rather than classification, there are applications in the emergency department and call centers where the classification outcome impacts the patient/customer service.

Under perfect information, we have proved that classification-based service differentiation may result in a misclassification incentive to minimize the waiting costs. In contrast, the optimal policy under type-dependent service differentiation does not result in any mis-

classification incentive. We have proved that the optimal threshold policy under imperfect information with classification-dependent service rates trades off a lower loss in specificity for a higher gain in sensitivity, i.e., lowers the threshold probability until the marginal gain in sensitivity does not exceed the marginal loss in specificity. However, we can no longer argue that this statement holds if service differentiation is type-dependent. Similarly, we have run numerical experiments showing that characteristics of the optimal policy that minimizes the waiting costs in the setting with type-dependent service differentiation are no longer valid if the service differentiation is class-dependent.

We emphasize that (i) machine learning methods tend to ignore the operational nature of the threshold selection problem and (ii) classification-dependent service rate differentiation occurs in real-life business applications, and the associated optimal policy deviates from the optimal policy established under the type-dependent service differentiation assumption. We have introduced a queuing system addressing these points and studied the optimal policy. However, there are other research directions to be explored in queue prioritization under imperfect information. First, service time can depend on true customer type and the classification outcome. Secondly, the current model assumes a stationary arrival rate and customer mix. In practice, customer arrival and mix are not necessarily stationary. For example, more patients arrive at ED late evening or night at UCM, while fewer patients arrive during the daytime. Patients, who arrive et the ED at night, are more likely to have more urgent conditions and are more likely to be admitted to the inpatient stay. Exploring the dynamic threshold policy and how it depends on the system parameters would be interesting.

This thesis has explored the patient prioritization problem at UCM ED using statistical tools such as the conditional logit model (CLM) and machine learning methods such as decision trees. We have mainly focused on understanding how the ED system and patient characteristics impact the prioritization rule and predicting which patient will be roomed next. However, the current prioritization rule is not necessarily optimal. A natural extension is to develop an improved prioritization rule by estimating the effect of waiting time on medical outcomes such as length of stay (LOS), Left Without Being Seen (LWBS), mortality, risk of admission, ED bounce-backs, etc.

In Chapter 5, we report that the ESI-FCFS policy is better at predicting the next patient

to be roomed than the MST Algorithm. At the same time, the CLM model with patient-level interaction terms is slightly better at predicting the set of patients who are most likely to be roomed. Since the performance of these methods does not significantly differ, we infer that triage nurses generally follow the ESI-FCFS policy.

We have first tested the performance of the CLM model and the MST Algorithm on an artificially generated dataset that includes choice incidents based on a deterministic optimal policy and have found that both methods with appropriate covariates explain most of the variation in the choice outcome. However, when we apply the MST algorithm and ESI-FCFS policy to the dataset that includes choice incidents at UCM ED, we find that they explain only a fraction of the variation in the choice outcome. This situation is either due to the possibly missing covariates in the model or some patient prioritization decisions made at random. To address the first matter, we have added pain score, conscious level as covariates, and other special condition indicators but found no significant improvement in prediction accuracy. It is possible that some patient attributes are observable to the triage nurse but not captured in the data. It is also possible that the MST algorithm might not have sufficient power to fully explain an extremely complex system such as an emergency department. Another possibility is that some patient routing decisions are indeed based on randomization, possibly due to different triage nurses. Here, randomization can be either complete or based on a subset of covariates, such as the ESI score. A future research direction is to explore whether we can estimate what fraction of the unexplained variation is due to the model structure versus randomization.

We start exploring this estimation question on the artificially generated dataset by assuming complete randomization. We first calculate the optimal deterministic policy as explained in Chapter 3 and then run a simulation study with randomization. In the simulation, the decision-maker selects the optimal action with probability $1 - p$ or selects a random patient with probability p whenever it is optimal to transfer a patient to a room from the waiting area. The decision-maker cannot violate the patient routing decisions. For example, if the decision-maker selects the random action, they cannot room a non-urgent patient in an urgency room. After we run the simulation, we generate the choice dataset. Next, we estimate $\hat{p}$ and compare how this estimate is close to the real p.

Let us introduce our preliminary approach to estimate the randomization probability. Suppose that $p = 1$, then any deterministic algorithm can correctly estimate on average r_{rand} fraction of choice incidents where

$$r_{rand} = \frac{1}{\sum_{w \in N_w} n_w} \sum_{w \in N_w} n_w \frac{1}{w}. \tag{6.1}$$

Suppose that $p = 0$, then there is an algorithm with perfect accuracy that can estimate the outcome of all incidents. We assume that if the estimation dataset is sufficiently large and the model includes the relevant covariates, the CLM model with interaction terms and the MST algorithm should be able to pick the optimal policy. The optimal policy should accurately predict the following percentage of the choice outcomes on average:

$$\hat{p} r_{rand} + (1 - \hat{p}) 1 = A_{clm}, \tag{6.2}$$

where A_{clm} is the fraction of incidents correctly predicted by the CLM model in the test data. Then, we have

$$\hat{p} = \frac{1 - A_{clm}}{1 - r_{rand}} \tag{6.3}$$

We run our simulation study and estimate the randomization probability at different levels of p. Figure 6.1 shows that our estimate is biased and tend to underestimate the actual p. More work is needed to explore obtaining an unbiased estimate of the randomization probability.

In Chapter 5, we introduce an algorithm that simultaneously performs the patient segmentation and prediction process so that the segmentation is (nearly) optimized for the prediction process. The algorithm is tree-based and, thus, can flexibly capture the interactions and nonlinear relationships between alternative attributes. We observe that this tree-based algorithm performs well on artificially generated datasets but cannot perform as well in the UCM ED dataset. Therefore, we infer that nonlinear and interaction impacts of the patient-level attributes (excluding the quadratic impact of the wait and interaction between the ESI score and wait, which are already included in the utility function) are relatively weak. Another possibility is that the algorithm is sufficiently flexible to capture the

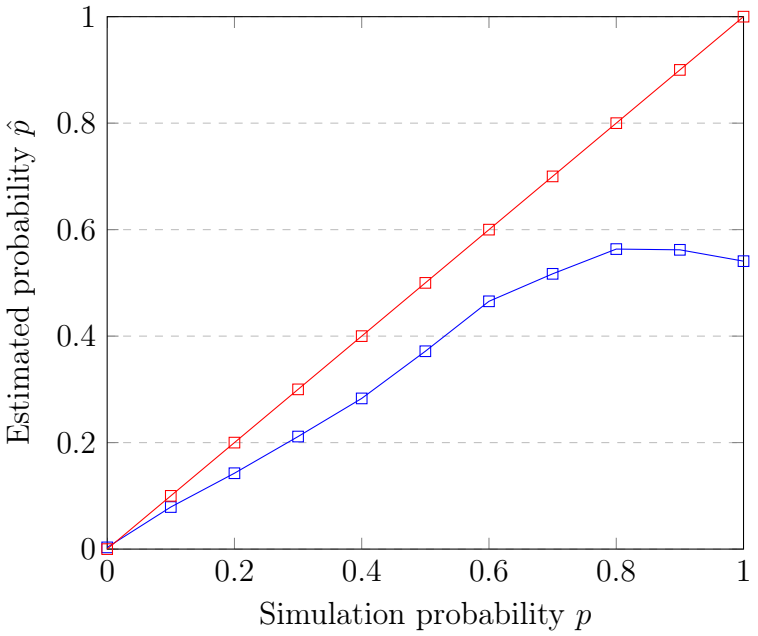


Figure 6.1: Randomization probability estimation

patient-level nonlinear effects and interaction terms in simple settings but not sufficiently flexible to capture these in more complex settings and needs improvement.

APPENDIX A

APPENDIX

A.1 Chapter 2 Proofs

A.1.1 Notation

For notational convenience, we write $W^a(z)$, $W^d(z)$, and $W(z)$ as

$$W^a(z) = \frac{\lambda\beta\mu}{2} R(z) H^a(z), \tag{A.1a}$$

$$W^b(z) = \frac{\lambda\beta\mu}{2} R(z) H^b(z), \tag{A.1b}$$

$$W(z) = \frac{\lambda\beta\mu}{2} R(z) H(z) \tag{A.1c}$$

where

$$R(z) = \left(z(e_1 - e_2) + e_2\right), \tag{A.2a}$$

$$H^a(z) = \frac{1}{\beta\mu - \lambda z}, \tag{A.2b}$$

$$H^b(z) = \frac{1}{\beta\mu - \lambda z}\left(\frac{z-\alpha}{1-\alpha} + \frac{1-z}{1-\alpha}\frac{\beta\mu}{\left(\beta\mu - \lambda\beta + \lambda\left(\beta - 1\right) z\right)}\right), \tag{A.2c}$$

$$H(z) = \frac{1}{\beta\mu - \lambda z}\left(\gamma\alpha + (z-\alpha) + \frac{\beta\mu(1-z)}{\left(\beta\mu - \lambda\beta + \lambda\left(\beta - 1\right) z\right)}\right). \tag{A.2d}$$

We will frequently refer to the first and the second derivative of $H(z)$ function:

$$\frac{d}{dz}H(z) = \frac{a(\gamma-1)\lambda + (\beta-1)\mu}{(\beta\mu - \lambda z)^2} - \frac{\mu(\beta-1)}{\left(\beta\mu - \lambda\beta + \lambda\left(\beta-1\right) z\right)^2}, \tag{A.3a}$$

$$\frac{d^2}{dz^2}H(z) = 2\lambda\left(\frac{a(\gamma-1)\lambda + (\beta-1)\mu}{(\beta\mu - \lambda z)^3} + \frac{\mu(\beta-1)^2}{\left(\beta\mu - \lambda\beta + \lambda\left(\beta-1\right) z\right)^3}\right). \tag{A.3b}$$

A.1.2 Proofs for Perfect Information

Proof of Proposition 1. Assume that $e_1\beta\mu \leq e_2(\beta\mu - \lambda)$. We first prove that $W^b(z)$ is decreasing function of z under the assumption. Let us take the first derivative of $W^b(z)$ with

respect to z by using product rule :

$$\frac{d}{dz}W^b(z) = \left(\frac{\lambda\beta\mu}{2}\right)\left(H^b(z)\frac{d}{dz}R(z) + R(z)\frac{d}{dz}H^b(z)\right). \tag{A.4}$$

Functions $R(z)$ and $H(z)$ are clearly positive. The first derivative of $R(z)$ is negative as a direct implication of the assumption:

$$\frac{d}{dz}R(z) = e_1 - e_2 < 0. \tag{A.5}$$

Now, we take the first derivative of $H^b(z)$ with respect to z and express it in terms of problem parameters and observe that it is negative for all $z \in [\alpha, 1)$:

$$\frac{d}{dz}H^b(z) = \frac{\lambda}{(1-\alpha)(\beta\mu - \lambda z)^2}\left(-\alpha - \frac{(\beta-1)\beta\mu(1-z)(2\beta\mu + \beta\lambda z - \beta\lambda - 2\lambda z)}{(\beta\mu - \lambda z - \lambda(1-z)\beta)^2}\right) < 0. \tag{A.6}$$

Note that $\beta > 1$, $\mu > \lambda$, and $z \in [\alpha, 1)$, the expression above is negative. By plugging inequalities (A.5 -A.6) into equation (A.4), we show that $W^b(z)$ is a decreasing function of z. To show that $W^a(z)$ is decreasing function of z under this assumption, we take the first derivative of $W^a(z)$ with respect to z:

$$\frac{d}{dz}W^a(z) = \left(\frac{\lambda\beta\mu}{2}\right)\frac{e_1\beta\mu - e_2(\beta\mu - \lambda)}{(\beta\mu - \lambda z)^2}.$$

The expression above is negative (nonpositive) for all $z \in [\alpha, 1]$ when the inequality on the higher moments is strict (non-strict). Since $W(z) = \alpha\gamma W^a(z) + (1-\alpha)W^b(z)$, i.e., $W(z)$ is summation of a positive decreasing function and nonincreasing function, then it is positive decreasing function of z, and thus $z^* = 1$ to minimize $W(z)$ under the assumption that $e_1\beta\mu \leq e_2(\beta\mu - \lambda)$. This concludes the proof. ■

Proof of Proposition 2. To show that W(z) is convex function if $e_1\beta\mu \geq e_2(\beta\mu - \lambda)$, we take the second derivative of $W(z)$ (eq. A.1c) by using the product rule:

$$\frac{d^2}{dz^2}W(z) = \frac{\lambda\beta\mu}{2}\left((e_1 - e_2)(2H'(z) + zH''(z)) + e_2H''(z)\right). \tag{A.7}$$

First, we divide interval $[\alpha, 1]$ into two disjoint intervals: $\mathcal{S}$ and $\bar{\mathcal{S}}$:

$$\mathcal{S} = \{z \in [\alpha, 1] : H'(z) > 0\},$$
$$\bar{\mathcal{S}} = \{z \in [\alpha, 1] : H'(z) \leq 0\}.$$

Next we will separately show that $W(z)$ is convex function of z on each interval.

Case I: Suppose that $z \in \bar{\mathcal{S}}$. By (A.7),

$$\begin{aligned}
\frac{d^2}{dz^2}W(z) &= \frac{\lambda\beta\mu}{2}\Big((e_1 - e_2)(2H'(z) + zH''(z)) + e_2H''(z)\Big) \\
&= \frac{\lambda\beta\mu}{2}\Big((e_1 - e_2)2H'(z) + H''(z)(ze_1 + (1-z)e_2)\Big) \\
&> 2(e_1 - e_2)2H'(z) \\
&\geq 0.
\end{aligned}$$

The first inequality is by $H''(z) > 0\ \forall z \in [\alpha, 1]$ (see eq. A.3b) and the second inequality is by $e_1 \leq e_2$ and $H'(z) \leq 0$ for all $z \in \bar{\mathcal{S}}$.

Case II: Suppose that $z \in \mathcal{S}$. By (A.7),

$$\frac{d^2}{dz^2}W(z) = \frac{\lambda\beta\mu}{2}\left((e_1 - e_2)(2H'(z) + zH''(z)) + e_2H''(z)\right). \tag{A.8}$$

By assumption:

$$e_1 \geq e_2\frac{(\beta\mu - \lambda)}{\beta\mu},$$

we immediately have

$$e_1 - e_2 \geq -e_2\frac{\lambda}{\beta\mu}. \tag{A.9}$$

Note that $(2H'(z) + zH''(z)) > 0$ for any $z \in \mathcal{S}$ since $H'(z) > 0$ and $H''(z) > 0$ for any $z \in \mathcal{S}$. After we multiply the both sides of inequality (A.9) by $(2H'(z) + zH''(z))$, we obtain

$$(e_1 - e_2)\left(2H'(z) + zH''(z)\right) \geq -e_2\frac{\lambda}{\beta\mu}(2H'(z) + zH''(z)). \tag{A.10}$$

By plugging (A.10) into (A.8), we obtain

$$\begin{aligned}\frac{d^2}{dz^2}W(z) &\geq \frac{\lambda\beta\mu}{2}e_2\left(\frac{-\lambda}{\beta\mu}\big(2H'(z)+zH''(z)\big)+H''(z)\right)\\ &= \frac{\lambda\beta\mu}{2}e_2\left(\frac{-\lambda}{\beta\mu}2H'(z)-\frac{\lambda z}{\beta\mu}H''(z)+\frac{\beta\mu}{\beta\mu}H''(z)\right)\\ &= \frac{\lambda e_2}{2}\underbrace{\Big(-2\lambda H'(z)+(\beta\mu-\lambda z)H''(z)\Big)}_{\star}.\end{aligned} \tag{A.11}$$

By using explicit expressions of $H'(z)$ and $H''(z)$ given by (A.3a,A.3b), we can write $-2\lambda H'(z)$ and $(\beta\mu-\lambda z)H''(z)$ in terms of problem parameters and decision variable z:

$$\begin{aligned}-2\lambda H'(z) &= -2\lambda\left(\frac{a(\gamma-1)\lambda+(\beta-1)\mu}{(\beta\mu-\lambda z)^2}-\frac{\mu(\beta-1)}{((\beta\mu-\lambda z-\lambda(1-z)\beta)^2}\right)\\ &= -2\lambda\frac{a(\gamma-1)\lambda+(\beta-1)\mu}{(\beta\mu-\lambda z)^2}+2\lambda\frac{\mu(\beta-1)}{((\beta\mu-\lambda z-\lambda(1-z)\beta)^2};\\ (\beta\mu-\lambda z)H''(z) &= (\beta\mu-\lambda z)2\lambda\left(\frac{a(\gamma-1)\lambda+(\beta-1)\mu}{(\beta\mu-\lambda z)^3}+\frac{\mu(\beta-1)^2}{(\beta\mu-\lambda z-\lambda(1-z)\beta)^3}\right)\\ &= 2\lambda\Bigg(\frac{a(\gamma-1)\lambda+(\beta-1)\mu}{(\beta\mu-\lambda z)^2}+\frac{\mu(\beta-1)^2}{((\beta\mu-\lambda z-\lambda(1-z)\beta)^2}\\ &\qquad\times\frac{(\beta\mu-\lambda z)}{(\beta\mu-\lambda z-\lambda(1-z)\beta)}\Bigg)\\ &\geq 2\lambda\left(\frac{a(\gamma-1)\lambda+(\beta-1)\mu}{(\beta\mu-\lambda z)^2}+\frac{\mu(\beta-1)^2}{((\beta\mu-\lambda z-\lambda(1-z)\beta)^2}\right).\end{aligned} \tag{A.12}$$

The inequality is by $(\beta\mu-\lambda z)\geq(\beta\mu-\lambda z-\lambda(1-z)\beta)>0$.

By combining terms of (A.12):

$$\begin{aligned}
\star &= -2\lambda H'(z) + (\beta\mu - \lambda z)H''(z) \\
&\geq -2\lambda \frac{a(\gamma-1)\lambda + (\beta-1)\mu}{(\beta\mu - \lambda z)^2} + 2\lambda \frac{\mu(\beta-1)}{((\beta\mu - \lambda z - \lambda(1-z)\beta)^2} \\
&\qquad + 2\lambda \frac{a(\gamma-1)\lambda + (\beta-1)\mu}{(\beta\mu - \lambda z)^2} + 2\lambda \frac{\mu(\beta-1)^2}{((\beta\mu - \lambda z - \lambda(1-z)\beta)^2} \\
&= 2\lambda \frac{\mu(\beta-1)}{((\beta\mu - \lambda z - \lambda(1-z)\beta)^2} + 2\lambda \frac{\mu(\beta-1)^2}{((\beta\mu - \lambda z - \lambda(1-z)\beta)^2} \\
&> 0.
\end{aligned} \tag{A.13}$$

The first inequality is by equation and the inequality given by (A.12); the last inequality is by $\beta > 1$.

Plugging (A.13) into (A.11), we obtain

$$\frac{d^2}{dz^2} W(z) > 0 \quad \forall z \in \mathcal{S}.$$

This concludes the proof. ■

Lemma 1. *Suppose that* $e_1\beta\mu > e_2(\beta\mu - \lambda)$. *Then*

$$\begin{aligned}
\frac{\partial}{\partial\alpha}\frac{d}{dz} W(z) > 0, \\
\frac{\partial}{\partial\gamma}\frac{d}{dz} W(z) > 0.
\end{aligned} \tag{A.14}$$

Proof of Lemma 1. Suppose that $e_1\beta\mu > e_2(\beta\mu - \lambda)$. Let us take the first derivative of $W(z)$:

$$\frac{d}{dz} W(z) = \frac{\lambda\beta\mu}{2}\Big(\big(e_1 - e_2\big)\big(H(z) + zB'(z)\big) + e_2 B'(z)\Big).$$

Next, we take the derivative of right hand side respect to γ and lower bound by zero:

$$\begin{aligned}
\frac{d}{d\gamma}\frac{d}{dz}W(z) &= \frac{\lambda\beta\mu}{2}\left((e_1-e_2)\left(\frac{d}{d\gamma}H(z)+z\frac{d}{d\gamma}H'(z)\right)+e_2\frac{d}{d\gamma}H'(z)\right)\\
&= \frac{\lambda\beta\mu}{2}\left((e_1-e_2)\left(\frac{\alpha}{\beta\mu-z\lambda}+z\frac{\alpha\lambda}{(\beta\mu-\lambda z)^2}\right)+e_2\frac{\alpha\lambda}{(\beta\mu-\lambda z)^2}\right)\\
&> \frac{\lambda\beta\mu e_2}{2}\left(-\frac{\lambda}{\beta\mu}\left(\frac{\alpha}{\beta\mu-z\lambda}+z\frac{\alpha\lambda}{(\beta\mu-\lambda z)^2}\right)+\frac{\alpha\lambda}{(\beta\mu-\lambda z)^2}\right)\\
&= \frac{\lambda\beta\mu e_2}{2}\frac{\alpha\lambda}{\beta\mu}\left(-\frac{1}{\beta\mu-z\lambda}-z\frac{\lambda}{(\beta\mu-\lambda z)^2}+\frac{\beta\mu}{(\beta\mu-\lambda z)^2}\right)\\
&= \frac{\lambda\beta\mu e_2}{2}\frac{\alpha\lambda}{\beta\mu}\left(\frac{(-(\beta\mu-\lambda z)-\lambda z+\beta\mu}{(\beta\mu-\lambda z)^2}\right)\\
&= 0.
\end{aligned}$$

The inequality is holds as direct implication of the assumption on second moments. Similarly, we can also show that

$$\begin{aligned}
\frac{d}{d\alpha}\frac{d}{dz}W(z) &= \frac{\lambda\beta\mu}{2}\left((e_1-e_2)\left(\frac{d}{d\alpha}H(z)+z\frac{d}{d\alpha}B'(z)\right)+e_2\frac{d}{d\alpha}B'(z)\right)\\
&= \frac{\lambda\beta\mu}{2}\left((e_1-e_2)\left(\frac{(\gamma-1)}{\beta\mu-z\lambda}+z\frac{(\gamma-1)\lambda}{(\beta\mu-\lambda z)^2}\right)+e_2\frac{(\gamma-1)\lambda}{(\beta\mu-\lambda z)^2}\right)\\
&> \frac{\lambda\beta\mu e_2}{2}\left(-\frac{\lambda}{\beta\mu}\left(\frac{(\gamma-1)}{\beta\mu-z\lambda}+z\frac{(\gamma-1)\lambda}{(\beta\mu-\lambda z)^2}\right)+\frac{(\gamma-1)\lambda}{(\beta\mu-\lambda z)^2}\right)\\
&= \frac{\lambda\beta\mu e_2}{2}\frac{(\gamma-1)\lambda}{\beta\mu}\left(-\frac{1}{\beta\mu-z\lambda}-z\frac{\lambda}{(\beta\mu-\lambda z)^2}+\frac{\beta\mu}{(\beta\mu-\lambda z)^2}\right)\\
&= \frac{\lambda\beta\mu e_2}{2}\frac{(\gamma-1)\lambda}{\beta\mu}\left(\frac{(-(\beta\mu-\lambda z)-\lambda z+\beta\mu}{(\beta\mu-\lambda z)^2}\right)\\
&= 0.
\end{aligned}$$

■

Proof of Proposition 3. The notation $\mathcal{M}_B$ denote the optimal misclassification rate among type-B customers for a given set of system parameters:

$$\mathcal{M}_B = \frac{z^*(\alpha)-\alpha}{1-\alpha}.$$

Case I: Suppose that $e_1\beta\mu \le e_2(\beta\mu-\lambda)$. By Proposition 1, $W(z)$ is decreasing function of

z on interval $[\alpha, 1]$ and thus $z^* = 1$ for any α and γ, implying optimal misclassification rate is hundred percent.

Case II: Suppose that $e_1\beta\mu > e_2(\beta\mu - \lambda)$. By Lemma 1, $W'(z)$ increases as γ and α increases. Thus, $W'(z)$ with higher γ or/and α is always positive or will hit the zero earlier, implying that z^* decreases as γ or/and α increases. As γ increases, z^* decreases, and thus misclassification rate decreases. To evaluate how change in α impacts the misclassification rate, we take the first derivative of $\mathcal{M}_B$ with respect to α:

$$\frac{d}{d\alpha}\mathcal{M}_B = \frac{\left(\frac{d}{d\alpha}z^*(\alpha) - 1\right)(1-\alpha) + (z^*(\alpha) - \alpha)}{(1-\alpha)^2}.$$

We upper-bound the numerator of the expression on the right hand can by zero by using the following inequalities:

$$\frac{d}{d\alpha}z^*(\alpha) < 0 \quad \text{by Lemma 1,}$$
$$z^* \leq 1 \quad \text{by feasibility.}$$

We conclude that misclassification rate decreases as α increases. ■

A.1.3 Proofs for Imperfect Information

For notational convenience, we define $W^a(x)$ and $W^b(x)$ as follows:

$$W^a(x) = \frac{\lambda\beta\mu}{2}R(x)A(x); \tag{A.15a}$$
$$W^b(x) = \frac{\lambda\beta\mu}{2}R(x)B(x), \tag{A.15b}$$

where

$$R(x) = G(x)(e_1 - e_2) + e_2, \tag{A.16a}$$
$$A(x) = \frac{1}{\beta\mu - \lambda G(x)}\left(g(x) + \frac{\beta\mu(1-g(x))}{\beta\mu - \lambda\beta + \lambda(\beta-1)G(x)}\right), \tag{A.16b}$$
$$B(x) = \frac{1}{\beta\mu - \lambda G(x)}\left(x + \frac{\beta\mu(1-x)}{\beta\mu - \lambda\beta + \lambda(\beta-1)G(x)}\right). \tag{A.16c}$$

Then we plig (2.11) into functions $R(x)$, $A(x)$, and $P(x)$ (see A.16) and obtain:

$$\begin{aligned}
R(x) &= (\alpha(f+x\kappa)+(1-\alpha)x)\,(e_1-e_2)+e_2 \\
A(x) &= \frac{1}{\beta\mu-\lambda\,(\alpha(f+x\kappa)+(1-\alpha)x)}\left((f+x\kappa)+\frac{\beta\mu(1-(f+x\kappa))}{\beta\mu-\lambda\beta+\lambda\,(\beta-1)\,(\alpha(f+x\kappa)+(1-\alpha)x)}\right), \\
B(x) &= \frac{1}{\beta\mu-\lambda\,(\alpha(f+x\kappa)+(1-\alpha)x)}\left(x+\frac{\beta\mu(1-x)}{\beta\mu-\lambda\beta+\lambda\,(\beta-1)\,(\alpha(f+x\kappa)+(1-\alpha)x)}\right).
\end{aligned} \tag{A.17}$$

For notational convenience, we define

$$A(x) = \frac{N_a(x)}{D(x)}, \tag{A.18a}$$

$$B(x) = \frac{N_b(x)}{D(x)}, \tag{A.18b}$$

where

$$\begin{aligned}
N_a(x) &= \beta\mu+(f+\kappa x)\Big(-\lambda\beta+\lambda\,(\beta-1)\,(\alpha(f+x\kappa)+(1-\alpha)x)\Big), \\
N_b(x) &= \beta\mu+x\Big(-\lambda\beta+\lambda\,(\beta-1)\,(\alpha(f+x\kappa)+(1-\alpha)x)\Big), \\
D(x) &= \Big(\beta\mu-\lambda\,(\alpha(f+x\kappa)+(1-\alpha)x)\Big) \\
&\qquad\times\Big(\beta\mu-\lambda\beta+\lambda\,(\beta-1)\,(\alpha(f+x\kappa)+(1-\alpha)x)\Big),
\end{aligned} \tag{A.19}$$

Let us take the derivative of $A(x)$ and $B(x)$:

$$\frac{d}{dx}A(x) = \frac{\phi_a(x)}{D^2(x)}, \tag{A.20a}$$

$$\frac{d}{dx}B(x) = \frac{\phi_b(x)}{D^2(x)}, \tag{A.20b}$$

where

$$\phi_a(x) = N_a'(x)D(x) - D'(x)N_a(x), \tag{A.21a}$$

$$\phi_b(x) = N_b'(x)D(x) - D'(x)N_b(x). \tag{A.21b}$$

Before we get into the proof of Proposition 5, we will prove two lemmas related to the gradient of $W^a(x)$. First, we will show that function $\phi_a(x)$ is an increasing function of x on

any interval and it takes only negative values as long as the the slope of ROC curve is equal or greater than 1 on the interval that it is evaluated. Note that derivative of $g(x)$ is not defined at breakpoints and thus function $\phi_a(x)$ can be discontinuous at breakpoints. Next, we will describe how function $\phi_a(x)$ behaves at breakpoints.

Suppose that $x \in \mathcal{I}_n = (x_{n-1}, x_n]$; and $g(x) = f + \kappa x$, where f is defined by equation (2.12). Since $g(x_n)$ cannot be greater than 1, $x_n \leq \frac{1-f}{\kappa}$. Instead of evaluating $\phi_a(x)$ on interval $(x_{n-1}, x_n]$, we will evaluate on interval $\left[0, \frac{1-f}{\kappa}\right]$ since $g(x)$ and thus $\phi_a(x)$ will exactly take same values on the interval of our interest.

Lemma 2. *Suppose that $g(x) = f + \kappa x$, where $x \in [\underline{x}, \bar{x}]$, where $\bar{x} = \frac{1-f}{\kappa}$. Function $\phi_a(x)$ is an increasing function of x on interval $[\underline{x}, \bar{x}]$. Suppose that $\kappa \geq 1$, then $\phi_a(x) < 0$ for all $x \neq 1$ and $\phi_a(x) \leq 0$ for $x = 1$ on that interval.*

Proof of Lemma 2. Let us take the derivative of $\phi_a(x)$:

$$\frac{d}{dx}\phi_a(x) = N_a''(x)D(x) - D''(x)N_a(x),$$

where,

$$\begin{aligned} N_a''(x) &= 2(\beta-1)(1-\alpha+\alpha\kappa)\kappa\lambda > 0, \\ D''(x) &= -2(\beta-1)(1-\alpha+\alpha\kappa)^2\lambda^2 < 0. \end{aligned} \tag{A.22}$$

By (A.22), it is clear that $\phi_a'(x) > 0$, and thus

$$\phi_a(x) \leq \phi_a\left(\frac{1-f}{\kappa}\right) \quad \forall x \in \left[\underline{x}, \frac{1-f}{\kappa}\right].$$

Now, we will show that $\phi_a(\bar{x})$ is negative, where $\bar{x} = \frac{1-f}{\kappa}$ given that $\kappa \geq 1$. After some algebraic manipulation, $\phi_a(\bar{x})$ can be written as

$$\phi_a\left(\frac{1-f}{\kappa}\right) = M_1 M_2, \tag{A.23}$$

where

$$M_1 = \lambda(1-\alpha)\beta(\mu-\lambda) + (\beta-1)(\alpha+\bar{x}(1-\alpha)) \tag{A.24a}$$

$$M_2 = -\beta^2\kappa\mu(1-\bar{x}) + (\beta-1)\lambda(1-\bar{x})(1-\kappa\bar{x})\alpha + \beta\left(\mu-\lambda+\lambda\bar{x}\right)(1-\kappa\bar{x}) - \lambda\bar{x}(1-\kappa\bar{x}). \tag{A.24b}$$

It is clear that $M_1 > 0$ because $\alpha \in [0,1)$, $\mu > \lambda$, and $\beta > 1$. Now, we will show that M_2 is negative. Note that $g(\bar{x}) = f + \kappa\bar{x} \leq 1$, and thus $\kappa\bar{x} \leq 1$ and observe that following inequalities hold:

$$\begin{aligned} -\lambda\bar{x}(1-\kappa\bar{x}) &\leq 0, \\ (\beta-1)\lambda(1-\bar{x})(1-\kappa\bar{x})\alpha &\leq (\beta-1)\lambda(1-\bar{x})(1-\kappa\bar{x}). \end{aligned} \tag{A.25}$$

Next, we put an upper bound on $\phi_a(\bar{x})$:

$$\begin{aligned} \phi_a(\bar{x}) &\leq -\beta^2\kappa\mu(1-\bar{x}) + (\beta-1)\lambda(1-\bar{x})(1-\kappa\bar{x}) + \beta(\mu-\lambda+\lambda\bar{x})(1-\kappa\bar{x})) \\ &= -\beta^2\kappa\mu(1-\bar{x}) + (\beta-1)\lambda(1-\bar{x})(1-\kappa\bar{x}) + \beta\mu(1-\kappa\bar{x}) - \beta\lambda(1-\bar{x})(1-\kappa\bar{x}) \\ &= -\beta^2\kappa\mu(1-\bar{x}) + \beta\mu(1-\kappa\bar{x}) - \lambda(1-\bar{x})(1-\kappa\bar{x}) \\ &= -\beta\mu\left(\beta\kappa(1-\bar{x}) - (1-\kappa\bar{x})\right) - \lambda(1-\bar{x})(1-\kappa\bar{x}) \\ &\leq -\beta\mu\left(\beta\kappa(1-\bar{x}) - (1-\kappa\bar{x})\right) \\ &\leq 0. \end{aligned}$$

The first inequality is by plugging inequalities given by (A.25) into (A.24b). The second inequality is trivial. The last inequality holds because $\kappa \geq 1$ and $\beta > 1$ allowing the following inequality hold:

$$\beta\kappa\left(1-\bar{x}\right) \geq 1-\bar{x} \geq 1-\kappa x \geq 0.$$

Observe that if $\bar{x} \neq 1$, then the first inequality is strict, i.e. $\phi_a(\bar{x}) < 0$ for all $\bar{x} \neq 1$. This concludes the proof. ■

Now we need to check how $\phi_a(x)$ changes at the breakpoints. Suppose that (r, w) is the n^{th} breakpoint, i.e., $(x_n, g(x_n)) = (r, w)$. To compare the value of left and right function at this breakpoint, we will take the derivative of function $\phi_a(r)$ with respect to κ.

Lemma 3. *Fix $x = r$ and $g(x) = w$. Suppose that $r \neq 1$.*

$$\frac{d}{d\kappa}\phi_a(r) = \mu^3\rho\Big(\beta + \rho q_1\Big)\Big(\big(q_1q_2 + \alpha\beta + \alpha w\rho q_1\big) + \big(\beta - 1\big)\alpha\beta q_2\big(w - 1\big)\Big) < 0,$$

where $h = \alpha w + (1 - \alpha)r$ and $\rho = \frac{\lambda}{\mu}$; $q_1 = -\beta + (\beta - 1)h$ and $q_2 = \beta - \rho h$.

Proof of Lemma 3. The value of function $\phi_a(x)$ at point (r, w) is given by (A.21a):

$$\phi_a(r) = N_a'(r)D(r) - D'(r)N_a(r).$$

We take the derivative of $\phi_a(r)$ with respect to κ. Noting that $N_a'(r)$ and $D'(r)$ are functions of κ while $N_a(r)$ and $D(r)$ are not, we have

$$\frac{d}{d\kappa}\phi_a(r) = D(r)\frac{d}{d\kappa}\left(\frac{dN_a(x)}{dx}\Big|_{x=r}\right) - N_a(r)\frac{d}{d\kappa}\left(\frac{dD(x)}{dx}\Big|_{x=r}\right). \tag{A.26}$$

After some algebraic manipulation, (A.26) can be written as

$$\frac{d}{d\kappa}\phi_a(r) = \mu^3\rho\Big(\underbrace{\big(\beta + \rho q_1\big)}_{\star}\underbrace{\big(q_1q_2 + \alpha\beta + \alpha w\rho q_1\big)}_{\star\star} + \underbrace{\big(\beta - 1\big)\,\alpha\beta q_2\,\big(w - 1\big)}_{\star\star\star}\Big).$$

Note that $0 > q_1 > -\beta$ and $\rho < 1$ and thus $\star$ is a positive expression. On the other hand, $\star\star\star$ is a nonpositive expression since $\beta > 1$, $\rho < 1$, $h \leq 1$ and $w \leq 1$. If we show that $\star\star$ is

also nonpositive, then we are done with the proof.

$$\begin{aligned}
\star\star &= \Big(q_1 q_2 + \alpha\beta + \alpha w\rho q_1\Big) \\
&= \Big(q_1 q_2 + \alpha\left(\beta + w\rho q_1\right)\Big) \\
&\leq \Big(q_1 q_2 + \beta + w\rho q_1\Big) \\
&= \Big(q_1(q_2 + w\rho) + \beta\Big) \\
&= \Big(\big(-\beta + (\beta-1)h\big)(\beta - \rho h + w\rho) + \beta\Big) \\
&= \Big(-\beta^2 + (\beta^2-\beta)h + \big(-\beta + (\beta-1)h\big)\rho\big(w-h\big) + \beta\Big) \\
&= \Big((\beta^2-\beta)(h-1) + \big(-\beta + (\beta-1)h\big)\rho\big(w-h\big)\Big) \\
&< 0.
\end{aligned}$$

The first inequality by the fact that $\beta + w\rho q_1$ are positive ($q_1 \geq -\beta$ and $\rho < 1$ and $w \leq 1$) and thus $\star\star$ increases as α increases. The last inequality is follows from $r \leq h = \alpha w + (1-\alpha)r < w \leq 1$ and $\beta > 1$. ■

Corollary 3. *$\phi_a(x)$ is an increasing function of x at all points including breakpoints. Moreover, $\phi_a(x) < 0$ for all $x \in [0, x_{\bar{n}}]$, where $\bar{n} = \max\{n : \kappa_n \geq 1\}$.*

It is possible that $\phi_a(x)$ is discontinuous at $x = r$. Therefore, to show that $\phi_a(r) < \phi_a(r+\epsilon)$, we need to show that the value of the right function is higher than the value of the left function at $x = r$. This is sufficient since $\phi_a(x)$ is increasing function of x on any interval by Lemma 2 and, thus $\phi_a(r+\epsilon)$ must be greater than the value of the right function at $x = r$. By Lemma 3, we know that as κ decreases the value of $\phi_a(x)$ function increases for a given $x = r$. Since κ value associated with right function is no greater than the κ value associated with left function, we conclude that $\phi_a(r) < \phi_a(r+\epsilon)$.

Proof of Proposition 5. Recall eq. 2.6b:

$$W^a(x) = \frac{\lambda\beta\mu}{2} R(x)A(x).$$

If we can show that $A(x)$ and $R(x)$ are nonnegative decreasing function of x where $x \in [0, x_{\bar{n}}]$,

then we are done. By equation (A.20a), the first derivative of A(x) is defined as

$$\frac{d}{dx}A(x) = \frac{\phi_a(x)}{(D(x))^2}.$$

By Corollary 3, we know that $\phi_a(x) < 0$ for all $x \in [0, x_{\bar{n}}]$. Therefore, $A(x)$ is a nonnegative decreasing function of x. It is straightforward that

$$R(x) = (\alpha g(x) + (1-\alpha)x)\,(e_1 - e_2) + e_2$$

is positive nonincreasing function of x for all $x \in [0, 1]$ because $e_1 \leq e_2$ by assumption. The function generated via multiplication of a positive nonincreasing and a positive decreasing function is decreasing. This concludes the proof. ■

Recall that the average waiting cost of a patient is given by

$$W(x) = \gamma\alpha W^a(x) + (1-\alpha)W^b(x) \text{ (see eq. 2.7).} \tag{A.27}$$

By (2.6b-2.6d) and (A.18a-A.18b), we have

$$\begin{aligned} W(x) &= \frac{\lambda\beta\mu}{2}R(x)\left(\gamma\alpha A(x) + (1-\alpha)P(x)\right), \\ &= \frac{\lambda\beta\mu}{2}R(x)\left(\gamma\alpha\frac{N_a(x)}{D(x)} + (1-\alpha)\frac{N_b(x)}{D(x)}\right). \end{aligned}$$

Let us take the derivative of $W(x)$:

$$\begin{aligned} \frac{d}{dx}W(x) &= \frac{\lambda\beta\mu}{2}\left(R'(x)\frac{(\gamma\alpha N_a(x) + (1-\alpha)N_b(x))}{D(x)} + R(x)\frac{(\gamma\alpha\phi_a(x) + (1-\alpha)\phi_b(x))}{D^2(x)}\right) \\ &= \frac{\lambda\beta\mu}{2}\left(\frac{R'(x)D(x)\left(\gamma\alpha N_a(x) + (1-\alpha)N_b(x)\right) + (\gamma\alpha\phi_a(x) + (1-\alpha)\phi_b(x))}{D^2(x)}\right) \\ &= \frac{\lambda\beta\mu}{2}\left(\frac{\psi(x)}{D^2(x)}\right), \end{aligned} \tag{A.28}$$

where,

$$\psi(x) = R'(x)D(x)N(x) + R(x)\phi(x), \tag{A.29a}$$

$$N(x) = \gamma\alpha N_a(x) + (1-\alpha)N_b(x), \tag{A.29b}$$

$$\phi(x) = N'(x)D(x) - D'(x)N(x), \tag{A.29c}$$

$$\phi(x) = \gamma\alpha\phi_a(x) + (1-\alpha)\phi_b(x). \tag{A.29d}$$

To comment on the properties of $W(x)$, we first need to understand how $\psi(x)$ function behaves. We have explicit expressions for $R'(x)$, $D(x)$, $N(x)$, and $R(x)$ while $\phi(x)$ is hard to express explicitly. We write $\phi(x)$ as weighted summation of functions $\phi_a(x)$ and $\phi_b(x)$. The following lemma shows that $\phi(x)$ increases as x increases on a given interval.

Lemma 4. *Suppose that $g(x) = f + \kappa x$, where $x \in [\underline{x}, \bar{x}]$, where $\bar{x} = \frac{1-f}{\kappa}$. Function $\phi(x)$ is an increasing function of x on interval $[\underline{x}, \bar{x}]$. If $\kappa \geq 1$, then $\phi(x)$ takes only negative values.*

Proof of Lemma 4. Let us take the derivative of $\phi_a(x)$:

$$\frac{d}{dx}\phi_b(x) = N_b''(x)D(x) - D''(x)N_b(x). \tag{A.30}$$

Let us take the second derivative of $N_b(x)$ and $D(x)$:

$$\begin{aligned} N_b''(x) &= 2(\beta-1)(1-\alpha+\alpha\kappa)\lambda > 0, \\ D''(x) &= -2(\beta-1)(1-\alpha+\alpha\kappa)^2\lambda^2 < 0. \end{aligned} \tag{A.31}$$

By (A.30) and (A.31), $\phi_b(x)$ is increasing function of x on any interval. Thus,

$$\phi_b(x) \leq \phi_b\left(\frac{1-f}{\kappa}\right) \quad \forall x \in \left[0, \frac{1-f}{\kappa}\right].$$

This inequality with together with Lemma 2 implies that

$$\begin{aligned} \phi(x) &= \gamma\alpha\phi_a(x) + (1-\alpha)\phi_b(x)\phi_b(x) \\ &\leq \gamma\alpha\phi_a\left(\frac{1-f}{\kappa}\right) + (1-\alpha)\phi_b\left(\frac{1-f}{\kappa}\right) \\ &= (\gamma-1)\alpha\phi_a\left(\frac{1-f}{\kappa}\right) + \alpha\phi_a\left(\frac{1-f}{\kappa}\right)(1-\alpha)\phi_b\left(\frac{1-f}{\kappa}\right) \\ &\leq \underbrace{\alpha\phi_a\left(\frac{1-f}{\kappa}\right) + (1-\alpha)\phi_b\left(\frac{1-f}{\kappa}\right)}_{\star} \quad \forall x \in \left[\underline{x}, \frac{1-f}{\kappa}\right]. \end{aligned}$$

The first inequality by $\phi_a(x)$ and $\phi_b(x)$ being increasing with x on any interval. The second inequality is by the fact that (i) $\phi_a(x)$ is negative for any associated κ value that is equal to or greater than 1 (See Lemma 2) and (ii) $\gamma > 1$ by assumption. After some algebraic manipulation $\star$ can be explicitly expressed as

$$\star = \underbrace{\beta\lambda\mu(1-\alpha)(1-\bar{x})(\beta-1)(\alpha\kappa+1-\alpha)}_{\dagger}\underbrace{\Big(\beta\lambda(1-\alpha)(1-\bar{x})+2\lambda(\alpha+\bar{x}-\alpha\bar{x})-2\beta\mu\Big)}_{\ddagger}$$

Observe that $\dagger$ is nonnegative. If we can show that $\ddagger$ is negative, then we are done.

$$\begin{aligned}
\ddagger &= \beta\lambda(1-\alpha)(1-\bar{x})+2\lambda(\alpha+\bar{x}-\alpha\bar{x})-2\beta\mu \\
&< \beta\lambda(1-\alpha)(1-\bar{x})+2\lambda(\alpha+\bar{x}-\alpha\bar{x})-2\beta\lambda \\
&< \lambda(1-\alpha)(1-\bar{x})+2\lambda(\alpha+\bar{x}-\alpha\bar{x})-2\lambda \\
&= \lambda(1-\alpha)(1-\bar{x})-2\lambda(1-\alpha)(1-\bar{x}) \\
&\leq 0.
\end{aligned}$$

The first inequality holds because $\ddagger$ decreases as μ increases and $\mu > \lambda$ by assumption. The second inequality holds because the right hand side of the inequality decreases as β increases. The inequality is by rearranging terms. The last inequality is trivial. ■

Suppose that (r, w) is a breakpoint where $x_n = r$ and $g(x_n) = w$. Next we will derive an explicit expression of the derivative of function $\phi_b(r)$ with respect to κ. Then use this expression together with Lemma 3 to analyze how function $\phi(x)$ behaves at the break point (r, w)

Lemma 5. *Fix $x = r$ and $g(x) = w$.*

$$\frac{d}{d\kappa}\phi_b(r) = \rho\mu^3\left(\beta\left(\beta-1\right)\alpha q_2\left(r-1\right)+\left(\beta+r\rho q_1\right)\alpha\left(\beta+\rho q_1\right)\right),$$

where $r < h = \alpha w + (1-\alpha)r < w \leq$ and $\rho = \frac{\lambda}{\mu}$; $q_1 = -\beta + (\beta-1)h$ and $q_2 = \beta - \rho h$.

Proof of Lemma 5. Let us first express $\phi_b(r)$ in terms of $N_b(r)$, $N_b'(r)$ $D(r)$, and $D'(r)$. For

this purpose, we explicitly write $N_b(x)$, $N_b'(x)$, $D(x)$, and $D'(x)$:

$$\begin{aligned}
N_b(x) &= \beta\mu + x\Big(-\lambda\beta + \lambda\left(\beta-1\right)\left(\alpha(f+x\kappa)+(1-\alpha)x\right)\Big),\\
N_b'(x) &= \Big(-\lambda\beta + \lambda\left(\beta-1\right)\left(\alpha(f+x\kappa)+(1-\alpha)x\right)\Big) + x\lambda\left(\beta-1\right)\left(\alpha\kappa+1-\alpha\right),\\
D(x) &= \Big(\beta\mu - \lambda\left(\alpha(f+x\kappa)+(1-\alpha)x\right)\Big)\Big(\beta\mu - \lambda\beta + \lambda\left(\beta-1\right)\left(\alpha(f+x\kappa)+(1-\alpha)x\right)\Big)\\
D'(x) &= -\lambda\left(\alpha\kappa+1-\alpha\right)\Big(\beta\mu - \lambda\beta + \lambda\left(\beta-1\right)\left(\alpha(f+x\kappa)+(1-\alpha)x\right)\Big)\\
&\qquad + \left(\beta\mu - \lambda\left(\alpha(f+x\kappa)+(1-\alpha)x\right)\right)\left(\lambda\left(\beta-1\right)\left(\alpha\kappa+1-\alpha\right)\right).
\end{aligned} \tag{A.32}$$

At $(x, g(x)) = (r, w)$,

$$f + \kappa x = w,$$

and value of functions given by (A.32) are

$$\begin{aligned}
N_b(r) &= \beta\mu + r\Big(-\lambda\beta + \lambda\left(\beta-1\right)h\Big),\\
N_b'(r) &= \left(-\lambda\beta + \lambda\left(\beta-1\right)h\right) + r\lambda\left(\beta-1\right)\left(\alpha\kappa+(1-\alpha)\right),\\
D(r) &= \left(\beta\mu - \lambda h\right)\left(\beta\mu - \lambda\beta + \lambda\left(\beta-1\right)h\right)\\
D'(r) &= -\lambda\left(\alpha\kappa+1-\alpha\right)\left(\beta\mu - \lambda\beta + \lambda\left(\beta-1\right)h\right)\\
&\qquad + \left(\beta\mu - \lambda h\right)\left(\lambda\left(\beta-1\right)\left(\alpha\kappa+1-\alpha\right)\right).
\end{aligned}$$

Note that $N_b(r)$ and $D(r)$ independent of κ and thus their derivative with respect to κ is zero. Then the derivative of $\phi_b(r)$ can be written as

$$\frac{d}{d\kappa}\phi_b(r) = \left(\frac{d}{d\kappa}N_b'(r)\right)D(r) - \left(\frac{d}{d\kappa}D'(r)\right)N_b(r) \tag{A.33}$$

After some algebraic manipulation, (A.33) can be written as

$$\begin{aligned}
\frac{d}{d\kappa}\phi_b(r) &= \rho\mu^3\Big(r\left(\beta-1\right)\alpha q_2\left(\beta+\rho q_1\right) - \big(-\alpha\left(\beta+\rho q_1\right) + q_2\left(\beta-1\right)\alpha\big)\big(\beta+r\rho q_1\big)\Big)\\
&= \rho\mu^3\Big(r\left(\beta-1\right)\alpha q_2\left(\beta+\rho q_1\right) + \left(\beta+r\rho q_1\right)\alpha\left(\beta+\rho q_1\right)\\
&\qquad - \left(\beta+r\rho q_1\right)q_2\left(\beta-1\right)\alpha\Big)\\
&= \rho\mu^3\Big(\beta\left(\beta-1\right)\alpha q_2\left(r-1\right) + \left(\beta+r\rho q_1\right)\alpha\left(\beta+\rho q_1\right)\Big).
\end{aligned}$$

■

The following lemma helps us understand how $\phi(x)$ behaves at breakpoints.

Lemma 6. *Fix $x = r$ and $g(x) = w$. Suppose that $r \neq 1$.*

$$\frac{d}{d\kappa}\phi(r) \leq 0.$$

Proof of Lemma 6. By using Lemma 3 and Lemma 5, we can write $\frac{d}{d\kappa}\phi(r)$ as

$$\begin{aligned}
\frac{d}{d\kappa}\phi(r) &= \gamma\alpha\frac{d}{d\kappa}\phi_a(r) + (1-\alpha)\frac{d}{d\kappa}\phi_b(r) \\
&\leq \alpha\frac{d}{d\kappa}\phi_a(r) + (1-\alpha)\frac{d}{d\kappa}\phi_b(r) \\
&= \alpha\mu^3\rho\Big((\beta+\rho q_1)(q_1q_2+\alpha\beta+\alpha w\rho q_1) + (\beta-1)\alpha\beta q_2(w-1)\Big) \\
&\quad + (1-\alpha)\rho\mu^3\Big(\beta(\beta-1)\alpha q_2(r-1) + (\beta+r\rho q_1)\alpha(\beta+\rho q_1)\Big).
\end{aligned} \tag{A.34}$$

The inequality is by $\frac{d}{d\kappa}\phi_a(r)$ is negative (as shown by Lemma 3) and $\gamma > 1$. Note that $r \leq h = \alpha w + (1-\alpha)r \leq w$ and $\rho = \frac{\lambda}{\mu}$; $q_1 = -\beta + (\beta-1)h$ and $q_2 = \beta - \rho h$. Then we have

$$\begin{aligned}
\alpha\mu^3\rho\Big((\beta-1)\alpha\beta q_2(w-1)\Big) &\leq 0 \\
(1-\alpha)\rho\mu^3\Big(\beta(\beta-1)\alpha q_2(r-1)\Big) &\leq 0.
\end{aligned} \tag{A.35}$$

By plugging inequalities (A.35) into (A.34), we obtain

$$\frac{d}{d\kappa}\phi(r) \leq \alpha\mu^3\rho\Big(\beta+\rho q_1\Big)\Big(q_1q_2+\alpha\beta+\alpha w\rho q_1\Big)+(1-\alpha)\rho\mu^3\Big((\beta+r\rho q_1)\alpha(\beta+\rho q_1)\Big). \tag{A.36}$$

After rearranging the terms on the right hand side of (A.36), we have

$$\frac{d}{d\kappa}\phi(r) \leq \alpha\mu^3\rho\Big(\beta+\rho q_1\Big)\Big(q_1q_2 + (\alpha w + (1-\alpha)r)\rho q_1 + \beta\Big).$$

Recall that $h = (\alpha w + (1-\alpha)r)$; $q_1 = -\beta + (\beta - 1)h$ and $q_2 = \beta - \rho h$.

$$\begin{aligned}
\frac{d}{d\kappa}\phi(r) &\leq \alpha\mu^3\rho\Big(\beta + \rho q_1\Big)\left(q_1 q_2 + (\alpha w + (1-\alpha)r)\rho q_1 + \beta\right) \\
&= \alpha\mu^3\rho\Big(\beta + \rho q_1\Big)\left(q_1(q_2 + h\rho) + \beta\right) \quad \text{by } h = (\alpha w + (1-\alpha)r) \\
&= \alpha\mu^3\rho\Big(\beta + \rho q_1\Big)\left(q_1\beta + \beta\right) \quad \text{by } q_2 = \beta - \rho h \\
&= \alpha\mu^3\rho\Big(\beta + \rho q_1\Big)\beta\left(q_1 + 1\right) \\
&= \alpha\mu^3\rho\Big(\beta + \rho q_1\Big)\beta\left(-\beta + (\beta - 1)h + 1\right) \quad \text{by } q_1 = -\beta + (\beta - 1)h \\
&= \alpha\mu^3\rho\Big(\beta + \rho q_1\Big)\beta(1-h)(1-\beta) \\
&< 0.
\end{aligned}$$

The last inequality by the fact that $\alpha in\,(0,1)$, $h < 1$, $\beta > 1$ and $\beta + \rho q_1 > 0$. ■

Corollary 4.

$$\frac{d}{dx}\phi(x) > 0 \quad \forall x \in [0,1).$$

By Lemma 4, function $\phi(x)$ increases on a given interval. Suppose that $x_n = r$ and $g(x_n) = w$ for some n. By Lemma 6, we know that $\frac{d}{d\kappa}\phi(r) \leq 0$, i.e.,function $\phi(r)$ increases as the associated κ value decreases. This implies that the value of the right function is higher than the value of the left function at breakpoint $x = r$. We can conclude that $\phi(x)$ increases as x increases, including at breakpoints.

Lemma 7. *Suppose that (r, w) is a breakpoint where $x_n = r$ and $g(x_n) = w$*

$$\frac{d}{d\kappa}\psi(r) < 0.$$

Proof. By (A.29a),

$$\psi(r) = \left(\frac{dR(x)}{dx}\Big|_{x=r}\right) D(r)N(r) + R(r)\phi(r).$$

Note that $D(r)$, $N(r)$ and $R(r)$ are positive and independent of κ.

$$\begin{aligned}
\frac{d}{d\kappa}\psi(r) &= D(r)N(r)\frac{d}{d\kappa}\left(\frac{dR(x)}{dx}\Big|_{x=r}\right) + R(r)\frac{d}{d\kappa}\phi(r) \\
&= D(r)N(r)\frac{d}{d\kappa}\left(\alpha\kappa + 1 - \alpha\right)\left(e_1 - e_2\right) + R(r)\frac{d}{d\kappa}\phi(r) \\
&= D(r)N(r)\alpha\left(e_1 - e_2\right) + R(r)\frac{d}{d\kappa}\phi(r) \\
&\leq D(r)N(r)\alpha\left(e_1 - e_2\right) \\
&\leq 0.
\end{aligned}$$

The first inequality follows from Lemma 6 and the second inequality holds because $e_1 \leq e_2$ by assumption. ■

Lemma 8. *Whenever $\psi(x) \geq 0$, $\psi'(x) > 0$ on a given interval $\mathcal{I}$.*

Proof of Lemma 8. By definition of $\psi(x)$ (see eq. A.29a),

$$\psi(x) \geq 0 \quad \text{iff} \quad \frac{R'(x)}{R(x)} \geq -\frac{\phi(x)}{N(x)D(x)}. \tag{A.37}$$

Recall that $R(x) = \left(\alpha\left(f + \kappa x\right) + (1-\alpha)x\right)\left(e_1 - e_2\right) + e_2$ is positive decreasing function of x since $e_1 \leq e_2$ $\forall x \in [0,1]$. Thus,

$$\frac{R'(x)}{R(x)} \leq 0 \ \forall x \in [0,1]\,.$$

Observe that if $\phi(x) < 0$, then the negation of the statement on the right hand will immediately hold. Since it is "iff", the negation of the statement on the left hand side must hold, i.e., $\psi(x) < 0$. Therefore, $W(x)$ is a decreasing function of x on that interval. Therefore, we focus on the subinterval that $\phi(x) \geq 0$.

$$\mathcal{I}_s = \{x \in \mathcal{I} : \phi(x) \geq 0\}\,.$$

Let us take the first derivative of $\psi(x)$ (see eq. A.29a) by using the product rule:

$$\frac{d}{dx}\psi(x) = R'(x)\frac{d}{dx}\Big(N(x)D(x)\Big) + R''(x)N(x)D(x) + R'(x)\phi(x) + R(x)\phi'(x). \tag{A.38}$$

By using the equality $R''(x) = 0$, and plugging (A.29c) into (A.38), we obtain

$$\begin{aligned}\frac{d}{dx}\psi(x) &= R'(x)\frac{d}{dx}\Big(N(x)D(x)\Big) + R'(x)\Big(N'(x)D(x) - D'(x)N(x)\Big) + R(x)\phi'(x)\\ &= R'(x)\Big(N'(x)D(x) + D'(x)N(x) + N'(x)D(x) - D'(x)N(x)\Big) + R(x)\phi'(x)\\ &= 2R'(x)N'(x)D(x) + R(x)\phi'(x).\end{aligned} \tag{A.39}$$

Next, we will show that $\psi'(x) > 0$ whenever $\psi(x) \geq 0$ by examining two complementary cases: (i) $N'(x)D(x) \leq 0$ and (ii)$N'(x)D(x) > 0$.

Case I: Suppose that $N'(x)D(x) \leq 0$. Plug this inequality into (A.39). then $\psi'(x) > 0$ since $R'(x) \leq 0$ and $R(x) > 0$ and $\phi'(x) > 0$ (see Corollary 4).

Case II: Suppose that $N'(x)D(x) > 0$. By (A.39), it is clear that

$$\frac{d}{dx}\psi(x) > 0 \quad \text{iff} \quad \frac{R'(x)}{R(x)} > -\frac{\phi'(x)}{2N'(x)D(x)}. \tag{A.40}$$

By Lemma 9(which we will prove afterwards), we have

$$-\frac{\phi(x)}{N(x)D(x)} > -\frac{\phi'(x)}{2N'(x)D(x)} \quad \forall x \in \mathcal{I}_s, \tag{A.41}$$

Whenever $\psi(x) \geq 0$, the inequality on the right hand side of (A.37) immediately holds:

$$\frac{R'(x)}{R(x)} \geq -\frac{\phi(x)}{N(x)D(x)} \quad \forall x \ \in \{x : \ \psi(x) \geq 0\}.$$

Note that $\{x : \ \psi(x) \geq 0\} \subset \mathcal{I}_s$. By plugging inequality (A.41) into (A.37), we immediately have

$$\frac{R'(x)}{R(x)} > -\frac{\phi'(x)}{2N'(x)D(x)} \quad \forall x \ \in \{x : \ \psi(x) \geq 0\}. \tag{A.42}$$

This inequality together with (A.40) implies $\psi'(x) > 0$. This concludes the proof. ■

Lemma 9. *Suppose that $x \in \mathcal{I}_s$.*

$$\Upsilon(x) = \phi'(x)N(x) - \phi(x)2N'(x) > 0. \tag{A.43}$$

Proof. Let us take the left derivative of $\Upsilon(x)$ with respect to x:

$$\begin{aligned}
\frac{d}{dx}\Upsilon(x) &= \phi'(x)N'(x) + \phi''(x)N(x) - 2\phi'(x)N'(x) - 2\phi(x)N''(x) \\
&= \phi''(x)N(x) - \phi'(x)N'(x) - 2\phi(x)N''(x) \\
&= \Big(N''(x)D'(x)N(x) - D''(x)N'(x)N(x)\Big) \\
&\qquad - \Big(N''(x)D(x)N'(x) - D''(x)N(x)N'(x)\Big) - 2\phi(x)N''(x) \\
&= N''(x)\Big(D'(x)N(x) - N'(x)D(x)\Big) - 2\phi(x)N''(x) \\
&= -3\phi(x)N''(x) \\
&\leq 0
\end{aligned} \tag{A.44}$$

The inequality holds because (i) $\phi(x) \geq 0$ by construction and $\phi(x)$; (ii) $N''(x) > 0$ (see eq. (A.22, A.31)). Since $\Upsilon(x)$ is a nonincreasing function of x in $\mathcal{I}_s$, we have

$$\Upsilon(x) \geq \Upsilon(1) \quad \forall x \in \mathcal{I}_s.$$

To calculate the value of $\Upsilon(1)$, we calculate the value of following terms:

$$\begin{aligned}
N_a(1) &= \beta\mu - \lambda\beta + \lambda(\beta - 1) \\
N_a'(1) &= \kappa\left(-\lambda\beta + \lambda(\beta - 1)\right) + \lambda(\beta - 1)\left(\alpha + (1-\alpha)\kappa\right), \\
N_a''(1) &= 2\lambda(\beta - 1)(1 - \alpha + \alpha\kappa)\kappa, \\
N_b(1) &= \beta\mu - \lambda, \\
N_b'(1) &= 1\lambda(\beta - 1)(1 - \alpha + \alpha\kappa) - \lambda\beta + \lambda(\beta - 1) \\
N_b''(1) &= 2\lambda(\beta - 1)(1 - \alpha + \alpha\kappa), \\
D(1) &= \big(\beta\mu - \lambda\big)\big(\beta\mu - \lambda\beta + \lambda(\beta - 1)\big), \\
D'(1) &= \big(-\lambda(1 - \alpha + \alpha\kappa)\big)\big(\beta\mu - \lambda\beta + \lambda(\beta - 1)\big) + \big(\beta\mu - \lambda\big)\big(\lambda(\beta - 1)(1 - \alpha + \alpha\kappa)\big), \\
D''(1) &= -2\lambda^2(1 - \alpha + \alpha\kappa)^2(\beta - 1).
\end{aligned} \tag{A.45}$$

To show that $\Upsilon(1) = \phi'(1)N(1) - 2\phi(1)N'(1)$ is bounded below by zero, we first lower bound

the value of $\phi'(1)N(1)$ by using the function values at $x = 1$ given by (A.45)::

$$\begin{aligned}
\phi'(1)N(1) &= \big(N''(1)D(1) - D''(1)N(1)\big)N(1) \\
&= \Big(2\lambda(1-\alpha+\alpha\kappa)\,(\beta-1)\,(\kappa\alpha\gamma+1-\alpha)\,(\beta\mu-\lambda)^2 + 2\lambda^2(1-\alpha+\alpha\kappa)^2 \\
&\qquad \times (\beta-1)\,(\beta\mu-\lambda)\,(\alpha\gamma+1-\alpha)\ \Big)\Big(\ (\beta\mu-\lambda)\,(\alpha\gamma+(1-\alpha))\ \Big) \\
&= \Big(2\lambda(1-\alpha+\alpha\kappa)\,(\beta-1)\,(\beta\mu-\lambda)^2\,(\alpha\gamma+1-\alpha)\Big) \\
&\qquad \times \Big((\kappa\alpha\gamma+1-\alpha)\,(\beta\mu-\lambda) + \lambda(1-\alpha+\alpha\kappa)\,(\alpha\gamma+1-\alpha)\ \Big) \qquad \text{(A.46)} \\
&\geq \Big(2\lambda(1-\alpha+\alpha\kappa)\,(\beta-1)\,(\beta\mu-\lambda)^2\,(\alpha\gamma+1-\alpha)\Big) \\
&\qquad \times \Big((\kappa\alpha\gamma+1-\alpha)\,(\beta\lambda-\lambda) + \lambda(1-\alpha+\alpha\kappa)\,(\alpha\gamma+1-\alpha)\ \Big) \\
&= \Big(2\lambda^2(1-\alpha+\alpha\kappa)\,(\beta-1)\,(\beta\mu-\lambda)^2\,(\alpha\gamma+1-\alpha)\Big) \\
&\qquad \times \Big((\kappa\alpha\gamma+1-\alpha)\,(\beta-1) + (1-\alpha+\alpha\kappa)\,(\alpha\gamma+1-\alpha)\Big).
\end{aligned}$$

The inequality is by $\mu > \lambda$. Next, we calculate the value of $\phi(1)N'(1)$ by using the function values at $x = 1$ given by (A.45):

$$\begin{aligned}
\phi(1)N'(1) &= \big(N'(1)D(1) - D'(1)N(1)\big)N'(1) \\
&= \Big(\big((1-\alpha+\alpha\kappa)\lambda(\beta-1)(\alpha\gamma+1-\alpha) - \lambda(\kappa\alpha\gamma+1-\alpha)\big)\big(\beta\mu-\lambda\big)^2 \\
&\qquad - \lambda(1-\alpha+\alpha\kappa)\,(\beta\mu-\lambda)^2\,(\beta-2)\,(\alpha\gamma+1-\alpha)\Big) \\
&\qquad \times \Big((1-\alpha+\alpha\kappa)\lambda(\beta-1)(\alpha\gamma+1-\alpha) - \lambda(\kappa\alpha\gamma+1-\alpha)\Big) \qquad \text{(A.47)} \\
&= \lambda^2\,(\beta\mu-\lambda)^2\,\Big((1-\alpha+\alpha\kappa)(\alpha\gamma+1-\alpha) - (\kappa\alpha\gamma+1-\alpha)\Big) \\
&\qquad \times \Big((1-\alpha+\alpha\kappa)(\beta-1)(\alpha\gamma+1-\alpha) - (\kappa\alpha\gamma+1-\alpha)\Big).
\end{aligned}$$

For notational convenience, let $q_1 = 1-\alpha+\alpha\kappa$, $q_2 = 1-\alpha+\alpha\gamma$, and $q_3 = 1-\alpha+\alpha\kappa\gamma$.

Using (A.46) and (A.47), we have

$$\begin{aligned}
\Upsilon(1) &= \phi'(1)N(1) - 2\phi(1)N'(1) \\
&\geq 2\lambda^2 (\beta\mu - \lambda)^2 \Big(q_1 q_2 q_3 (\beta - 1)^2 + (\beta - 1) q_1^2 q_2^2 - (q_1 q_2 - q_3)\left(q_1 q_2 (\beta - 1) - q_3\right) \Big) \\
&= 2\lambda^2 (\beta\mu - \lambda)^2 q_3 \Big(q_1 q_2 (\beta - 1)^2 + q_1 q_2 \beta - q_3 \Big) \\
&\geq 2\lambda^2 (\beta\mu - \lambda)^2 q_3 (q_1 q_2 - q_3) \\
&= 2\lambda^2 (\beta\mu - \lambda)^2 q_3 (1 - \alpha)\alpha(1 - \kappa)(\gamma - 1) \\
&> 0.
\end{aligned} \tag{A.48}$$

The first equality is by definition of $\Upsilon(1)$. The first inequality is by inequality (A.44) and equality (A.48). The second equality is by rearranging the terms of the right hand side. The second inequality trivially holds because q_1, q_2, q_3 are nonnegative functions. The third equality by plugging $q1, q1$ into the right hand side and rearranging the terms. The final inequality holds because $\gamma > 1$, $q_3 > 0$, $\alpha \in (0, 1)$ and $\kappa < 1$ ($\phi(x) \geq 0$ only if $\kappa < 1$ by Corollary 4). ■

Proof of Theorem 1. By Proposition 8, we know that $\psi'(x) > 0$ whenever $\psi(x) \geq 0$ on a given interval. Next, we will show that the right value of $\psi(x)$ should be no less than the left value of the left value of the function at breakpoint $(x, g(x)) = (r, w)$. Since κ value associated the right function is less than the κ value associated the left function, we only need to show $\frac{d}{d\kappa}\psi(r) \leq 0$.

$$\begin{aligned}
\frac{d}{d\kappa}\psi(r) &= \frac{d}{d\kappa}\left(R'(r)D(r)N(r)\right) + \frac{d}{d\kappa}\left(R(r)\phi(r)\right) \\
&= D(r)N(r)\frac{d}{d\kappa}R'(r) + R(r)\frac{d}{d\kappa}\phi(r) \\
&= D(r)N(r)\left(e_1 - e_2\right)\alpha + R(r)\frac{d}{d\kappa}\phi(r) \\
&\leq R(r)\frac{d}{d\kappa}\phi(r) \\
&< 0.
\end{aligned}$$

The second equation holds because that $N(r)$, $D(r)$ and $R(r)$ are constant at breakpoint $(x, g(x)) = (r, w)$, thus their derivative with respect to κ is zero. The first inequality is by

$e_1 \leq e_2$ and the second equality is by Lemma 6. We can finally conclude that $\psi(x)$ increases once the objective function starts the increase. This is enough to show that there is a unique minimizer x^*. ■

A.1.4 Proofs for the Expected Value of Perfect Information

Proof of Proposition 6. We have already showed that value of perfect information is zero if $e_1\beta\mu \leq e_2(\beta\mu - \lambda)$ since it is optimal to misclassify all class-b customers for any feasible problem parameters. For the remaining of the proof, suppose that $e_1\beta\mu > e_2(\beta\mu - \lambda)$. Suppose that for a given set of parameters and α, the optimal solution to problem $\mathcal{P}_p$ is $z_p^*(\alpha)$.

Case I: Suppose that $z_p^*(\alpha) \geq \alpha + \frac{1-\alpha}{\kappa}$. This is feasible to problem $\mathcal{P}_i$, and thus value of perfect information is zero. Then, the claim trivially holds.

Case I: Suppose that $z_p^*(\alpha) < \alpha + \frac{1-\alpha}{\kappa}$. Assume the same set of parameters and $\alpha' > \alpha$, and let $z_p^*(\alpha')$ be the optimal solution to the new perfect information problem. By Proposition 1, we know that

$$z_p^*(\alpha') \leq z_p^*(\alpha) \quad \forall \alpha' \geq \alpha,$$

and, thus we have

$$z_p^*(\alpha') \leq z_p^*(\alpha) < \alpha + \frac{1-\alpha}{\kappa} \leq \alpha' + \frac{1-\alpha'}{\kappa}. \tag{A.49}$$

The last inequality is by $\alpha' \geq \alpha$ and $\kappa > 1$. This implies that $z_p^*(\alpha')$ is not a feasible solution to imperfect information problem with α'. Let us compare the value of perfect information for settings with α and α'.

$$\begin{aligned}
W\left(z_i^*(\alpha')\right) - W\left(z_p^*(\alpha')\right) &= \int_{z_p^*(\alpha')}^{\alpha' + \frac{1-\alpha'}{\kappa}} dW(z, \alpha') \\
&\geq \int_{z_p^*(\alpha)}^{\alpha + \frac{1-\alpha}{\kappa}} dW(z, \alpha') \\
&\geq \int_{z_p^*(\alpha)}^{\alpha + \frac{1-\alpha}{\kappa}} dW(z, \alpha) \\
&= W\left(z_i^*(\alpha)\right) - W\left(z_p^*(\alpha)\right).
\end{aligned}$$

The first equation is given by (2.16). The first inequality holds because $dW(z,\alpha) \geq 0$ for all $z \geq z_p^*(\alpha')$ by convexity (see Proposition 2) and the inequality chain (A.49) holds. The second inequality is by $dW(z,\alpha)$ being an increasing function of α (see Lemma 1). The final equation is by (2.16). The details of the proof for γ is similar, so we omit the details. ■

A.1.5 Proofs for Classification when the Service Rate is Type-dependent

Proof of Proposition 4. Suppose perfect information. Recall that there is only incentive to misclassify type-A customers as nonpriority in the setting with class-dependent service rates because if nonpriority service time is relatively less variable and not too much more variable than the priority service time, then the expected residual time can decrease and result in decrease in type-A customers waiting time. However, if the service rates are type-dependent then, expected residual time is the same regardless of any classification outcome as shown in Argon and Ziya [2009]. Therefore, we conclude that there would be no misclassification incentive to misclassify type-A customers as nonpriority in their setting. Next we will show that there is also no misclassification incentive to misclassify type-B customers as priority. First, let us introduce some notation:

- $(1-\eta)$ fraction of true type-A customers who classified as priority.
- $(\eta - s)$ fraction of true type-B customers who are classified as priority.
- s fraction of true type-B customers who are classified as nonpriority.

Suppose perfect information. Then, the mean waiting time of customers who are identified as priority (class-1) and nonpriority (class-2) are respectively given by $W_1(s)$ and $W_2(s)$:

$$\begin{aligned} W_1(s) &= \frac{\lambda\left((1-\eta)e_1 + \eta e_2\right)}{2\left(1-\lambda(1-\eta)a_A - \lambda(\eta - s)a_B\right)}, \\ W_2(s) &= \frac{W_1(s)}{1-\rho}, \end{aligned} \tag{A.50}$$

where

$$\rho = \lambda\left(\eta a_B + (1-\eta)a_A\right).$$

The cost function $C(s)$ is given by

$$\begin{aligned} C(s) &= \lambda\left((1-\eta)h_A W_1(s) + (\eta - s)h_B W_1(s) + s h_B W_2(s)\right) \\ &= \lambda W_1(s)\left((1-\eta)h_A + (\eta - s)h_B + \frac{s h_B}{1-\rho}\right) \\ &= \frac{\left((1-\eta)h_A + (\eta - s)h_B + \frac{s h_B}{1-\rho}\right)}{(1 - \lambda(1-\eta)a_A - \lambda(\eta - t)a_B)} \\ &= \frac{\lambda^2\left((1-\eta)e_1 + \eta e_2\right)}{2}\frac{\left((1-\eta)h_A + (\eta - s)h_B + \frac{s h_B}{1-\rho}\right)}{(1 - \lambda(1-\eta)a_A - \lambda(\eta - t)a_B)}. \end{aligned} \tag{A.51}$$

Define $\kappa,\ \kappa_A,\ \kappa_B,\ c_1,\ c_2$ as

$$\begin{aligned} \kappa &= \frac{\lambda^2\left((1-\eta)e_A + \eta e_B\right)}{2}, \\ \kappa_A &= (1-\eta)h_A + \eta h_B, \\ \kappa_B &= 1 - \lambda(1-\eta)a_A - \lambda\eta a_B = 1 - \rho, \\ c_1 &= h_B\left(\frac{1}{1-\rho} - 1\right) = h_B\left(\frac{\rho}{1-\rho}\right), \\ c_2 &= \lambda a_B \end{aligned} \tag{A.52}$$

Now, we can rewrite the cost function $C(s)$ as

$$C(s) = \kappa\frac{\kappa_A + c_1 s}{\kappa_B + c_2 s}. \tag{A.53}$$

The first and second derivative of the function are Define $\kappa,\ \kappa_A,\ \kappa_B,\ c_1,\ c_2$ as

$$\begin{aligned} \frac{d}{ds}C(s) &= \frac{\kappa_B c_1 - \kappa_A c_2}{(\kappa_B + c_2 s)^2}, \\ \frac{d^2}{ds^2}C(s) &= 2c_2\frac{\kappa_A c_2 - \kappa_B c_1}{(\kappa_B + c_2 s)^3}. \end{aligned} \tag{A.54}$$

Next, we figure out the sign of the nominator

$$\begin{aligned} \kappa_A c_2 - \kappa_B c_1 &= (1-\eta)\lambda a_B h_A + \eta\lambda a_B h_B - \left(\frac{\rho}{1-\rho}\right)(1-\rho)h_B \\ &= (1-\eta)\lambda a_B h_A + \eta\lambda a_B h_B - \rho h_B \end{aligned} \tag{A.55}$$

By assumption $h_A/a_A > h_B/a_B$, i.e. $h_A a_B > h_B a_A$, we can rewrite the above expression as

$$\begin{aligned} \kappa_A c_2 - \kappa_B c_1 &> (1-\eta)\lambda a_A h_B + \eta\lambda a_B h_B - \rho h_B \\ &= h_B\left(\rho - \rho\right) \\ &= 0. \end{aligned} \tag{A.56}$$

We conclude that for all $s \in [0, \eta]$, we have

$$\begin{aligned} &\frac{d}{ds}C(s) < 0, \\ &\frac{d^2}{ds^2}C(s) > 0. \end{aligned} \tag{A.57}$$

This implies that $C(s)$ is a convex decreasing function of s, and thus $s^* = \eta$. We conclude that there is no incentive to misclassify type 2 customers. ■

A.2 Chapter 4: Chief Complaint Categorization

Chief complaints (CCs) are brief statements that explain why the patient initiates an ED encounter. A patient can have a single or multiple CCs. For example, the patient can say "I have abdominal pain", "I have nausea" or "I have abdominal pain and nausea". We consider two approaches to model multiple chief complaints. The first approach is adding the main and interaction effect of each chief complaint into the model. For example, a patient with abdominal pain and nausea takes value of 1 for abdominal pain, nausea and abdominal pain + nausea. We observe 598 distinct CCs in our dataset after data cleaning. Even if we want to add the simplest type of interaction – two-way interaction –, we need to add $(598-1)^2 \approx 350,000$ terms into the model. This approach causes overfitting because our dataset includes $40,000$ choice incidents. Even if we want to utilize dimension reduction techniques to prevent overfitting, we can come across run-time and memory issues given the immense nature of the feature space. Therefore, we suggest the following clustering approach for CC modeling.

1. For each patient encounter, combine all CCs into a single statement. For example, if the patient has abdominal pain and nausea, their CC is neither abdominal pain nor

nausea. It is abdominal pain + nausea, which is a new category.

2. Group modified CCs into two groups: $\mathbb{C}_s$ — the set of CCs associated with 400 or more encounters (the most common 104 CCs) — and $\mathbb{C}_s$ — the set of CCs associated with less than 104 encounters.

3. We replace less common CCs with more common CCs based on the natural language and clinical information, in particular, associated ESI scores with CCs.

Let us introduce some notation before presenting the CC clustering algorithm. We define $\text{Cover}(c_i, c_j)$ as the ratio of the number of the shared items in c_i and c_j to the number of items in c_j. Suppose that c_i = abdominal pain + nausea + vomiting + fatigue and c_j = abdominal pain. Then $\text{Cover}(c_i, c_j) = 1$ and $\text{Cover}(c_j, c_i) = 0.25$. We let ESI_c as the average ESI score of patients with chief complaint c.

We should not assign a patient with abdominal pain + nausea + vomiting to a chief complaint that includes an item that the patient does not present such as chest pain or back pain. For each $c_s \in \mathbb{C}_s$, we let $M(c_s)$ denote the set of $c_h \in \mathbb{C}_h$ such that all items in c_h is covered by c_s, i.e, $\text{Cover}(c_s, c_h) = 1$. Assume a less common c_s = abdominal pain + nausea + vomiting + fatigue and a more common CC c_o = abdominal pain + chest pain. Observe that $\text{Cover}(c_s, c_0) = 0.5$; thus c_s cannot be assigned to common c_o.

If the set $M(c_s)$ contains multiple CCs, we select c_h that looks like c_s the most. Assume, for instance, that $M(c_s) = \{c_1, c_2, c_3\}$ where c_1 = abdominal pain + nausea + vomiting, c_2 = abdominal pain+nausea, and c_3 = abdominal pain. Among these three CCs, c_1 has the highest number of shared items with c_s, i.e., highest value of $\text{Cover}(c_i, c_s)$. $\text{Cover}(c_1, c_s) = 0.75 > \text{Cover}(c_2, c_s) = 0.50 > \text{Cover}(c_3, c_s) = 0.25$. If there is a tie between multiple items in $M(c_s)$ in terms of the number of shared items, we select the common CC associated with the highest urgency. The Algorithm 4 presents the formalization of our CC clustering approach.

Algorithm 4 CC-REASSIGNMENT

Input: $\mathbb{C}_s$, $\mathbb{C}_h$, $\left\{\text{ESI}_{c_h}\right\}_{c_h \in \mathbb{C}_s}$
for all $c_s \in \mathbb{C}_s$ **do**
    $M = \langle\rangle$
    for all $c_h \in \mathbb{C}_h$ **do**
        $r_h, r_s = \text{Cover}(c_s, c_h), \text{Cover}(c_h, c_s)$
        if $r_h = 1$, i.e., all components of c_h is covered by c_s **then**
            $M \leftarrow M + \langle c_h, r_s, \text{ESI}_{c_h}\rangle$
        end if
    end for
    if $M = \emptyset$ **then**
        $c_s \leftarrow$ other
    else
        Sort M in descending order of r_s and then in ascending order of ESI_{c_h}
        Select c^* as the c_h in the top row of M
        $c_s \leftarrow c^*$
    end if
end for

REFERENCES

ACEP. Policy Statements crowding, 2019. URL https://www.acep.org/patient-care/policy-statements/crowding/.

Saed Alizamir, Francis De Véricourt, and Peng Sun. Diagnostic accuracy under congestion. *Management Science*, 59(1):157–171, 2013.

Ali Aouad, Adam N Elmachtoub, Kris J Ferreira, and Ryan McNellis. Market segmentation trees. *arXiv preprint arXiv:1906.01174*, 2019.

Theo Arentze and Harry Timmermans. Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice. *Transportation Research Part B: Methodological*, 41(7):772–783, 2007.

Nilay Tanık Argon and Serhan Ziya. Priority assignment under imperfect information on customer type identities. *Manufacturing & Service Operations Management*, 11(4):674–693, 2009.

American Heart Association. High blood pressure, a. URL https://www.heart.org/en/health-topics/high-blood-pressure.

Emergency Nurses Association. Emergency severity index (esi): A triage tool for emergency department care version 4, b.

Iain Beardsell and Sarah Robinson. Can emergency department nurses performing triage predict the need for admission? *Emergency Medicine Journal*, 28(11):959–962, 2011.

Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.

Timothy Brathwaite and Joan L Walker. Asymmetric, closed-form, finite-parameter models of multinomial choice. *Journal of choice modelling*, 29:78–112, 2018.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

Allan Cameron, Kenneth Rodgers, Alastair Ireland, Ravi Jamdar, and Gerard A McKay. A simple tool to predict admission at the time of triage. *Emergency Medicine Journal*, 32(3):174–179, 2015.

Mayo Clinic. Patient Care and Health Information fever, 2020. URL https://www.mayoclinic.org/diseases-conditions/fever/symptoms-causes/syc-20352759.

Alan Cobham. Priority assignment in waiting line problems. *Journal of the Operations Research Society of America*, 2(1):70–76, 1954.

Yichuan Ding, Eric Park, Mahesh Nagarajan, and Eric 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.

Gregory Dobson and Arvind Sainathan. On the impact of analyzing customer information and prioritizing in a service system. *Decision Support Systems*, 51(4):875–883, 2011.

Andrea Freyer Dugas, Thomas D Kirsch, Matthew Toerper, Fred Korley, Gayane Yenokyan, Daniel France, David Hager, and Scott Levin. An electronic emergency triage system to improve patient distribution by critical outcomes. *The Journal of emergency medicine*, 50 (6):910–918, 2016.

L Durr. A single-server priority queuing system with general holding times, poisson input, and reverse-order-of-arrival queuing discipline. *Operations Research*, 17(2):351–358, 1969.

Marta Fernandes, Susana M Vieira, Francisca Leite, Carlos Palos, Stan Finkelstein, and João MC Sousa. Clinical decision support systems for triage in the emergency department using intelligent systems: a review. *Artificial Intelligence in Medicine*, 102:101762, 2020.

Nicki Gilboy, Paula Tanabe, Debbie Travers, and Alexander M. Rosenau. *Emergency Severity Index Version 4: Implementation Handbook*. Emergency Nurses Association (ENA), 2020.

Farrokh Habibzadeh, Parham Habibzadeh, and Mahboobeh Yadollahie. On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochemia medica: Biochemia medica*, 26(3):297–307, 2016.

Woo Suk Hong, Adrian Daniel Haimovich, and R Andrew Taylor. Predicting hospital admission at emergency department triage using machine learning. *PloS one*, 13(7):e0201016, 2018.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370.

Jin-Hyung Kim and Mijung Kim. Two-stage multinomial logit model. *Expert Systems with Applications*, 38(6):6439–6446, 2011.

Mijung Kim. Two-stage logistic regression model. *Expert Systems with Applications*, 36(3): 6727–6734, 2009.

Niels Landwehr, Mark Hall, and Eibe Frank. Logistic model trees. *Machine learning*, 59 (1-2):161–205, 2005.

Seung-Yup Lee, Ratna Babu Chinnam, Evrim Dalkiran, Seth Krupp, and Michael Nauss. Prediction of emergency department patient disposition decision for proactive resource allocation for admission. *Health care management science*, 23(3):339–359, 2020.

Scott Levin, Matthew Toerper, Eric Hamrock, Jeremiah S Hinson, Sean Barnes, Heather Gardner, Andrea Dugas, Bob Linton, Tom Kirsch, and Gabor Kelen. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Annals of emergency medicine*, 71 (5):565–574, 2018.

Alix Lhéritier, Michael Bocamazo, Thierry Delahaye, and Rodrigo Acuna-Agost. Airline itinerary choice modeling using machine learning. *Journal of choice modelling*, 31:198–209, 2019.

Dan-Ling Li, Jun-Xiang Peng, Chong-Yang Duan, and Ju-Min Deng. Weighted product index and its two-independent-sample comparison based on weighted sensitivity and specificity. *Communications in Statistics-Theory and Methods*, 47(22):5445–5459, 2018.

Wenhao Li, Zhankun Sun, and L Jeff Hong. Who is next: Patient prioritization under emergency department blocking. *Operations Research*, 2021.

Pablo Martínez-Camblor and Juan Carlos Pardo-Fernández. The youden index in the generalized receiver operating characteristic curve context. *The international journal of biostatistics*, 15(1), 2019.

Daniel McFadden et al. Conditional logit analysis of qualitative choice behavior. 1973.

Velibor V Mišić. *Data, models and decisions for large-scale stochastic optimization problems*. PhD thesis, Massachusetts Institute of Technology, 2016.

University of Rochester Medical Center. Vital signs (body temperature, pulse rate, respiration rate, blood pressure), 2020. URL `https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID=85&ContentID=P00866`.

Members of the Public Health and Injury Prevention Committee. Public health impact of ed crowding and boarding of inpatients. American College of Emergency Pyhcisians, 2009.

Chirag R. Parikh and Heather Thiessen Philbrook. Chapter 2 - statistical considerations in analysis and interpretation of biomarker studies. In Charles L. Edelstein, editor, *Biomarkers of Kidney Disease*, pages 25–37. Academic Press, San Diego, 2011. ISBN 978-0-12-375672-5. doi:https://doi.org/10.1016/B978-0-12-375672-5.10002-7. URL `https://www.sciencedirect.com/science/article/pii/B9780123756725100027`.

Clare Allison Parker, Nan Liu, Stella Xinzi Wu, Yuzeng Shen, Sean Shao Wei Lam, and Marcus Eng Hock Ong. Predicting hospital admission at the emergency department triage: A novel prediction model. *The American journal of emergency medicine*, 37(8):1498–1504, 2019.

John R Quinlan et al. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348. World Scientific, 1992.

Yoshihiko Raita, Tadahiro Goto, Mohammad Kamal Faridi, David FM Brown, Carlos A Camargo, and Kohei Hasegawa. Emergency department triage prediction of clinical outcomes using machine learning models. *Critical care*, 23(1):64, 2019.

John R Richards, M Christien van der Linden, and Robert W Derlet. Providing care in emergency department hallways: demands, dangers, and deaths. *Advances in Emergency Medicine*, 2014, 2014.

Gerta Rücker and Martin Schumacher. Summary roc curve based on a weighted youden index for selecting an optimal cutpoint in meta-analysis of diagnostic accuracy. *Statistics in medicine*, 29(30):3069–3078, 2010.

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

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

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

Yiwen Shen, Carri Chan, Fanyin Zheng, and Gabriel J Escobar. Structural estimation of intertemporal externalities on icu admission decisions. *Available at SSRN 3564776*, 2020.

Simrita Singh, Itai Gurvich, and Jan A Van Mieghem. Feature-based design of priority queues: Digital triage in healthcare. *Available at SSRN*, 2020.

Zhankun Sun, Nilay Tanık Argon, and Serhan Ziya. Patient triage and prioritization under austere conditions. *Management Science*, 64(10):4471–4489, 2018.

Ilker Unal. Defining an optimal cut-point value in roc analysis: an alternative approach. *Computational and mathematical methods in medicine*, 2017, 2017.

S van Cranenburgh, Shenhao Wang, Akshay Vij, F Pereira, and J Walker. Choice modelling in the age of machine learning [arxiv]. 2021.

SP Van der Zee and H Theil. Priority assignment in waiting-line problems under conditions of misclassification. *Operations Research*, 9(6):875–885, 1961.

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

Zlata K Vlodaver, Jeffrey P Anderson, Brittney E Brown, and Michael D Zwank. Emergency medicine physicians' ability to predict hospital admission at the time of triage. *The American journal of emergency medicine*, 37(3):478–481, 2019.

Shari J Welch. Patient segmentation: Redesigning flow. *Emergency Medicine News*, 31(8), 2009.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Achim Zeileis, Torsten Hothorn, and Kurt Hornik. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514, 2008.

Xingyu Zhang, Joyce Kim, Rachel E Patzer, Stephen R Pitts, Aaron Patzer, and Justin D Schrager. Prediction of emergency department hospital admission based on natural language processing and neural networks. *Methods Inf Med*, 56(5):377–89, 2017.