Data-driven Stochastic Optimization Approaches to Determine Decision Thresholds for Risk Estimation Models

Gian-Gabriel P. Garcia[1], Mariel S. Lavieri[1], Ruiwei Jiang[1],

Michael A. McCrea[2], Thomas W. McAllister[3], Steven P. Broglio[4], CARE Consortium Investigators

[1]Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI

[2]Departments of Neurosurgery and Neurology, Medical College of Wisconsin, Milwaukee, WI

[3]Department of Psychiatry, Indiana University School of Medicine, Indianapolis, IN

[4]School of Kinesiology, University of Michigan, Ann Arbor, MI

The Version of Record of this manuscript has been published and is available in *IISE Transactions* February 6, 2020 https://www.tandfonline.com/doi/abs/10.1080/24725854.2020.1725254

Abstract: The increasing availability of data has popularized risk estimation models in many industries, especially healthcare. However, properly utilizing these models for accurate diagnosis decisions remains challenging. Our research aims to determine when a risk estimation model provides sufficient evidence to make a positive or negative diagnosis, or if the model is inconclusive. We formulate the two-threshold problem (TTP) as a stochastic program which maximizes sensitivity and specificity while constraining false-positive and false-negative rates. We characterize the optimal solutions to TTP as either two-threshold or one-threshold and show that its optimal solution can be derived from a related linear program (TTP*). We also derive utility-based and multi-class classification frameworks for which our analytical results apply. We solve TTP* using data-driven methods: quantile estimation (TTP*-Q) and distributionally robust optimization (TTP*-DR). Through simulation, we characterize the feasibility, optimality, and computational burden of TTP*-Q and TTP*-DR and compare TTP*-Q to an optimized single threshold. Finally, we apply TTP* to concussion assessment data and find that it achieves greater accuracy at lower misclassification rates compared to traditional approaches. This data-driven framework can provide valuable decision support to clinicians by identifying "easy" cases which can be diagnosed immediately and "hard" cases which may require further evaluation before diagnosing.

Key Words: data-driven optimization, stochastic programming, distributionally robust optimization, quantile estimation, diagnosis decisions, risk estimation models

1 Introduction

The growing availability of data has led to the rapid development of risk estimation models in several industries including healthcare, finance, and manufacturing. These models can be especially impactful in healthcare, where risk estimation models can be used as decision aids to supplement medical diagnosis and treatment decisions. For example, methods such as logistic regression, survival analysis, neural networks, and other machine learning models have been used to assess emergency department admission risk (Peck et al., 2012), estimate hospital readmission risk (Xue et al., 2019), detect glaucoma progression (Schell et al., 2014), predict adverse coronary heart disease events (Anderson, 1991), and aid diagnosis of cancer (Kourou et al., 2015). Yet, applying these models in clinical practice is challenging (Moons et al., 2009; Degeling et al., 2017). One way to bridge this gap between research and practice is by determining decision thresholds to help clinicians interpret these models (Ebell, 2010).

In this research, we aim to determine suitable decision thresholds for risk estimation models. For instance, if a model estimates that a patient has a 55% chance of having a cancerous tumor, does that estimate provide sufficient evidence to diagnose the patient with cancer? What if the model estimates a 45% chance? When the consequences associated with misdiagnoses are great, how one determines these decision boundaries is critical.

Determining appropriate decision thresholds is not straightforward. First, these decision boundaries should reflect the decision-maker's risk attitude, i.e., willingness to take on consequences associated with misdiagnoses (Felder and Mayrhofer, 2014). While some patients and physicians may be more willing to accept highly consequential outcomes, others may be far more conservative in their decision-making (based on the perceived consequences associated with each decision). Yet, traditional methods for interpreting risk estimates do not reflect such risk attitudes (Degeling et al., 2017). Even worse, some may provide arbitrary risk stratifications.

Another key issue is the application of these risk estimation models to populations for which the underlying population does not match the population used to parameterize the model (Altman et al., 2009). For example, consider risk scores from the Framingham Heart Study (Wang et al., 2003), which was parameterized on a cohort of roughly 800 participants from Framingham, Mas-

sachussettes between 1948 and 2000. This model has been used to inform clinical practice for cardiovascular disease management, which is applied to populations which are far more diverse (Perk et al., 2012). Therefore, decision boundaries should not only depend on the risk estimation model at hand but also on the population to which it will be applied.

Furthermore, traditional binary classification models may not sufficiently address uncertainty in diagnosis decisions. This sentiment is reflected in diagnosis guidelines for conditions such as multiple sclerosis (McDonald et al., 2001), Alzheimer's disease (AD) (McKhann et al., 2011), diabetes (American Diabetes Society, 2016), and concussion (Kutcher and Giza, 2014), where diagnosis is divided into risk classifications rather than a dichotomous outcome. For instance, for sports-related concussion, the diagnosis may be broken up into (1) Possible, (2) Probable, and (3) Definite concussion depending on the clinician's diagnostic certainty (Kutcher and Giza, 2014). In guidelines which use such risk classifications, intermediate risk classifications arise from conflicting diagnostic assessment results or a lack of definitive evidence to make strong diagnostic conclusions. Analogously, there may be ranges along the risk spectrum where a risk estimation model's estimates are not "certain enough" and call for more information before a diagnosis decision can be made. In particular, these intermediate ranges reflect cases in which qualitative information, which may not be easily implemented in risk estimation models, should be used to guide diagnosis decisions. Since risk estimation models are typically designed to supplement clinical diagnosis, identifying these ranges is critical. Yet, few methods create decision boundaries which account for this risk estimation uncertainty (Degeling et al., 2017).

This research aims to bridge the gap between risk estimation models and clinical application by presenting a rigorous approach to determine diagnosis decision thresholds. These thresholds (1) reflect the decision-maker's risk attitude, (2) jointly depend on the risk estimation model and patient population to which it is applied, and (3) identify ranges in the risk continuum in which the risk estimation model is most and least accurate. We apply our method to acute concussion assessment, a field where diagnostic decisions must be made accurately and quickly to mitigate prolonged injury recovery and post-concussion symptom severity (Asken et al., 2018).

The key contributions of this work are as follows:

1. We introduce a data-driven stochastic optimization framework to determine diagnostic decision thresholds based on the application of a risk estimation model to patients from a fixed population. Compared to previous methods (see Section 2), we avoid the need to estimate outcome-based utilities or make distributional assumptions to account for uncertainty in risk estimates.

2. In our analytical study, we show that the optimal solution to our proposed model can be characterized by extreme-point solutions of a related linear program. Thus, our model can be solved using quantile estimation — bypassing the need for advanced optimization software. We also identify additional modeling frameworks, including utility-based and multi-class classification frameworks, for which our analytical results can be applied. Specifically, for our utility-based extensions, we formulate a model for which utilities such as quality-adjusted life-years may be used. In our extensions to multi-class classification, we develop frameworks for both multi-label and ordinal classification.

3. Through an analytical study and numerical analysis using both real and simulated data, we determine when two decision thresholds, which allow for a deferred diagnosis decision, will outperform a single decision threshold, which only allows for binary classification.

4. We perform extensive numerical analysis to determine how the modeling parameters should be chosen based on the general characteristics of the population that undergoes diagnostic testing. Our analysis also gives insight to guide the choice of data-driven solution methodology based on sample size and the quality of the underlying risk estimation model.

5. We are one of the first groups to apply an optimization framework to develop data-driven diagnostic thresholds for acute concussion based on data from the CARE Consortium — a nationwide collaboration comprising 29 National Collegiate Athletic Association (NCAA) universities and military service academies. By incorporating feedback from concussion experts across the CARE Consortium, we ensure that, in the case of acute concussion assessment, our modeling framework outperforms methods which are commonly used in practice. Furthermore, we provide a valuable framework which quantifies the uncertainty in diagnosis decisions using real data rather than subjective clinical experience. The models developed

in this research have the potential to be developed into tools which can supplement clinical decision-making.

The remainder of this paper is organized as follows. In Section 2, we present a review of related research literature. In Section 3, we present our modeling approach and the analytical properties of this model, along with model extensions to utility-based and multi-class classification frameworks. In Section 4, we present our data-driven solution methods and in Section 5, we evaluate each of these solution methods using simulation. In Section 6, we apply our models to concussion assessment data. We perform sensitivity analysis on the model parameters and analyze the performance of this model compared to existing methods. Finally, we present managerial insights and other concluding remarks in Section 7. The proofs for all analytical results in this manuscript may be accessed in the online supplementary material.

2 Relevant Literature

This research falls within the domains of (1) operations research in disease screening and diagnosis decisions and (2) determining diagnosis decision boundaries.

2.1 Operations Research in Disease Screening and Diagnosis Decisions

Operations Research has been applied to many areas of disease screening. Specifically, applications to cancer screening are reviewed extensively by Pierskalla and Brailer (1994) and Alagoz et al. (2011) while more recent works include Maillart et al. (2008); McLay et al. (2010); Ayer et al. (2012); Erenay et al. (2014); Li et al. (2014); Güneş et al. (2015); Lee et al. (2015); Tejada et al. (2015); Ayer et al. (2016); Bertsimas et al. (2016) and Barnett et al. (2017). Other applications include obesity (Yang et al., 2013), Glaucoma (Helm et al., 2015), Chlamydia (Teng et al., 2015), Ebola (Jacobson et al., 2016), blood screening (El-Amine et al., 2018), and HIV (Deo et al., 2015; Deo and Sohoni, 2015; Jónasson et al., 2017). Like our study, these works consider the imperfect nature of diagnostic tests in their models. However, they determine optimal strategies based on estimated utilities while we focus on diagnostic accuracy. Further, they focus on sequential decisions while we consider the immediate diagnosis decision.

Operations Research has also been applied to medical diagnosis decisions, where such problems typically optimize pre-diagnosis decisions or follow-up decisions after an initial diagnostic test.

For example, Bayati et al. (2018) determine the least-cost set of biomarker tests which allow for sufficient diagnostic power while Ayvaci et al. (2012) and Zhang et al. (2012) optimize biopsy follow-up decisions for cancer. However, our work focuses on the actual diagnosis decision at hand instead of pre- or post-diagnosis decisions. To this end, Ayvaci et al. (2017) and Ahsen et al. (2019) study when and how bias-inducing information should be incorporated in breast cancer diagnostic decisions. Similarly to Ahsen et al. (2019), we also study the incorporation of clinical decision support systems in diagnostic decisions. However, they focus on the design of such systems while we focus on the interpretation of these decision support systems, i.e., a risk estimation model. Furthermore, their work focuses on balancing two sources of information (i.e., mammogram risk and clinical-risk information) and deriving one diagnostic threshold whereas our work focuses on a single source of information (i.e., risk estimates) and deriving two diagnostic decision thresholds.

2.2 Determining Diagnosis Decision Boundaries

A number of methods have been developed to optimize a single decision threshold. In this literature, it is typical to assign a utility to each possible diagnostic outcome and determine an optimal threshold which maximizes utilities (Pauker and Kassirer, 1975; Deneef and Kent, 1993; Moons et al., 1997; Jund et al., 2005; Felder and Mayrhofer, 2014; van Giessen et al., 2018). For instance, van Giessen et al. (2018) develop a stepwise method to optimize a risk threshold based on multiple criteria, including quality-adjusted life-years, cost of treatment, and net health benefit. However, the use of such utilities has been questioned (McGregor and Caro, 2006; Nord et al., 2009). To circumvent this difficulty, methods based on the receiver operating characteristic (ROC) have been developed (Vermont et al., 1991; Somoza and Mossman, 1992; Greiner et al., 1995, 2000; Odetola et al., 2016). These methods typically optimize a single threshold based on measures of diagnostic accuracy. However, single thresholds do not specify regions for which the risk estimation models perform poorly, i.e., risk estimate ranges over which false-positive and false-negative rates are especially high and diagnostic decisions should be avoided. Therefore, we also review methods to determine multiple decision thresholds.

Utility-based methods for deriving multiple thresholds include Pauker and Kassirer (1980); Glasziou and Hilden (1986) and Nease et al. (1989). In contrast, Hartz et al. (1986) determine thresholds based on uncertainty in different physicians' decision thresholds and Mangasarian et al. (1995) apply

linear programming to determine diagnostic decision thresholds based on tumor characteristics. Zhu and Fang (2016) and Si et al. (2017) develop tree-based approaches which categorize diagnoses as positive, negative, or uncertain, where those who are uncertain are not predicted well by their classification tree and hence should be further evaluated. We apply this same classification scheme in our work but we develop thresholds along the probability spectrum. Our work most closely relates to Si et al. (2017) since we both employ optimization frameworks to determine diagnosis thresholds based on ROC statistics. However, they determine decision boundaries for each biomarker in a sequence of biomarkers, rather than a one-time threshold in probability space which can potentially incorporate multiple biomarkers at once. Furthermore, they assume that their biomarker readings follow a Gaussian distribution while we do not make any assumptions on the distribution of risk scores. Additionally, in their treatment of non-Gaussian biomarkers, they solve the optimization problem using an iterative approximation. In contrast, we solve tractable data-driven optimization problems to global optimality.

Several methods based on Bayesian decision theory have also been used to derive decision thresholds (Sheppard and Kaufman, 2005; Weise et al., 2006; Yao, 2010; Yao and Zhou, 2016). Among these approaches, both Sheppard and Kaufman (2005) and Weise et al. (2006) derive thresholds based on distributional information. Specifically, Sheppard and Kaufman (2005) compare posterior likelihood ratios to determine whether a risk estimate is more likely to come from a true-positive or a true-negative patient. However, this approach requires estimation of the full distribution of risk estimates to compute, whereas we find that our thresholds only depend on quantiles of these distributions. To this end, Weise et al. (2006) also derive thresholds based on quantiles. However, they assume that these quantiles are derived from a standard normal distribution whereas we do not make distributional assumptions. Finally, the decision thresholds derived by Yao (2010) and Yao and Zhou (2016) are most similar to our research since they explicitly aim to determine whether an object should be classified as either positive, negative, or on the boundary (i.e., too uncertain to make a decision). To this end, the thresholds derived in both Yao (2010) and Yao and Zhou (2016) are determined entirely by "risk" associated with each decision and do not rely on any distributional information at all. Moreover, all of the aforementioned approaches determine decision thresholds based on unconstrained optimization problems, whereas our approach requires solving a constrained

stochastic program.

Overall, our work differs from previous research in three ways: (1) we take a constrained optimization approach, allowing us to simultaneously maximize sensitivity and specificity while limiting false-positive and false-negative rates, (2) we limit false-positive and false-negative rates based on the decision-maker's risk attitude, creating personalized decision thresholds, and (3) we consider uncertainty without making distributional assumptions. In Section 3, we detail our stochastic programming approach to determining these decision thresholds.

3 Modeling Approach

In this section, we describe our modeling approach, its related analytical properties, and several modeling extensions. Specifically, we provide our general problem setting and notation in Section 3.1 and the model formulation in Section 3.2. In Section 3.3, we develop an approximation to our stochastic programming model and in Section 3.4, we characterize its optimal solution based on the extreme-point solutions of a related linear program. We then identify conditions under which the optimal solution for the approximation is also optimal for the original stochastic programming model. In Section 3.5 and Section 3.6, we formulate and analyze utility-based and multi-class diagnosis extensions, respectively, to our modeling framework.

3.1 Problem Setting and Notation

A summary of our notation is provided in Table 1. We consider a patient population for a chosen disease, which can be divided into two mutually exclusive populations of true-positives (e.g., has a concussion) and true-negatives (e.g., does not have a concussion). A randomly chosen patient is associated with the random variables (X, Y). The random vector $X \in \mathcal{X}$ represents a vector of patient characteristics (e.g., age, sex, concussion assessment results) which have been transformed into a numerical representation (e.g., using one-hot encoding for categorical variables). We let $\mathcal{X} \subseteq \mathbb{R}^p$ denote the set of p-length numerical representations of patient characteristics. The random variable $Y \in \{0, 1\}$ represents the patient's label, which indicates whether he or she is from the true-positive (i.e., $Y = 1$) or true-negative (i.e., $Y = 0$) population. A risk estimation model $f : \mathcal{X} \to (0, 1)$ approximates the conditional probability $\mathbb{P}(Y = 1|X)$. Such models include logistic regression, classification and regression trees, and artificial neural networks.

Table 1: Model Notation

Notation	Description
$\xi^+ \sim \mathcal{P}^+$, $\xi^- \sim \mathcal{P}^-$	Random risk estimate belonging to true-positives and true-negatives, respectively, along with their corresponding distributions
t	General diagnostic threshold
u, l	Upper and lower diagnostic decision thresholds, respectively
$se(u, \xi^+)$	Event that a true-positive patient is correctly classified as positive given upper threshold u and risk estimate ξ^+ (sensitivity)
$sp(l, \xi^-)$	Event that a true-negative patient is correctly classified as negative given lower threshold l and risk estimate ξ^- (specificity)
$fp(u, \xi^-)$	Event that a true-negative patient is incorrectly classified as positive given upper threshold u and risk estimate ξ^- (false-positive)
$fn(l, \xi^+)$	Event that a true-positive patient is incorrectly classified as negative given lower threshold l and risk estimate ξ^+ (false-negative)
λ, ϕ	Weighting parameters to balance sensitivity and specificity in TTP and TTP*, respectively
γ^{fp}, γ^{fn}	Maximum levels of false-positive and false-negative rates, respectively

Throughout the remainder of the paper, we focus on the risk estimates $\xi^+ := f(X|Y = 1)$ and $\xi^- := f(X|Y = 0)$, which denote the risk estimates belonging to a patient from the population of true-positives and true-negatives, respectively. Note that ξ^+ and ξ^- are expressed as random variables since they are functions of X, a random vector. Therefore, we also write that $\xi^+ \sim \mathcal{P}^+$ (i.e., ξ^+ has distribution $\mathcal{P}^+$) and $\xi^- \sim \mathcal{P}^-$.

Given a diagnostic threshold t, a patient is classified as positive if his or her risk estimate exceeds t. Otherwise, the patient is classified as negative. Let $\mathbb{1}(\cdot)$ denote the indicator function. Then, given a threshold t and a true-positive patient with risk estimate ξ^+, we define sensitivity $se(t, \xi^+) := \mathbb{1}(\xi^+ \geq t)$ and false-negative $fn(t, \xi^+) := \mathbb{1}(\xi^+ < t)$. Similarly, given a threshold t and a true-negative patient with risk estimate ξ^-, we define specificity $sp(t, \xi^-) := \mathbb{1}(\xi^- \leq t)$ and false-positive $fp(t, \xi^-) := \mathbb{1}(\xi^- > t)$. The conditional expectations $\mathbb{E}[se(t, \xi^+)]$ and $\mathbb{E}[sp(t, \xi^-)]$ represent the expected sensitivity and specificity, respectively, under threshold t based on each population's risk estimates. The expected false-positive rate, $\mathbb{E}[fp(t, \xi^-)]$, and expected false-negative rate, $\mathbb{E}[fn(t, \xi^+)]$, are defined similarly.

3.2 Stochastic Programming Formulation

We consider an upper threshold u and lower threshold l such that any risk estimate above u is classified as positive and any risk estimate below l is classified as negative. We aim to determine

the values of u and l which balance sensitivity and specificity while also limiting the rate of false-positive and false-negative classifications. The region between u and l defines a range of risk estimates for which diagnostic decisions may be deferred due to elevated risk of false-positive or false-negative diagnoses. This region also reflects a range over which risk scores were not estimated well and clinical judgment should be favored. Since we model patient risk estimates as random variables, we formulate the Two-Threshold Problem (TTP) as the following stochastic program.

$$\begin{aligned}
(\text{TTP}) \quad \max_{u,l} \quad & \lambda \mathbb{E}[se(u, \xi^+)] + (1-\lambda)\mathbb{E}[sp(l, \xi^-)] && \text{(1a)} \\
\text{s.t.} \quad & \mathbb{E}[fp(u, \xi^-)] \leq \gamma^{fp} && \text{(1b)} \\
& \mathbb{E}[fn(l, \xi^+)] \leq \gamma^{fn} && \text{(1c)} \\
& 0 \leq l \leq u \leq 1. && \text{(1d)}
\end{aligned}$$

The objective function (1a) uses the parameter $\lambda \in (0, 1)$ to specify the relative importance of sensitivity to specificity. Higher values of λ imply that greater importance is placed on correctly classifying true-positives. Alternatively, setting λ equal to the proportion of true-positives in the overall patient population equates (1a) to maximizing the probability of making a correct diagnosis decision. The value of λ is chosen by the decision-maker and making an appropriate choice can be difficult. Fortunately, we show, in Section 3.4, that when it is optimal to use two distinct decision thresholds, the choice of λ does not affect the optimal solution. The constraints (1b) and (1c) imply that, based on the thresholds u and l, the false-positive and false-negative rates should not exceed the bounds γ^{fp} and γ^{fn}, respectively, where $\gamma^{fp}, \gamma^{fn} \in (0, 1)$. These parameters can reflect clinically acceptable levels of diagnostic accuracy. We provide guidelines for choosing these parameters in Section 7.1. Finally, constraint (1d) ensures that the upper threshold u remains above the lower threshold l, and that both are between 0 and 1.

3.3 Approximating the Two-threshold Problem

Very little can be said about the form of TTP's objective function (1a), even if the distributions $\mathcal{P}^+$ and $\mathcal{P}^-$ are known exactly. For example, in general, (1a) may not be concave, continuous, or differentiable everywhere. However, the functions $se(u, \xi^+)$ and $sp(l, \xi^-)$ are monotone in u

and l for every $\xi^+, \xi^- \in (0,1)$, respectively. Using these properties, we approximate TTP with TTP*:

$$\begin{aligned} (\text{TTP}^*) \quad & \min_{u,l} \quad \phi u - (1-\phi) l \\ & \text{s.t.} \quad \text{(1b)-(1d)}, \end{aligned} \tag{2}$$

where $\phi \in (0,1)$ is also a weighting parameter which may not necessarily equal λ but also serves the purpose of defining the relative importance of one threshold to the other. In Section 3.4, we analyze TTP* and identify cases in which the optimal solutions to TTP and TTP* coincide.

3.4 Structural Properties

In this section, we highlight structural properties which are useful in understanding TTP through TTP*. First, we show that by expressing the constraints (1b) and (1c) in terms of quantiles, TTP* is equivalent to a linear program. Based on this linear program, we show that the optimal solution to TTP* is either a two-threshold solution or a one-threshold solution, and that the optimal solution can be characterized based on the parameters of TTP*. We then relate the optimal solution of TTP* with that of TTP, showing that a two-threshold solution of TTP* is optimal in TTP, but not necessarily for one-threshold solutions. These results indicate that TTP can be solved, in two-threshold cases, using quantile estimation.

Throughout this section, we assume that probabilities of any true-positive or true-negative risk estimate falling on a fixed threshold t is zero. That is,

A1 For any fixed t, $\mathbb{P}(t = \xi^+) = 0$ and $\mathbb{P}(t = \xi^-) = 0$.

Assumption A1 implies that the distributions $\mathcal{P}^+$ and $\mathcal{P}^-$ are continuous which is not too restrictive since most risk estimation models output risk scores along a continuum.

Now, we relate the optimal solution to TTP* with the parameters ϕ, γ^{fp} and γ^{fn}. Define

$$u(\gamma^{fp}) := \inf\{u \in [0,1] : g_1(u) \leq \gamma^{fp}\} \text{ and} \tag{3}$$

$$l(\gamma^{fn}) := \sup\{l \in [0,1] : g_2(l) \leq \gamma^{fn}\}, \tag{4}$$

where we define $g_1(u) := \mathbb{E}[fp(u, \xi^-)]$ and $g_2(l) := \mathbb{E}[fn(l, \xi^+)]$ for notational convenience. The quantity $u(\gamma^{fp})$ specifies the smallest value of u which ensures that the false-positive rate is no more than γ^{fp} and similarly, $l(\gamma^{fn})$ denotes the greatest value of l such that the false-negative rate is no more than γ^{fn}. Note that $u(\gamma^{fp})$ and $l(\gamma^{fn})$ are quantiles. Specifically, $u(\gamma^{fp})$ is the $(1-\gamma^{fp})$-quantile of $\mathcal{P}^-$ and $l(\gamma^{fn})$ is the negative value of the $(1-\gamma^{fn})$-quantile of $\mathcal{P}^+$ (see online supplementary material for the derivation). This fact forms the basis for the solution methodology described in Section 4.1.

Now, by Theorem 7.48 in Shapiro et al. (2009), the functions $g_1(u)$ and $g_2(l)$ are finite-valued and continuous over $[0,1]$, so the function $g(u,l) := \max\{g_1(u), g_2(l)\}$ is finite-valued and continuous on $[0,1]^2$. The continuity of these functions along with the fact that the interval $[0,1]$ is compact and convex implies that $u(\gamma^{fp})$ and $l(\gamma^{fn})$ are attained on $[0,1]$. Now, in Proposition 1, we show that TTP* is equivalent to a linear program with constraints written in terms of $u(\gamma^{fp})$ and $l(\gamma^{fn})$.

Proposition 1. *TTP* is equivalent to the following linear program.*

$$\begin{aligned} \min_{u,l} \quad & \phi u - (1-\phi) l \\ s.t. \quad & u \geq u(\gamma^{fp}), \quad l \leq l(\gamma^{fn}), \quad 0 \leq l \leq u \leq 1. \end{aligned}$$

By recasting TTP* as a linear program, we can focus our analysis on the extreme-point solutions of TTP*. The following results characterize the types of optimal solution to TTP*.

Proposition 2 (Two-threshold Solutions)**.** *If $\phi \in (0,1)$ and both γ^{fp} and γ^{fn} are chosen such that $u(\gamma^{fp}) > l(\gamma^{fn})$, then $\left(u(\gamma^{fp}), l(\gamma^{fn})\right)$ is the unique optimal solution to TTP*.*

Proposition 2 implies that the parameter $\phi \in (0,1)$ has no effect if the optimal solution consists of two thresholds. Additionally, the conditions specified in Proposition 2 are satisfied if both γ^{fp} and γ^{fn} are "low enough". Finally, we have shown that the optimal solution may be obtained without solving an optimization problem if the values $u(\gamma^{fp})$ and $l(\gamma^{fn})$ can be computed exactly and the parameters γ^{fp} and γ^{fn} are chosen to satisfy the sufficient conditions. However, the distributions $\mathcal{P}^+$ and $\mathcal{P}^-$ are typically unknown, making $u(\gamma^{fp})$ and $l(\gamma^{fn})$ difficult to ascertain.

While we have identified conditions on γ^{fp} and γ^{fn} for which the optimal solution will consist of

two separate thresholds, we also seek to determine conditions under which it is optimal to have a single threshold, i.e., $u^* = l^*$. We specify these conditions in Proposition 3.

Proposition 3 (One-threshold Solutions)**.** *If γ^{fp} and γ^{fn} are chosen such that $u(\gamma^{fp}) \leq l(\gamma^{fn})$, then the optimal solution to TTP* consists of a single threshold, i.e., $u^* = l^*$. Furthermore,*

(a) if $\phi > 0.5$, the optimal threshold will be given by $t^* = u^* = l^* = u(\gamma^{fp})$,

(b) if $\phi < 0.5$, the optimal threshold will be given by $t^* = u^* = l^* = l(\gamma^{fn})$, and

(c) if $\phi = 0.5$, every $t \in [u(\gamma^{fp}), l(\gamma^{fn})]$ is optimal.

The implication of Proposition 3 is that if at least one of the bounds γ^{fp} or γ^{fn} is "loose" enough, a one-threshold solution is better than a two-threshold solution. Therefore, if it is desired to constrain just one of false-positive or false-negative rates, then a one-threshold solution will be at least as good as two-threshold solutions. Alternatively, this result can be interpreted to mean that using one decision threshold will not be optimal if aiming to keep both false-positive and false-negative rates low while maximizing sensitivity and specificity. Furthermore, Proposition 3 implies that in parameterizing TTP*, one only needs to consider the three choices of ϕ described. This characterization is extremely useful in a practical sense, since choosing appropriate values for parameters is often challenging when implementing such models in practice.

Based on the results established in Propositions 2 and 3, we can relate the optimal solutions of TTP* and TTP based on the parameter choices.

Theorem 1 (Relation Between TTP* and TTP Optimal Solutions)**.** *Suppose that γ^{fp} and γ^{fn} are chosen identically for TTP* and TTP.*

1. *If γ^{fp} and γ^{fn} are chosen such that $u(\gamma^{fp}) > l(\gamma^{fn})$, then, for any $\phi \in (0, 1)$, the optimal solution to TTP* is also optimal in TTP for any $\lambda \in (0, 1)$.*

2. *If γ^{fp} and γ^{fn} are chosen such that $u(\gamma^{fp}) \leq l(\gamma^{fn})$, then the following statements hold.*

   (a) If $\phi > 0.5$, then the optimal solution to TTP is optimal in TTP with $\lambda = 1$,*

   (b) If $\phi < 0.5$, then the optimal solution to TTP is optimal in TTP with $\lambda = 0$, and*

   (c) For any $\lambda \in (0, 1)$, there exists a one-threshold optimal solution to TTP. Furthermore

any one-threshold optimal solution to TTP is optimal in TTP with* $\phi = 0.5$.

From Theorem 1, we find that the optimal solution to TTP* is optimal for TTP in the two-threshold case, regardless of how the parameters ϕ and λ are chosen. Specifically, for two-threshold solutions, the optimal thresholds to TTP* at some initial value of ϕ are already tight on their constraints (1b)-(1c). Since changing the value of ϕ will not change the direction of improvement for the objective function (2), the optimal thresholds will also remain unchanged. Therefore, the optimal solution to TTP* does not depend on the value of ϕ. Since the optimal solution to TTP* is optimal in TTP (regardless of λ) for the two-threshold solution case, the optimal solution also does not depend on λ.

However, the one-threshold case for TTP* does not translate well to TTP. To derive stronger results for the one-threshold case, we need additional assumptions regarding the form of the objective function (1a). For instance, if (1a) is monotone over the interval $[u(\gamma^{fp}), l(\gamma^{fn})]$, then the optimal solution will coincide with one of the endpoints. Unfortunately, the form of (1a) is difficult to know *a priori* and it may be the case that (1a) has many maxima over $[u(\gamma^{fp}), l(\gamma^{fn})]$. However, if it is known that TTP has a one-threshold solution, then it suffices to use existing one-threshold methods (see Section 2). Since such instances are well-studied, we emphasize our analysis of TTP over two-threshold solutions.

Propositions 2 and 3 characterize the optimal solution to TTP* based on ϕ, γ^{fp}, and γ^{fn}. Specifically, the optimal thresholds can be constructed if $u(\gamma^{fp})$ and $l(\gamma^{fn})$ are known or estimated. Practically speaking, this result implies that one can solve TTP without sophisticated optimization software by using quantile estimation methods instead (see Section 4.1).

3.5 Utility-based Frameworks

Our formulation of TTP does not take into account any utilities because we aimed to determine a method which generalizes well across many applications, including those for which utilities are poorly estimated. However, utilities are an important component of decision-theoretic frameworks. In this section, we extend our analysis and analyze two utility-based frameworks for identifying diagnosis decision thresholds. Specifically, in Section 3.5.1, we provide a general set of conditions which ensure that the analytical results for TTP (as derived in Section 3.4) still hold. Based on

these conditions, we provide a specific utility-based framework for TTP. Then, in Section 3.5.2, we study a utility-based objective function which does not fit the conditions provided in Section 3.5.1. Nevertheless, we derive conditions which must hold such that a two-threshold solution (e.g., one obtained by solving TTP*) is better than a fixed one-threshold solution.

3.5.1 Utility-based Two-Threshold Problem

In this section, we identify a general class of objective functions and constraints which, when used in the TTP framework, ensure that the structural results from Section 3.4 still hold.

Remark 1. *The results shown in Section 3.4 hold for general optimization problems of the form*

$$\begin{aligned} \max_{u,l} \quad & g(\mathbb{E}[se(u,\xi^+)], \mathbb{E}[sp(l,\xi^-)]) \\ s.t. \quad & h_1(\mathbb{E}[fp(u,\xi^-)]) \leq b_1 \\ & h_2(\mathbb{E}[fn(l,\xi^+)]) \leq b_2 \\ & 0 \leq l \leq u \leq 1, \end{aligned}$$

where $g(\cdot)$ is non-increasing in u and non-decreasing in l over the feasible region, $h_1(\cdot)$ is non-increasing in u, $h_2(\cdot)$ is non-decreasing in l, and $b_1, b_2 > 0$.

Through Remark 1, we now show how to incorporate utilities into the TTP framework. Let

- q be the proportion of true-positives in the population,
- $r^+ > 0$ and $r^- > 0$ be the utilities associated with correctly classifying true-positives and true-negatives, respectively,
- $c^+ > 0$ and $c^- > 0$ be the costs associated with false-negatives and false-positives, respectively, and
- $b^+ > 0$ and $b^- > 0$ be the maximum allowable costs associated with false-negatives and false-positives, respectively.

Examples of utilities for r^+, r^-, c^+ and c^- include quality-adjusted life-years — a common utility measure used in medical decision-making and public health. Now, we can formulate the Utility-

based Two-Threshold Problem (UTTP) as:

$$\begin{aligned} \text{(UTTP)} \quad \max_{u,l} \quad & r^+ q\mathbb{E}[se(u,\xi^+)] + r^-(1-q)\mathbb{E}[sp(l,\xi^-)] \\ \text{s.t.} \quad & c^-(1-q)\mathbb{E}[fp(u,\xi^-)] \leq b^- \\ & c^+ q\mathbb{E}[fn(l,\xi^+)] \leq b^+ \\ & 0 \leq l \leq u \leq 1. \end{aligned}$$

UTTP aims to determine thresholds u and l which maximize the expected utility associated with correct classification without exceeding expected misclassification costs. Since the monotonicity of the objective function and constraints for UTTP match the monotonicity for TTP, Theorem 1 applies to UTTP.

3.5.2 Utility-based Cost Function

In this section, we consider a more general objective function than the one considered in Section 3.5.1. For a fixed set of thresholds (u, l), define

$$\begin{aligned} p^+(u,l) &:= \mathbb{P}(\text{Correct Classification}|u,l) = \mathbb{E}[se(u,\xi^+)]q + \mathbb{E}[sp(l,\xi^-)](1-q) \\ p^-(u,l) &:= \mathbb{P}(\text{Misclassification}|u,l) = \mathbb{E}[fn(l,\xi^+)]q + \mathbb{E}[fp(u,\xi^-)](1-q) \\ p^D(u,l) &:= \mathbb{P}(\text{Defer}|u,l) = 1 - p^+(u,l) - p^-(u,l), \end{aligned}$$

where q specifies the proportion of people who are true-positives in the population subject to diagnosis. Furthermore, consider three utilities c^+, c^-, and c^D associated with a correct classification, misclassification, and deferred diagnosis decision, respectively. We assume that $c^+ \geq c^D \geq c^-$, i.e., it is better to correctly classify a patient than to defer a patient and it is better to defer a patient than to misclassify a patient. Given this ordering of utilities, we can express c^D as $c^D = \delta c^- + (1-\delta)c^+$ for some $\delta \in [0,1]$. Then, we can define an expected utility function of the form

$$J(u,l) = p^+(u,l)c^+ + p^-(u,l)c^- + p^D(u,l)c^D. \tag{5}$$

In general, $J(u,l)$ does not satisfy the conditions in Remark 1. More specifically, $J(u,l)$ is not generally non-increasing in u and non-decreasing in l. We note that (5) is nearly identical to the

regret-based framework in Section 5.1 of Shapiro (1999), except we include a utility associated with deferred diagnosis decisions. With the function (5), one may be interested in comparing a fixed two-threshold solution (u, l) with a one-threshold solution (t, t). In particular, how high must the utility c^D be for (u, l) to be a better solution than (t, t)? We provide intuitive conditions to answer this very question in Proposition 4.

Proposition 4. *Consider a one-threshold solution (t, t) and a two-threshold solution (u, l) such that $p^D(u, l) > 0$. Then, $J(u, l) \geq J(t, t)$ if and only if*

$$\delta \leq \frac{p^-(t, t) - p^-(u, l)}{p^D(u, l)}. \tag{6}$$

Proposition 4 provides insights in the trade-off between two-threshold and one-threshold solutions. Let us first examine the expression on the right-hand side of (6). Note that the numerator of (6) specifies the misclassifications which are saved by switching from (t, t) to (u, l) and the denominator specifies the overall probability of being deferred under (u, l). Thus, we can interpret the right-hand side of (6) as the proportion of deferred decisions under (u, l) which consist of would-be misclassifications under (t, t). Therefore, Proposition 4 tells us that the utility associated with deferred decisions c^D can only be as low as the proportion of saved misclassifications by switching from (t, t) to (u, l). Furthermore, if we interpret δ to be the expected proportion of deferred decisions which end up being misclassified, Proposition 4 tells us that it is better to switch from (t, t) to (u, l) if the expected proportion of misclassifications after a deferred decision is no more than the misclassifications saved by switching from (t, t) to (u, l). To summarize, this analysis shows that two-threshold solutions are particularly useful when either: (i) the utility c^D is relatively high compared to c^+ and c^-, (ii) if most of deferred decisions consist of would-be misclassifications, or (iii) if ultimately, few deferred decisions end up being misclassified.

3.6 Extensions to Multi-class Diagnosis

In our formulation of TTP, we focused on the case in which the population consists of two subpopulations. In this section, we extend our modeling framework to handle the case when there are three or more subpopulations, i.e., multi-class diagnosis. We specifically focus on the cases of (1) multi-label classification and (2) ordinal classification.

3.6.1 Multi-label Classification via Binary Relevance

In this section, we extend TTP to address multi-label classification problems through the Binary Relevance (BR) method. In multi-label classification through BR, one aims to determine which labels should be assigned to a patient given an estimate for the likelihood that the patient should be assigned each label. Examples of this problem include determining which concussion subtypes a patient may have based on a multi-dimensional clinical assessment (Collins et al., 2014; Maruta et al., 2018) or determining which chronic diseases a person may have based on electronic health record information (Zufferey et al., 2015).

The specific problem setting is as follows. The patient population consists of patients who may belong to any subset of $K \geq 2$ classes. Hence, in this patient population, there are 2^K mutually exclusive subpopulations. A randomly drawn patient from this population is associated with random vectors (X, Y), where $X \in \mathcal{X}$ consists of patient characteristics (and $\mathcal{X}$ is the set of patient characteristics) and $Y \in \{0,1\}^K$ is a vector where the k^{th} entry y_k is equal to 1 if the patient is associated with class k and 0 otherwise. Furthermore, there are K risk estimation models $f_1, ..., f_K$ where each risk estimation model $f_k(X) : \mathcal{X} \to (0,1)$ approximates $\mathbb{P}(y_k = 1|X)$. Examples of such models include multinomial regression, multi-class support vector machines, and multi-class perceptrons. The BR method associates a patient with class label k if $f_k(X) \geq t_k$, where $t_k \in [0,1]$ is a class k decision threshold.

Note that BR treats risk estimates for each class label as if they were independent of risk estimates from all other class labels. This simplifying assumption is a known limitation to the BR framework (Zhang et al., 2018), though it has been shown to outperform more intricate modeling approaches for some applications (Zufferey et al., 2015). To this end, ensemble learning methods have been developed which account for dependence between labels in the creation of risk estimation models $f_1, ..., f_k$ (Godbole and Sarawagi, 2004; Zhang and Zhang, 2010; Read et al., 2011). These existing methods may be applied prior to determining each of the k decision thresholds.

The performance of a set of thresholds $t_1, ..., t_K$ can be evaluated similarly to the binary classification problem which is the focus of this manuscript. For brevity, we denote class k risk estimates as

$$\xi_k^+ := f_k(X|y_k = 1) \text{ and}$$

$$\xi_k^- := f_k(X|y_k = 0),$$

for each $k = 1, ..., K$. Given a patient with label $y_k = 1$ and some threshold t_k, we define the functions

$$se_k(t_k, \xi_k^+) := \mathbb{1}\{\xi_k^+ \geq t_k\}$$
$$fn_k(t_k, \xi_k^+) := \mathbb{1}\{\xi_k^+ < t_k\}$$

as class k sensitivity and false-negative, respectively. Similarly, for patients with $y_k = 0$ and a given threshold t_k, we define

$$sp_k(t_k, \xi_k^-) := \mathbb{1}\{\xi_k^- \leq t_k\}$$
$$fp_k(t_k, \xi_k^-) := \mathbb{1}\{\xi_k^- > t_k\}$$

as class k sensitivity and false-positive, respectively.

In our extension to BR, we consider an upper threshold u_k and lower threshold l_k such that a risk estimate above u_k labels a patient as class k and below l_k does not label a patient as class k. For patients with risk estimates between u_k and l_k, their association with label k is inconclusive. For each class k, we aim to determine the thresholds u_k and l_k by maximizing the expected class k sensitivity and specificity, respectively, while restricting the expected false-positive and false-negative rates. This problem can be represented with the Multi-label Two Threshold Problem (MTTP), which we define as

$$\text{(MTTP)} \quad \max_{u_1,...,u_K,l_1,...,l_K} \quad \sum_{k=1}^{K} \left(\lambda_k \mathbb{E}[se_k(u_k, \xi_k^+)] + (1 - \lambda_k)\mathbb{E}[sp_k(l_k, \xi_k^-)] \right) \tag{7a}$$

$$\text{s.t.} \quad \mathbb{E}[fp_k(u_k, \xi_k^-)] \leq \gamma_k^{fp} \quad \text{for all } k = 1, ..., K \tag{7b}$$

$$\mathbb{E}[fn_k(l_k, \xi_k^+)] \leq \gamma_k^{fn} \quad \text{for all } k = 1, ..., K \tag{7c}$$

$$0 \leq l_k \leq u_k \leq 1, \text{ for } k = 1, ..., K, \tag{7d}$$

where the parameters $\lambda_k \in (0, 1)$ indicate the preference for sensitivity over specificity in class k for

all $k = 1, ..., K$. If we take λ_k to be the proportion of patients with $y_k = 1$ in the overall population, then the objective function (7a) can be interpreted as maximizing the labeling accuracy.

Following the BR framework, the thresholds in each class k are treated independently of thresholds in class $j \neq k$. Through this simplification, one could decompose MTTP into solving TTP K times. Therefore, the results derived for TTP can be applied directly to MTTP. Specifically, we show in Corollary 1 that the optimal solution can be written in terms of the following quantiles:

$$u_k(\gamma_k^{fp}) := \min\{u : \mathbb{E}[fp(u, \xi_k^-)] \leq \gamma_k^{fp}\} \text{ for all } k = 1, ..., K$$
$$l_k(\gamma_k^{fn}) := \max\{l : \mathbb{E}[fn(l, \xi_k^k)] \leq \gamma_k^{fn}\} \text{ for all } k = 1, ..., K.$$

Corollary 1. *If $u_k(\gamma_k^{fp}) > l_k(\gamma_k^{fn})$ for all $k = 1, ..., K$, the optimal solution to MTTP is given by $u_k^* = u_k(\gamma_k^{fp})$ and $l_k^* = l_k(\gamma_k^{fn})$ for all $k = 1, ..., K$.*

Remark 2. *If it were the case that each patient belongs to only 1 of the K classes, then this classification problem is no longer a multi-label classification problem. Instead, it is a multi-class classification problem. To this end, the One-versus-All approach to multi-class classification is similar to the BR approach studied in this section. Specifically, both methods create K risk estimation models $f_1, ..., f_K$ with one model for each class. The key distinction between these two methods is that in the One-versus-All framework, each patient may only be assigned a single label. For example, one may wish to classify which type of skin lesion a patient has among 7 different possibilities (Tschandl et al., 2019). While MTTP could certainly be applied directly to this related problem, it may not be well-suited for this scenario since there exists the possibility that MTTP recommends a patient for more than one class. This result would be difficult to interpret and could limit its application in practice.*

3.6.2 Ordinal Classification

In this section, we present a model similar to TTP which can be used to identify thresholds for ordinal classification problems. Ordinal classification aims to determine which severity level a patient belongs to for a certain disease. Currently, there is not consensus on severity ratings for concussion, though one may try to predict whether a patient has no concussion, a concussion with normal recovery, or a concussion with pro-longed recovery (Lau et al., 2011). Another example of

ordinal classification is in predicting a patient's degree of recovery six months after traumatic brain injury (Roozenbeek et al., 2011).

In this problem setting, there are $K \geq 2$ mutually exclusive patient classes ordered by severity. That is, class 1 is the least severe class and class K is the most severe. A randomly chosen patient is associated with the tuple (X, Y), where $X \in \mathcal{X}$ is a random vector of patient characteristics from the set of patient characteristics $\mathcal{X}$ and the random variable $Y \in \{1, ..., K\}$ describes his or her class. Rather than directly estimating Y based on patient characteristics X, ordinal classification methods estimate a continuous latent variable $\xi \in [0, 1]$. Then, given $K - 1$ ordered thresholds $t_1, ..., t_{K-1}$ such that $t_0 := 0 \leq t_1 \leq ... \leq t_{K-1} \leq t_K := 1$, a patient is classified as being in class k if $t_{k-1} \leq \xi \leq t_k$. Specifically, we assume that there is a severity score model $f : \mathcal{X} \to [0, 1]$ which estimates the latent variable ξ based on patient characteristics X. Such models include ordinal logistic regression and ordinal support vector machines. Throughout the remainder of this section, we define $\xi^k := f(X|Y = k)$ to be the severity score for a random patient from class $Y = k$ with characteristics X.

We aim to identify the thresholds $t_1, ..., t_{K-1}$ which correctly classify as many patients as possible into the correct severity class while limiting the number of patients misclassified into lower and higher severity classes. Using the same functions defined in Section 3.1, we define the Ordinal Threshold Problem (OTP) as follows:

$$\text{(OTP)} \quad \max_{t_1,...,t_{K-1}} \quad \sum_{k=0}^{K-1} \lambda_{k+1} \mathbb{E}[se(t_k, \xi^{k+1})] \tag{8a}$$

$$\text{s.t.} \quad \mathbb{E}[fp(t_k, \xi^j)] \leq \gamma_{k,j}^{fp} \text{ for all } j \leq k, k = 1, ..., K-1 \tag{8b}$$

$$\mathbb{E}[fn(t_k, \xi^j)] \leq \gamma_{k,j}^{fn} \text{ for all } j > k, k = 1, ..., K-1 \tag{8c}$$

$$t_k \leq t_{k+1} \text{ for all } k = 1, ..., K-2, \tag{8d}$$

where $t_0 = 0$ and the weight parameters $\lambda_k \in (0, 1)$ satisfy $\sum_{k=1}^{K} \lambda_k = 1$. These weight parameters can be interpreted as the relative importance of correctly classifying one class to the others. In the objective function, we focus on modeling sensitivity — rather than combining sensitivity and specificity — since there is generally more urgency in correctly identifying conditions which are

more severe compared to those which are less severe. For example, this urgency may arise in classification of traumatic brain injury since the health-related costs (resp., quality of life) for patients with mild or moderate traumatic brain injury are estimated to be less (resp., greater) compared to patients with severe traumatic brain injury (Andelic et al., 2009; Humphreys et al., 2013). The constraint parameters $\gamma_{k,j}^{fp}$ (resp., $\gamma_{k,j}^{fn}$) specify the maximum expected false-positive rate (resp., false-negative rate) for class j against threshold t_k. We assume that for each $k = 1, ..., K$, the constraint parameters are chosen such that $\gamma_{k+1,j}^{fp} \leq \gamma_{k,j}^{fp}$ for all $j \leq k$ and $\gamma_{k+1,j}^{fn} \geq \gamma_{k,j}^{fn}$ for all $j > k$. If these parameters are not chosen in this way, then OTP is guaranteed to be infeasible.

We now show that, much like TTP, the optimal solution to OTP can be defined in terms of quantiles. Specifically, for all $k = 1, ..., K$, we define the following quantiles:

$$\underline{t}^k(\gamma^{fp}) := \min\{t : \mathbb{E}[fp(t, \xi^k)] \leq \gamma^{fp}\}$$
$$\bar{t}^k(\gamma^{fn}) := \max\{t : \mathbb{E}[fn(t, \xi^k)] \leq \gamma^{fn}\}.$$

For ease of notation, we also define

$$\underline{t}_k := \max_{j \leq k} \underline{t}^j(\gamma_{k,j}^{fp}) \tag{9}$$
$$\bar{t}_k := \min_{j > k} \bar{t}^j(\gamma_{k,j}^{fn}), \tag{10}$$

for all $k = 1, ..., K-1$. Notice that $\bar{t}_k \leq \bar{t}_{k+1}$ and $\underline{t}_k \geq \underline{t}_{k+1}$ based on how the parameters $\gamma_{k,j}^{fp}$ and $\gamma_{k,j}^{fn}$ need to be chosen.

Now, we show that if OTP is feasible, the optimal solution to OTP can be stated in terms of the quantiles defined in (9)-(10).

Theorem 2. *OTP is feasible if and only if $\underline{t}_k \leq \bar{t}_k$ for all $k = 1, ..., K$. Furthermore, if OTP is feasible, then the optimal solution is given by $t_k^* = \underline{t}_k$ for all $k = 1, ..., K-1$.*

Theorem 2 tells us that the optimal solution to OTP can be stated in terms of quantiles and therefore, computed via quantile estimation. Furthermore, OTP is guaranteed to be feasible if we set $\gamma_{k,j}^{fn} = 1$ for all $j > k$. Thus, this formulation of OTP actually emphasizes constraints on the false-positive rate more than it does the false-negative rate. However, one could also formulate

OTP based on maximizing specificity with more emphasis on constraining false-negative rates. Specifically, if (8a) was replaced with

$$\max_{t_1,...,t_K} \quad \sum_{k=1}^{K} \lambda_k \mathbb{E}[sp(t_k, \xi^k)],$$

where $t_K = 1$, then it is straightforward to show that the feasibility result in Theorem 2 holds and the optimal solution to this new problem is given by $t_k^* = \bar{t}_k$ for $k = 1, ..., K-1$. That is, with this alternative objective function, the constraints on false-negative rates are emphasized. Using the results from Theorem 2 to develop solutions for OTP, we use simulation to assess the performance of OTP in the online supplementary material.

While restricting the objective function to containing only sensitivity (or specificity) was motivated by practical reasons, it serves a technical purpose as well. Specifically, the objective function would no longer be monotone in the thresholds $t_1, ..., t_K$ if both sensitivity and specificity were included. As a result, little can be said about the form taken by the optimal solution to OTP without making assumptions about the distributions of risk estimates $\xi^1, ..., \xi^K$. To this end, finding ways to incorporate both sensitivity and specificity in OTP would be an interesting direction for future research.

4 Data-driven Solution Methods

In Section 3.4, we showed important analytical properties which relate TTP* and TTP. In particular, TTP* provides an optimal solution to TTP if the optimal solution is a two-threshold solution. However, solving TTP* is not straightforward since obtaining the functions (1b) and (1c) may not be possible, even if the distribution $\mathcal{P}$ or $\mathcal{P}^+, \mathcal{P}^-$ are known exactly. On the other hand, data may be available which can be used to estimate (1b) and (1c). In this section, we propose two data-driven methods for solving TTP* tractably: quantile estimation and distributionally robust optimization.

In the remainder of this section, we denote a data sample of size N as $\hat{P}_N = \{\hat{\xi}_1, .., \hat{\xi}_N\}$. Let the set $\hat{P}_N^+ = \{\hat{\xi} \in \hat{P}_N : \text{true-positive}\} = \{\hat{\xi}_1^+, ..., \hat{\xi}_{N^+}^+\}$ be the part of the data sample consisting of true-positives. Similarly, let $\hat{P}_N^- = \{\xi \in \hat{P}_N : \text{true-negative}\} = \{\hat{\xi}_1^-, ..., \hat{\xi}_{N^-}^-\}$ be the true-negatives. The sets $\hat{P}_N^+$ and $\hat{P}_N^-$ are mutually exclusive and $\hat{P}_N = \hat{P}_N^+ \cup \hat{P}_N^-$, so $N^+ + N^- = N$.

4.1 Quantile Estimation

In Section 3.4, we have shown that the optimal solution to TTP* is characterized by the quantiles $u(\gamma^{fp})$ and $l(\gamma^{fn})$. Thus, we propose to solve TTP* using Harrell-Davis quantile estimation (Harrell and Davis, 1982), as it has been shown to outperform the standard quantile estimation method across various distributions. We also performed auxiliary analyses and found that the Harrell-Davis quantile estimation method outperformed standard sample average approximation techniques. Once we have estimated $u(\gamma^{fp})$ and $l(\gamma^{fn})$, we apply Propositions 2 and 3 to estimate the optimal solution to TTP*. We refer to this solution as TTP*-Q.

Remark 3. *Different application needs may call for different quantile estimation methods. For instance, if γ^{fp} or γ^{fn} are set very close to 0, it may be more appropriate to use a method for estimating extreme quantiles (Daníelsson and de Vries, 1997).*

An important factor in obtaining high out-of-sample feasibility for TTP*-Q is the choice of constraint parameters γ^{fp} and γ^{fn}. Specifically, solving TTP*-Q with constraint parameters $\gamma_\tau^{fp} < \gamma^{fp}$ and $\gamma_\tau^{fn} < \gamma^{fn}$ can improve its out-of-sample feasibility. However, choosing values of γ_τ^{fp} and γ_τ^{fn} which are too small can result in overly conservative thresholds. In the online supplementary appendix, we describe how the values of γ_τ^{fp} and γ_τ^{fn} can be calibrated using data.

4.2 Data-driven Distributionally Robust Optimization

A drawback to the quantile estimation approach in Section 4.1 is that it may require a very large sample size N to sufficiently approximate the distribution $\mathcal{P}$. However, obtaining a large enough sample of data may not be possible, e.g., when data collection is very costly or time-consuming. Hence, we consider the context in which the data samples $\hat{P}_N$ do not sufficiently represent the true population distribution $\mathcal{P}$ and we wish to create thresholds which are robust to the worst-case differences between $\hat{P}_N$ and $\mathcal{P}$. We consider the following distributionally robust TTP* (TTP*-DR) model:

$$
\begin{aligned}
\text{(TTP*-DR)} \quad \min_{u,l} \quad & \phi u - (1-\phi) l \\
\text{s.t.} \quad & \sup_{\mathcal{Q}^- \in D^-} \mathbb{E}_{\mathcal{Q}^-}[fp(u, \xi^-)] \leq \gamma^{fp} && \text{(11a)} \\
& \sup_{\mathcal{Q}^+ \in D^+} \mathbb{E}_{\mathcal{Q}^+}[fn(l, \xi^+)] \leq \gamma^{fn} && \text{(11b)}
\end{aligned}
$$

$$0 \leq l \leq u \leq 1,$$

where D^- (respectively, D^+) represents a family of probability distributions that are plausible candidates of $\mathcal{P}^-$ (respectively, $\mathcal{P}^+$) and is termed the ambiguity set. For TTP*-DR, we propose to construct D^- and D^+ in a data-driven fashion by considering all distributions which are "close" to the data sample $\hat{P}_N$ based on the Wasserstein distance metric. We opt to construct our ambiguity set based on the Wasserstein distance metric since such ambiguity sets have been shown, under mild conditions, to satisfy strong finite sample guarantees, asymptotic consistency, and tractability (Mohajerin Esfahani and Kuhn, 2018).

To define the Wasserstein-based ambiguity set, we must first define some preliminary notation. Let $\mathcal{M}(\Xi)$ represent the set of all probability distributions over the support $\Xi = [0, 1]$, i.e., all distributions which generate risk scores. Then, the Wasserstein distance between any two distributions $\mathcal{Q}_1, \mathcal{Q}_2 \in \mathcal{M}(\Xi)$ is defined by

$$W(\mathcal{Q}_1, \mathcal{Q}_2) := \inf_{\Pi \in \Xi \times \Xi} \left\{ \int_{\Xi^2} |\xi - \xi'| \Pi(d\xi, d\xi') : \begin{array}{l} \Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \\ \text{with marginals } \mathcal{Q}_1 \text{ and } \mathcal{Q}_2, \text{ respectively,} \end{array} \right\} \quad (12)$$

where $\xi, \xi' \in \Xi$. Now, let $\hat{\mathcal{P}}_N$ be an empirical uniform distribution centered on the data sample $\hat{P}_N$. Then, for any Wasserstein radius $\epsilon > 0$, we define our ambiguity set as

$$D_\epsilon(\hat{\mathcal{P}}_N) := \{\mathcal{Q} : W(\mathcal{Q}, \hat{\mathcal{P}}_N) \leq \epsilon\}. \quad (13)$$

The ambiguity set (13) represents the set of all probability distributions which are within an ϵ distance, based on (12), to the data-driven empirical distribution $\hat{\mathcal{P}}_N$. Intuitively, larger choices of ϵ lead to more robust (i.e., conservative) solutions. Now, let ϵ^+ and ϵ^- denote chosen Wasserstein radii for the data samples $\hat{P}_N^+$ and $\hat{P}_N^-$, respectively. In the online supplementary appendix, we detail a cross-validation scheme to calibrate the selection of Wasserstein radii ϵ^+ and ϵ^- based on the data $\hat{P}_N^+$ and $\hat{P}_N^-$. Setting the data-driven ambiguity sets as $D^+ := D_{\epsilon^+}(\hat{\mathcal{P}}_N^+)$ and $D^- := D_{\epsilon^-}(\hat{\mathcal{P}}_N^-)$,

TTP*-DR becomes

$$\begin{aligned}
\min_{u,l} \quad & \phi u - (1-\phi) l \\
\text{s.t.} \quad & \sup_{\mathcal{Q} \in D^-} \mathbb{E}_{\mathcal{Q}}[fp(u, \xi^-)] \leq \gamma^{fp} && \text{(14a)} \\
& \sup_{\mathcal{Q} \in D^+} \mathbb{E}_{\mathcal{Q}}[fn(l, \xi^+)] \leq \gamma^{fn} && \text{(14b)} \\
& 0 \leq l \leq u \leq 1.
\end{aligned}$$

Due to the optimization problems embedded in the constraints (14a)-(14b), solving TTP*-DR is not straightforward. Furthermore, this optimization occurs over the space of all probability distributions in D^+ and D^-. To this end, we show in the online supplementary material that the constraints (14a) and (14b) can be reformulated into tractable problems. Specifically, when the thresholds u and l are fixed, these reformulations are linear programs which scale in size based on the size of the data sample P_N and can be handled efficiently by most modern commercial solvers (e.g., GUROBI, AMPL, and MOSEK). However, since u and l are not fixed in TTP*-DR, the reformulations present bilinear constraints. Nevertheless, since the left-hand sides of (14a) and (14b) are monotone in u and l, respectively, the optimal values of u and l can still be determined in polynomial time using algorithms such as bisection line search.

5 Simulation Analysis

In this section, we use simulation to analyze the performance of TTP under various conditions. Specifically, we first analyze the feasibility and optimality of TTP*-Q and TTP*-DR with respect to known distributions in Section 5.1. Then in Section 5.2, we perform a numerical analysis to compare the performance of TTP*-Q with an optimized single threshold.

5.1 Feasibility and Optimality of TTP*-Q and TTP*-DR

In this section, we use simulation to estimate the performance of TTP*-Q and TTP*-DR under different underlying risk estimation distributions and sample sizes. We aim to identify the situations for which each solution methodology performs well.

We performed our simulation analysis under the following settings.

Underlying risk estimation distributions: The Beta distribution is commonly used to simulate random variables with the support $[0, 1]$ and its shape parameters can be manipulated to change the distribution's mean, variance, and skewness. Therefore, we assumed that $\mathcal{P}^+ = \text{Beta}(0.55v, 0.45v)$ and $\mathcal{P}^- = \text{Beta}(0.45v, 0.55v)$ where $v \in \{1, 10, 50, 100\}$. Larger values of v reflect a higher quality of risk estimation model as measured by area under the receiver operating characteristic curve (AUROC). Specifically, v being equal to $1, 10, 50$, and 100 is equivalent to an AUROC of $0.585, 0.680, 0.844$, and 0.922, respectively.

Sample size: We assumed that N^+ and N^- are equal, with $N^+, N^- \in \{100, 500, 1000\}$.

For each scenario given by AUROC and sample size, we calibrated and solved TTP*-Q and TTP*-DR 100 times under the constraints $\gamma^{fp} = 0.10$ and $\gamma^{fn} = 0.05$.

Since this simulation analysis utilized Beta distributions with known parameters, we were able to compute the true optimal solution to TTP* for each simulation scenario. We then compared these true optimal solutions to the solutions obtained by solving TTP*-Q and TTP*-DR. Specifically, we compared these methods on the basis of optimality and feasibility. To measure optimality, we computed the optimality gap, which we define by

$$1 - \frac{\mathbb{E}[se(u, \xi^+)] + \mathbb{E}[sp(l, \xi^-)]}{\mathbb{E}[se(u^*, \xi^+)] + \mathbb{E}[sp(l^*, \xi^-)]}, \tag{15}$$

where (u, l) denotes a candidate solution from TTP*-Q and (u^*, l^*) denotes the true optimal solution. From (15), it can be seen that a set of thresholds (u, l) achieving a positive optimality gap near 0 indicates near-optimal performance and near 1 indicates poor performance. By definition, optimality gaps below 0 imply that the thresholds are infeasible to the true underlying distribution.

To measure feasibility, we computed the maximum constraint violation, which is defined by

$$\max \left\{ \mathbb{E}[fp(u, \xi^-)] - \gamma^{fp}, \mathbb{E}[fn(l, \xi^-)] - \gamma^{fn} \right\}. \tag{16}$$

From (16), it can be seen candidate thresholds (u, l) are feasible for the true problem only if their maximum constraint violations are non-positive. For both TTP*-Q and TTP*-DR, we plot joint

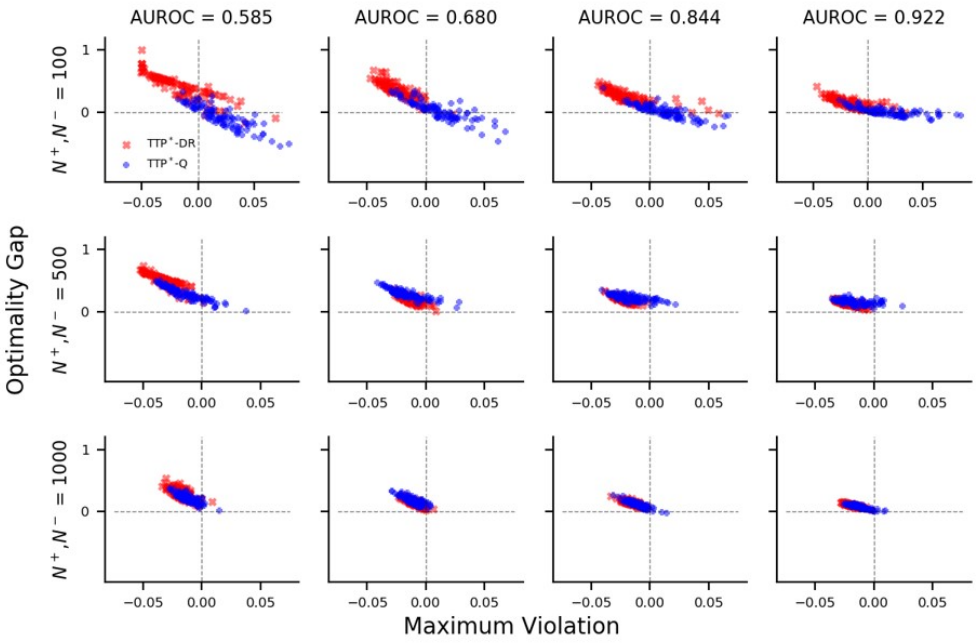

Figure 1: Distribution of Optimality Gap and Maximum Constraint Violation under varying quality of risk estimation model (AUROC) and sample size (N^+, N^-).

distribution of optimality gap and maximum constraint violation in Figure 1.

Several insights can be drawn from this analysis. For TTP*-Q, we find that its optimality and feasibility vary greatly based on sample size and AUROC. Specifically, TTP*-Q has low feasibility when the sample size is small (i.e., 17%-34% of solutions are feasible when $N^+, N^- = 100$). However, feasibility greatly improves when the sample size reaches 500 (i.e., 78%-87% of solutions are feasible) and 1000 (i.e., 93%-95% of solutions are feasible). For fixed sample sizes, we find that feasibility is the lowest when the AUROC is 0.585 and rises as AUROC increases to 0.680. However, additional increases in AUROC do not result in additional improvements in feasibility. Additionally, we find that the optimality gap (regardless of whether it is conditional on being feasible or not) decreases as the AUROC increases. On the other hand, increasing sample size does not seem to provide a noticeable decrease in the optimality gap.

The feasibility of TTP*-DR is largely driven by sample size. Specifically, 76%-96% of solutions are feasible when $N^+, N^- = 100$ and the feasibility rises between 96%-100% once the sample size increases to 500 and 1000. AUROC does not have a discernible effect on feasibility but much like for TTP*-Q, the optimality gap for TTP*-DR decreases as AUROC increases. When AUROC is

at least 0.680, the optimality gap decreases as the sample size increases.

Comparing the two solution methods, TTP*-DR achieves a greater level of feasibility than TTP*-Q in every combination of AUROC and sample size tested. Furthermore, the difference in feasibility between the two approaches is greatest when the sample size is small. Conditional on being feasible, TTP*-Q achieves a smaller optimality gap than TTP*-DR when the sample size is small and the AUROC is low. However, as the sample sizes increase and the AUROC increases, results are mixed in terms of optimality gap. However, the two methods appear to achieve similar optimality gaps as the AUROC and sample size increase.

Overall, sample size and AUROC play an important role in feasibility and optimality. Increasing sample size greatly improves feasibility, especially for TTP*-Q. Increasing AUROC decreases the optimality gap for both methods, and may slightly improve feasibility for TTP*-Q. TTP*-DR appears to be superior to TTP*-Q for small sample sizes due to its greater rate of feasibility. Since both methods perform similarly for large sample sizes and sufficiently high AUROC, computational time can be an important deciding factor in choosing which method to employ in practice. In an auxiliary analysis (see online supplementary material), we found that TTP*-DR is several orders of magnitude higher in computational time compared to TTP*-Q. While the overall computation time is not prohibitive for small sample sizes, TTP*-DR takes a median time of 74.20 seconds once the sample size reaches 5000 (compared to 0.04 seconds for TTP*-Q). Hence, TTP*-Q appears to be more practical for large sample sizes given the need to solve each model several times through the calibration procedure and their similarity in out-of-sample performance at large sample sizes. While our simulation study assumed that the underlying distributions for $\mathcal{P}^+$ and $\mathcal{P}^-$ were Beta distributions, we performed the same analysis under different distributional assumptions and found that similar trends follow (see online supplementary material). However, this auxiliary analysis does suggest that the variance of the distributions $\mathcal{P}^+$ and $\mathcal{P}^-$ play an important role in the feasibility of TTP*-Q.

5.2 Comparing One- and Two-threshold Classification Schemes

In this section, we use simulation to compare the accuracy of a one-threshold classification scheme and the two-threshold classification scheme determined by TTP*. Specifically, we obtain an optimal

one-threshold solution (denoted 1T*) by using sample average approximations to solve

$$(1\text{T}^*) \quad \max_t \lambda \mathbb{E}[se(t, \xi^+)] + (1-\lambda)\mathbb{E}[sp(t, \xi^-)], \tag{17}$$

where $\lambda \in (0,1)$ plays the same role as the λ used in the objective function of TTP. Notice that (17) is the same as solving an unconstrained version of TTP. Furthermore, the resulting threshold from solving (17) is akin to one chosen from the ROC curve, as is commonly done in practice (see Section 2). We evaluated both of these models based on classification accuracy, which can be interpreted as the probability of correctly classifying a randomly chosen patient. Specifically, we define accuracy as

$$q\Big(\mathbb{E}[se(u, \xi^+)] + \eta^+(u,l)\Big) + (1-q)\Big(\mathbb{E}[sp(u,l)|\xi^-] + \eta^-(u,l)\Big),$$

where q represents the proportion of true-positives in the overall population and the functions $\eta^+(u,l)$ and $\eta^-(u,l)$ denote the probabilities of correctly classifying a true-positive and true-negative patient, respectively, after they are deferred based on the thresholds u and l. Without the functions $\eta^+(u,l)$ and $\eta^-(u,l)$, 1T* would always have a higher accuracy since no proportion of the population is deferred. Note that $\eta^+(t,t) = 0$ and $\eta^-(t,t) = 0$ for 1T* since the upper and lower thresholds are equal.

Throughout this simulation analysis, we set

$$\eta^+(u,l) := \int_l^u \mathcal{L}^+_{k,\xi_0^+} f_{\mathcal{P}^+}(\xi^+) d\xi^+ \text{ and } \eta^-(u,l) := \int_l^u \mathcal{L}^-_{k,\xi_0^-} f_{\mathcal{P}^-}(\xi^-) d\xi^+,$$

where $f_{\mathcal{P}^+}$ and $f_{\mathcal{P}^-}$ are density functions based on the Beta distributions specified in Section 5.1 and the logistic functions

$$\mathcal{L}^+_{k,\xi_0} = \Big(1 + \exp\Big(-k(\xi^+ - \xi_0)\Big)\Big)^{-1} \text{ and } \mathcal{L}^-_{k,\xi_0} = \Big(1 + \exp\Big(-k(\xi_0 - \xi^-)\Big)\Big)^{-1}$$

are parameterized by scale $k \geq 0$ and centers $\xi_0^+, \xi_0^- \in [0,1]$. We use $\mathcal{L}^+_{k,\xi_0}$ (resp., $\mathcal{L}^-_{k,\xi_0}$) to approximate the likelihood of correctly classifying a deferred true-positive (resp., true-negative) patient given risk estimate ξ^+ (resp., ξ^-).

In our analysis, we solved TTP*-Q with $\gamma^{fp} = \gamma^{fn} = 0.10$ and 1T* with $\lambda = 0.5$ by sampling from each distribution specified in Section 5.1 with sample sizes $N^+ = N^- = 1000$. To compute accuracy, we set $q = 0.5$ and parameterized the functions L_{k,ξ_0^+} and L_{k,ξ_0^-} with fixed centers $\xi_0^+ = 0.45, \xi_0^- = 0.55$ and varying $k \in [0, 20]$. Higher values of k indicate greater likelihood of correctly classifying a patient after they are initially deferred.

The results of our simulation analysis are shown in Figure 2. In all cases, the accuracy of TTP*-Q increases quadratically in k, indicating that small improvements in the accuracy of post-defer classification will initially result in large gains of overall accuracy. However, there is an inflection point at which there are diminishing returns on the gains in overall accuracy. To this end, the inflection point appears to be increasing in the AUROC of the underlying risk estimation model. This result implies that the marginal value of increasing the accuracy of post-defer classification remains high for risk estimation models which are already accurate. Practically speaking, if the underlying risk estimation model is accurate, then increasing the accuracy of post-defer classification has high utility if the marginal cost of increasing post-defer accuracy is low (e.g., ordering low-cost lab tests).

Comparing TTP*-Q and 1T*, our results indicate that when few patients are correctly classified after they are first deferred (i.e., k is near 0), 1T* achieves a greater classification accuracy. However, as the likelihood of correctly classifying deferred patients increases, TTP*-Q begins to achieve a greater level of accuracy than 1T*. To this end, the minimum classification accuracy for deferred patients (i.e., k) needed such that TTP*-Q outperforms 1T* also increases in the quality of the underlying risk estimation model (i.e., AUROC) improves. This result indicates that if the underlying risk estimation model can already distinguish between true-positives and true-negatives with high accuracy, more effort is required in ensuring that deferred patients can actually be correctly identified after they are initially deferred. Practically speaking, a one-threshold classification scheme may be more useful when the underlying risk estimation model is already accurate and there is a high cost associated with increasing the accuracy of classifying patients who are deferred (e.g., ordering assessments which require expensive equipment). This analysis illustrates that the utility of TTP increases as the accuracy of a risk estimation model decreases. This result stems from the fact that the underlying advantage to using TTP is its ability to identify patients who cannot be

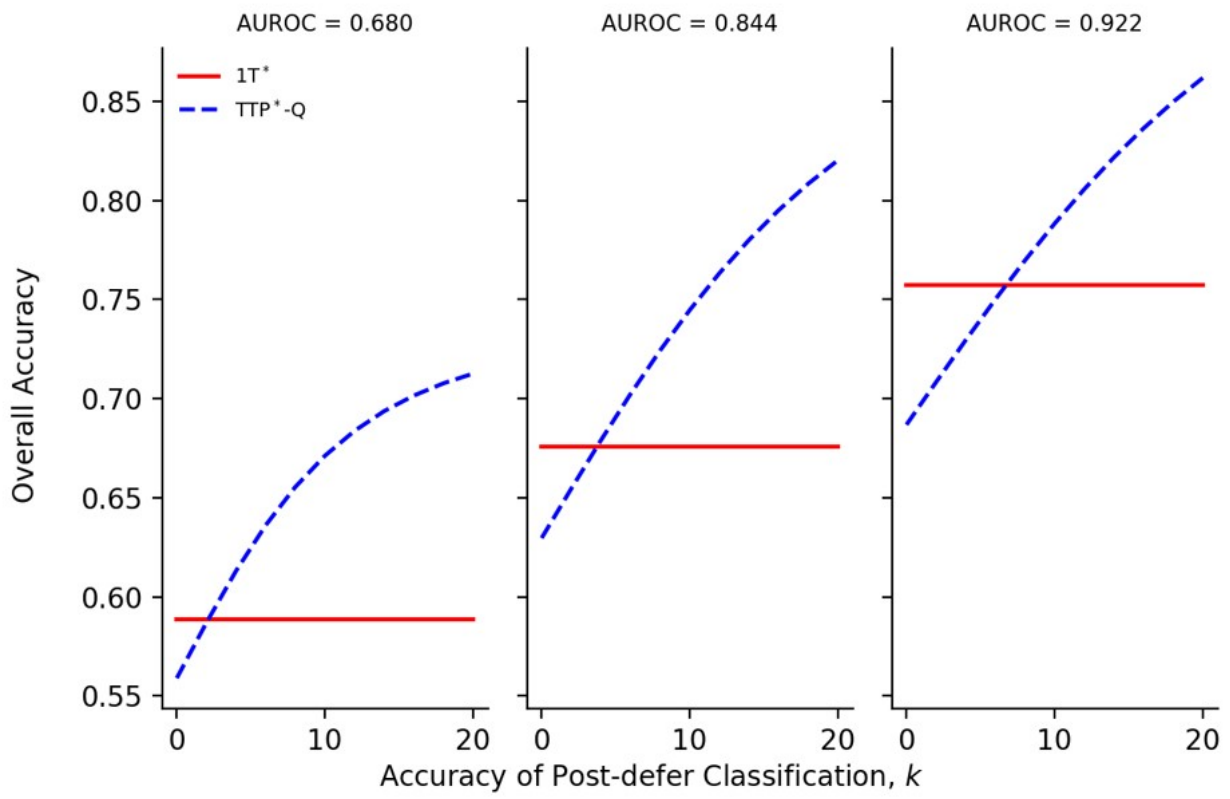

Figure 2: Comparison of overall accuracy between one-threshold (1T*) and two-threshold (TTP*-Q) classification schemes under varying quality of risk estimation model (AUROC) and post-defer classification accuracy (k).

accurately diagnosed using the risk estimation model. Therefore, these patients are better served by the expert judgment of clinicians.

6 Case Study: Acute Concussion Assessment

In this section, we present numerical results based on the application of TTP* to concussion assessment data. We first describe our data and choice of risk estimation model. Then, we define three efficiency measures and analyze the relationship between modeling parameters and these efficiency measures. Finally, we compare the performance of TTP* to (1) an optimized one-threshold solution which is commonly used to determine decision thresholds for risk estimation models and (2) a normative value comparison method which is commonly used in acute concussion assessment.

6.1 Concussion Assessment Data

Concussions, a type of traumatic brain injury, are an emerging public health issue. Accurate diagnosis is an important part of the injury management process as delayed removal from play following concussion is known to lengthen recovery time and lead to more severe post-concussion symptoms (Asken et al., 2018). Furthermore, while the exact relationship is still unclear, concussions are postulated to develop into more serious long-term consequences which may include cognitive impairment, depression, and neurodegenerative disease (Guskiewicz et al., 2005, 2007; Kerr et al., 2012, 2014; McCrory et al., 2017). Our numerical analysis uses multi-center longitudinal data provided by the Concussion Assessment, Research, and Education (CARE) Consortium (Broglio et al., 2017). To our knowledge, this nationwide study, with 29 National Collegiate Athletic Association universities and military service academies, is the first of its scale for concussion assessment.

Athletes diagnosed with concussion are assessed at five post-injury timepoints, beginning from within 6 hours of injury (< 6 hours) and 6 months (± 1 week) after the athlete has been cleared to return to play. We focus on evaluations from < 6 hours, which we denote as "concussion" (i.e., true-positive), and the first time at which the athlete is cleared to return to play, which we denote "non-concussion" (i.e., true-negative).

Data was received in two batches. Because some data were missing at the < 6 hours timepoint, the first batch (i.e., training data) contained 560 concussions and 707 non-concussions. Similarly, the second batch contained 539 concussions and 629 non-concussions. We used a randomly chosen 40% of the training data to form a risk estimation model and the remaining 60% to solve TTP* using each method in Section 4. We validated our thresholds and assessed out-of-sample performance using the second batch of data (i.e., validation data). We summarize this data in Table 2 for males and females in the training and validation set with respect to the Standard Assessment of Concussion (SAC), the Sport Concussion Assessment Tool (SCAT) symptom assessments, and the Balance Error Scoring System (BESS).

6.2 Risk Estimation Model

Multivariate logistic regression can be used to create a risk estimation model of the form

$$f(\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{x}^T\mathbf{b})},$$

Table 2: Description of Training and Validation Data

	Training Data							
	Male				**Female**			
	Concussion (n=361)		*Non-concussion* (n=423)		*Concussion* (n=199)		*Non-concussion* (n=284)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
SAC Total Score	25.90	3.12	27.87	1.85	26.74	2.42	28.00	1.60
SCAT Symptom Severity	27.33	20.88	0.43	1.55	31.67	20.02	0.93	2.46
SCAT Total Number of Symptoms	10.53	5.50	0.33	1.17	11.81	5.02	0.67	1.65
BESS Total Score	16.58	8.88	10.92	5.93	14.86	7.90	9.86	5.78
	Validation Data							
	Male				**Female**			
	Concussion (n=332)		*Non-concussion* (n=340)		*Concussion* (n=207)		*Non-concussion* (n=289)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
SAC Total Score	25.52	3.52	28.07	1.91	26.38	2.86	28.31	1.62
SCAT Symptom Severity	27.42	21.54	0.29	0.92	30.31	21.2	0.69	2.24
SCAT Total Number of Symptoms	10.62	5.75	0.22	0.70	11.03	5.17	0.51	1.46
BESS Total Score	17.86	8.97	11.12	5.02	16.99	8.67	10.76	5.73

SAC = Standard Assessment of Concussion — a neurocognitive assessment; SCAT = Sport Concussion Assessment Tool — a standard concussion assessment battery including a 22-item symptom checklist; BESS = Balance Error Scoring System — a postural stability assessment; SD = standard deviation

where $f(\mathbf{x})$ is a risk estimate, $\mathbf{x}$ is a vector of patient characteristics, and $\mathbf{b}$ is a vector of corresponding coefficients. We used the multivariate logistic regression model developed by Garcia et al. (2018) for acute concussion assessment, in which the relevant patient characteristics include sex, whether the injury was reported immediately, whether the athlete was removed from play immediately, and scores from the SAC, SCAT symptom assessment, and BESS.

6.3 Example Solution and Post-hoc Analysis

We solved TTP*-Q and TTP*-DR using the training data and evaluated their optimal solutions against the validation data. For various choices of γ^{fp} and γ^{fn}, we found that TTP*-Q generally remains feasible. In some cases (e.g,. $\gamma^{fp} = 0.01$ and $\gamma^{fn} = 0.03$), TTP*-Q violated one of its constraints — though the magnitude of this violation was generally small (i.e., less than 0.006). In contrast, TTP*-DR was feasible against the validation data in every parameter combination tested. Figure 3 shows an example optimal solution from TTP*-Q obtained by setting $\gamma^{fp} = 0.028$ and $\gamma^{fn} = 0.02$. In this example, u^* captures many of the true-positives (sensitivity= 0.91) while maintaining a low false-positive rate (false-positive= 0.016). On the other hand, l^* captures fewer true-negatives (specificity= 0.74) but also maintains a low false-negative rate (false-negative= 0.015). Only 16.7% of the validation data fell between u^* and l^*. That is, only 16.7% of those

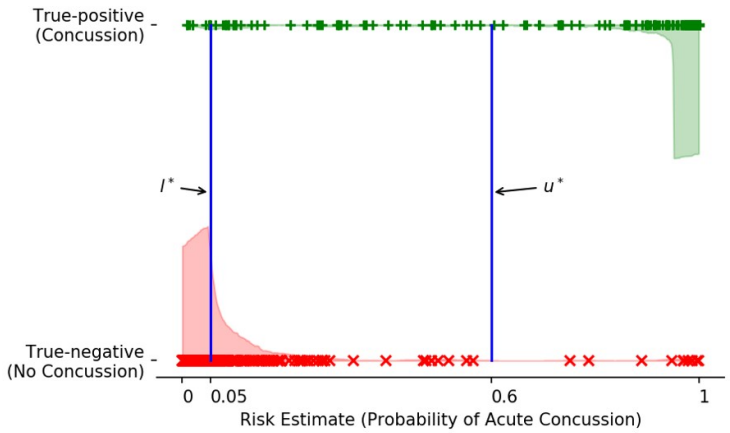

Figure 3: Optimal upper threshold (u^*) and lower threshold (l^*) when solving TTP*-Q with $\gamma^{fp} = 0.028$ and $\gamma^{fn} = 0.020$. Risk estimates $> u^*$ are classified as true-positives while risk estimates $< l^*$ are classified as true-negatives. Diagnosis decisions are deferred for risk estimates between u^* and l^*. True positives (+) and true-negatives (x) from validation data are shown with estimated kernel densities.

undergoing diagnosis would result in deferred diagnosis.

In practice, it may be of interest to perform post-hoc analysis on athletes who were ultimately deferred. For example, in Table 3, we characterize the deferred athletes based on their risk estimates and clinical assessment variables. We also compare them to athletes who were correctly diagnosed. It can be seen that risk estimates were lower for deferred athletes with concussion compared to those that were correctly diagnosed. Likewise, risk estimates were higher for deferred athletes without concussion compared to those who were correctly diagnosed. We also identified statistically significant differences in the SAC total score, SCAT symptom severity, and SCAT total number of symptoms between deferred athletes and their correctly diagnosed counterparts. Practically speaking, these results confirm just how different the "easy" and the "hard" cases are. Furthermore, these characterizations can be used to inform future clinical decisions. For example, deferred athletes with higher symptom presentation are more likely to have concussion than those with lower symptom presentation. However, as these clinical variables were already included in our risk estimation model, it may be of greater value to consider additional assessments (e.g., vestibular/ocular-motor screening) for the group of deferred athletes.

Table 3: Comparison of Athletes Who Were Correctly Diagnosed and Deferred Under Example TTP*-Q Solution

True Outcome	**Concussion**				**No Concussion**			
Diagnosis Decision	*Positive*		*Defer*		*Negative*		*Defer*	
	n=490		n=41		n=465		n=154	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Risk Estimate†,‡	0.98	0.06	0.28	0.16	0.01	0.01	0.14	0.10
SAC Total Score †,‡	25.71	3.36	27.24	2.25	28.26	1.73	27.97	1.85
BESS Total Score†,‡	18.02	8.92	13.00	6.19	10.60	5.10	11.88	5.86
SCAT Total Symptom Severity†,‡	31.04	20.83	3.85	3.35	0.14	0.57	0.88	1.68
SCAT Total Number of Symptoms†,‡	11.64	5.01	2.37	1.84	0.11	0.42	0.69	1.27

†Mean value is significantly different ($p < 0.05$) between concussions with positive and deferred diagnosis decisions using two-sample Student's t-test; ‡Mean value is significantly different ($p < 0.05$) between non-concussions with negative and deferred diagnosis decisions using two-sample Student's t-test; SAC = Standard Assessment of Concussion — a neurocognitive assessment; SCAT = Sport Concussion Assessment Tool — a standard concussion assessment battery including a 22-item symptom checklist; BESS = Balance Error Scoring System — a postural stability assessment; SD = standard deviation

6.4 Analyzing The Efficiency of Two-threshold Solutions

In practice, decision thresholds which defer too many diagnosis decisions can be problematic. For example, if the necessary treatment following a diagnosis decision must be performed in a timely manner, then deferring too many decisions can cause a delay in receiving that treatment. Additionally, if the action following a deferred decision requires valuable resources (e.g., performing additional diagnostic tests), then deferring too many decisions could put potentially strain those resources. Thus, we evaluated TTP*-Q for acute concussion assessment based on three efficiency measures:

$$e_1 = \frac{p^+(u,l)}{p^D(u,l)}, e_2 = \frac{p^-(u,l)}{p^D(u,l)}, e_3 = \frac{p^+(u,l)}{p^-(u,l)},$$

where the $p^+(u,l)$, $p^-(u,l)$ and $p^D(u,l)$ are defined in Section 3.5. The measure e_1 is the rate of correct classifications per deferred decision, e_2 is the rate of misclassifications per deferred decision, and e_3 is the rate of correct classifications per misclassification. We computed each of these measures for various parameter combinations. We excluded any results with single-threshold solutions, since we would have $p^D(u,l) = 0$ and thus, e_1 and e_2 cannot be computed. We also varied the proportion of true-positives in the population undergoing acute concussion assessment, q.

From this efficiency analysis, we found that when there is a high proportion of true-positives in the testing population (i.e., q is high), parameter combinations which allow for higher false-positive rates and ultimately end up with high sensitivity result in greater efficiency in terms of correct

classifications per misclassification (i.e., $e1$) and deferred decision (i.e., $e3$). Similarly, when most of the testing population is composed of true-negatives (i.e., q is low), parameter combinations which allow for higher false-negative rates result in greater specificity. These parameter combinations are also more efficient in terms of $e1$ and $e3$. Likewise, parameter combinations which have low false-positive rates tend to have lower misclassification rates per deferred decision (i.e., $e2$) when there are fewer true-negatives in the population (i.e., q is lower). Parameter combinations with low false-negative rates are more efficient in terms of $e2$ when there are more true-positives in the population (i.e., q is higher).

We also consider the tradeoff between e_1 and e_3 since decision makers may aim to correctly classify as many patients as possible, while minimizing deferred and incorrect classifications. We plot the values of e_1 and e_3 at different values of q for various combinations of γ^{fp} and γ^{fn} in Figure 4.

Table 4: Selected Pareto Optimal Parameter Combinations For Efficiency Analysis

	Parameters		*Opt Solution*		*Out-of-sample Performance*			
Label	γ^{fp}	γ^{fn}	u^*	l^*	**Sensitivity**	**Specificity**	**False-positive**	**False-negative**
0	0.01	0.01	0.94	0.02	0.83	0.57	0.01	0.01
1	0.01	0.02	0.94	0.05	0.83	0.74	0.01	0.01
2	0.01	0.03	0.94	0.12	0.83	0.89	0.01	0.04
3	0.01	0.04	0.94	0.22	0.83	0.95	0.01	0.05
12	0.02	0.01	0.73	0.02	0.89	0.57	0.02	0.01
13	0.02	0.02	0.73	0.05	0.89	0.74	0.02	0.01
14	0.02	0.03	0.73	0.12	0.89	0.89	0.02	0.04
15	0.02	0.04	0.73	0.22	0.89	0.95	0.02	0.05
16	0.02	0.05	0.73	0.31	0.89	0.97	0.02	0.06
21	0.03	0.02	0.54	0.05	0.91	0.74	0.02	0.01
23	0.03	0.04	0.54	0.22	0.91	0.95	0.02	0.05
24	0.03	0.05	0.54	0.31	0.91	0.97	0.02	0.06
25	0.03	0.06	0.54	0.38	0.91	0.97	0.02	0.07
29	0.04	0.02	0.40	0.05	0.93	0.74	0.03	0.01
30	0.04	0.03	0.40	0.12	0.93	0.89	0.03	0.04
31	0.04	0.04	0.40	0.22	0.93	0.95	0.03	0.05
32	0.04	0.05	0.40	0.31	0.93	0.97	0.03	0.06
33	0.04	0.06	0.40	0.38	0.93	0.97	0.03	0.07
35	0.05	0.02	0.32	0.05	0.94	0.74	0.03	0.01
36	0.05	0.03	0.32	0.12	0.94	0.89	0.03	0.04
39	0.06	0.02	0.27	0.05	0.95	0.74	0.04	0.01
41	0.06	0.04	0.27	0.22	0.95	0.95	0.04	0.05

Certain parameter combinations are only Pareto optimal for one value of q, suggesting the importance of choosing the parameters γ^{fp} and γ^{fn} based on the proportion of true-positives in the testing population rather than the underlying population. In contrast, a few parameter combina-

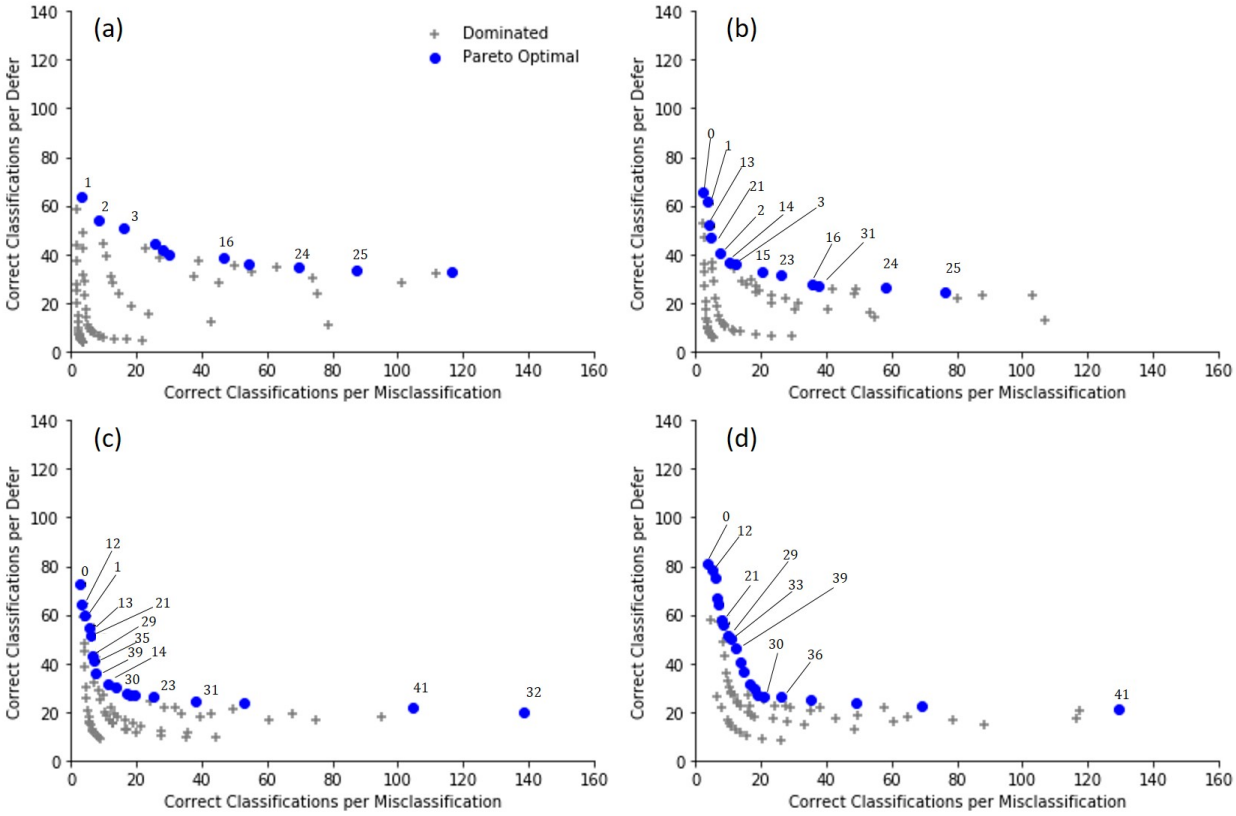

Figure 4: (a) $q = 0.2$. (b) $q = 0.4$. (c) $q = 0.6$. (d) $q = 0.8$. Rate of correct classifications per deferred decision (e_1) vs. Rate of correct classifications per misclassification (e_3) for TTP*-Q under various parameter combination and proportion of true-positives, q. Pareto optimal (circles) and dominated (crosses) parameter combinations are shown. Labels are shown only for parameter combinations which are Pareto optimal for at least two values of q. These labels correspond to parameter combinations in Table 4.

tions were relatively unaffected by changes in q, i.e., parameter combinations which remained on the Pareto optimal frontier for at least 2 different values of q. These parameter combinations are summarized in Table 4. From this analysis, we find that parameter combinations with low values of γ^{fp} (e.g., labels 1, 2, 3) tend to do well when the proportion of true-positives is low, i.e., $q = 0.2$ and $q = 0.4$. The opposite seems to hold as well — parameter combinations with low γ^{fn} (e.g., labels 12, 30, 39) tend to do well when the proportion of true-positives is high, i.e., $q = 0.6$ and $q = 0.8$. Some parameter combinations performed well for at least 3 values of q (e.g., labels 0, 1, 21). These parameter combinations strike a balance between γ^{fp} and γ^{fn}, though this pattern may not always hold (e.g., label 13). We also find that more conservative parameter combinations, i.e., low γ^{fp} and γ^{fn} such as labels 0 and 1, will achieve more correct classifications for each misclassification. In

contrast, less conservative parameter combinations, i.e., higher values of γ^{fp} and γ^{fn} such as labels 33 and 41, will have more correct classifications for each deferred decision.

6.5 Comparing TTP* Performance To Existing Methods

We evaluated the model performance for TTP*-Q and TTP*-DR under different parameter combinations, i.e., ϕ, γ^{fp}, and γ^{fn}. We considered the following performance criteria:

(i) False-positive vs. false-negative rate satisfying minimum levels of sensitivity,

(ii) Sensitivity vs. false-positive rate satisfying maximum levels of false-negative rate, and

(iii) Probability of correct classification vs. probability of misclassification,

where the probabilities in criterion (iii) correspond to the definitions of $p^+(u,l)$ and $p^-(u,l)$ in Section 3.5. We computed $p^+(u,l)$ and $p^-(u,l)$ at different values of q to evaluate criterion (iii). Furthermore, we compared the performance of these two-threshold solutions against two baseline cases: an optimal one-threshold solution (1T*) and normative value comparison.

Optimal One-Threshold Solution: Applying sample average approximation on the training set of data (i.e., $N^+ = 560$ and $N^- = 706$), we estimated the solution to 1T* as defined in (17). We compare TTP to this threshold since the resulting threshold is akin to one chosen from the ROC curve as is commonly done in determining decision thresholds (see Section 2).

Normative Value Comparison: While the clinical examination remains to be the golden standard for concussion diagnosis, clinicians commonly use Normative Value Comparison (NC) as an initial acute concussion screening tool (Broglio et al., 2008; Chin et al., 2016; Hänninen et al., 2016; Zimmer et al., 2015). In NC, the clinician compares the performance of an athlete suspected of concussion on various standard assessments, e.g., the SAC, SCAT symptom checklist, and BESS. If the athlete's performance is too many standard deviations (SD) away from a given normative (i.e., mean) value, then the athlete is treated as concussed and otherwise non-concussed. Using the normative values for the training data in Table 2, we analyzed a number of one-threshold and two-threshold NC schemes by varying the number of SD away from the normative value (from 0.5 to 3 in increments of 0.25) on the SAC, SCAT symptom checklist, and the BESS. For example, we considered the case where an athlete is assessed as non-concussed if his or her performance on at

least one assessment is within 1 SD of the normative value, concussed if beyond 2 SD, or deferred otherwise.

The results of our analysis for criteria (i) and (ii) are presented in Figure 5. Only Pareto optimal parameter combinations for TTP*-Q and TTP*-DR are shown. From Figure 5(a), we see that, for TTP*-Q and TTP*-DR, at least one of the Pareto optimal parameter combinations achieves a lower false-positive and false-negative rate than 1T* and NC while maintaining the same minimum level of sensitivity. However, when the sensitivity must be ≥ 0.96, TTP*-DR is too conservative and cannot produce a solution which satisfies this requirement. Furthermore, no solution from 1T* or NC is able to achieve at least 0.96 sensitivity while maintaining false-positive and false-negative rates below 0.15.

In Figure 5(b), we find that different parameter combinations for TTP*-Q and TTP*-DR are able to achieve similar or better levels of discrimination (i.e., high sensitivity and low false-positive rates) compared to 1T* and NC while keeping lower false-negative rates. Specifically, we point out that when the false-negative rate is constrained to be ≤ 0.01, no 1T* solution is able to achieve a similar performance to TTP*-Q and TTP*-DR. However, several TTP*-Q and TTP*-DR solutions still satisfy the maximum false-negative limit. We also point out that while NC is able to produce a solution with good discrimination at low false-negative rates, it is dominated by the TTP* solutions. That is, TTP*-Q and TTP*-DR are able to achieve similar levels of sensitivity and false-negative rates at lower false-positive rates.

The results for criterion (iii) are presented in Figure 6. We find that 1T* is able to match the performance achieved by TTP*-Q and TTP*-DR for one parameter combination. However, TTP* offers solutions with lower probability of misclassification compared to 1T*. Similarly to the analysis of criterion (ii), we find that the TTP* solutions dominate NC in terms of performance. That is, for the same probability of misclassification, each of the TTP* methods achieve a higher probability of correct classification. This finding demonstrates the utility of two-threshold solutions in combination with risk estimation models compared to traditional methods.

Overall, these results suggest that TTP* outperforms both 1T* and NC when comparing across multiple criteria. In particular, the two-threshold solutions are able to maintain high diagnostic

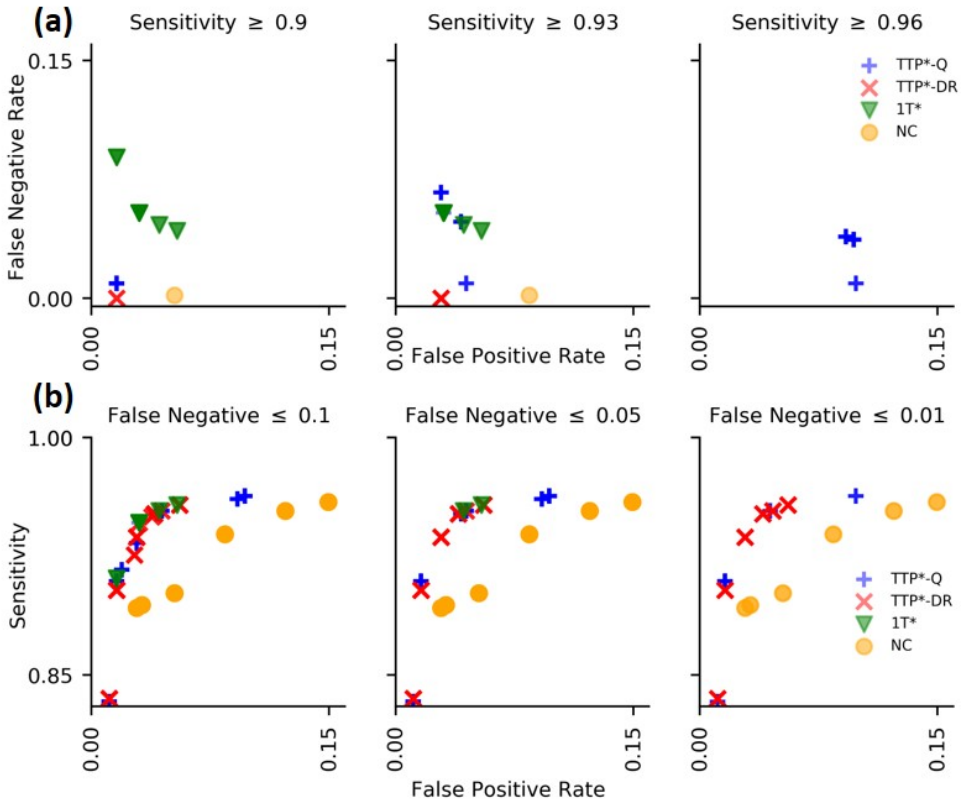

Figure 5: (a) false-positive vs. false-negative rate satisfying minimum levels of sensitivity. (b) sensitivity vs. false-positive rate satisfying maximum levels of false-negative rate. Comparison of Pareto optimal parameter combinations for TTP*-Q, TTP*-DR, optimal single-threshold solutions (1T*), and normative value comparison (NC).

accuracy while maintaining low levels of false-positive and false-negative rates.

7 Conclusion

In conclusion, we have designed and analyzed a data-driven method for optimizing an upper and lower threshold for diagnosis decisions. This method (1) reflects the decision maker's risk attitude, (2) determines data-driven thresholds based on risk estimates specific to a certain risk estimation model and target population, and (3) identifies a range of risk estimates for which the risk estimation model faces elevated false-positive and false-negative rates. Through data-driven solution approaches, we avoid needing assumptions on the distribution of risk estimates.

7.1 Managerial Insights

Our analysis provides insight into the nature of our modeling framework and its application to acute concussion assessment.

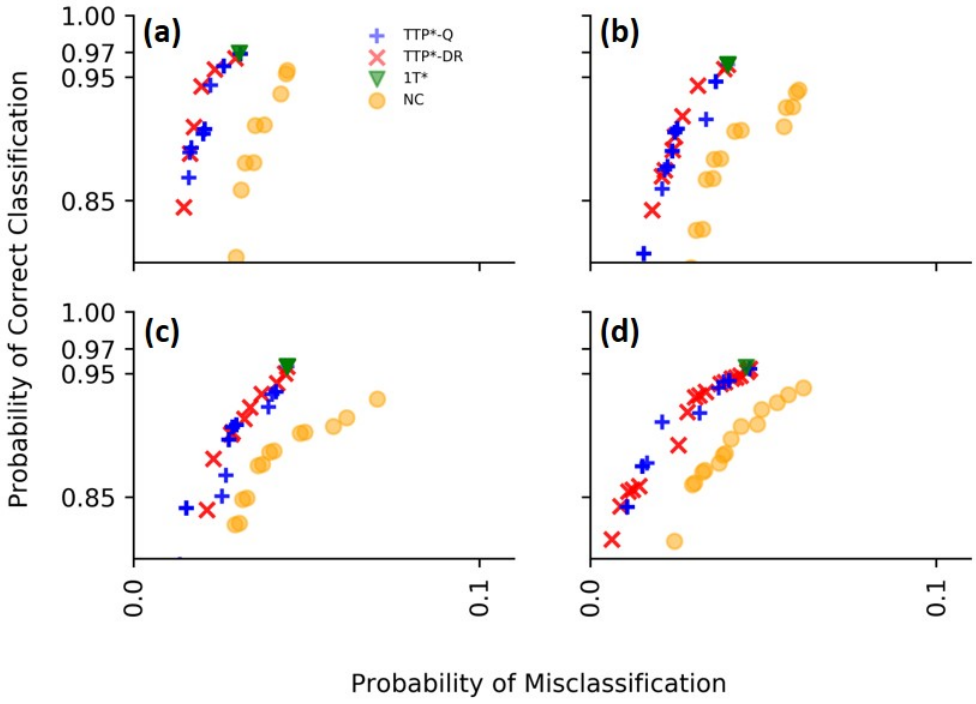

Figure 6: (a) $q = 0.2$. (b) $q = 0.4$. (c) $q = 0.6$. (d) $q = 0.8$. Probability of correct classification vs. probability of misclassification for varying proportions of true-positives, q, and Pareto optimal parameter combinations of TTP*-Q , TTP*-DR, optimal single-threshold solutions (1T*), and normative value comparison (NC).

When should two diagnosis decision thresholds be used instead of one? Since the advantage of using two decision thresholds is in its ability to defer patients, the benefit of using two thresholds (instead of one) depends on the risk estimation model's accuracy and the accuracy of the assessments which follow a deferred diagnosis decision. Specifically, if a risk estimation model is inaccurate and patients who are deferred can eventually be diagnosed with sufficient accuracy, then two decision thresholds should be used. We also find that two diagnosis decision thresholds are needed (e.g., possible, probable, or definite concussion) when it is important to restrict both the false-positive and false-negative rates. However, if it is only important to limit either the false-positive or false-negative rate, then a single decision threshold (e.g., concussion or no concussion) suffices. Fortunately, for acute concussion assessment, the added conservatism by restricting misclassification rates does not drastically reduce sensitivity and specificity. Furthermore, two thresholds should be used if the cost associated with misdiagnosis is sufficiently high compared to a deferred diagnosis decision. Alternatively, if the majority of deferred decisions consists of pa-

tients who ultimately would have been misdiagnosed or if very few deferred decisions are ultimately misdiagnosed, then two decision thresholds should be used instead of one.

How does one choose parameters which generate good thresholds? For TTP, the decision-maker only needs to consider the constraints on false-positive and false-negative rates, since the relative importance of sensitivity to specificity is trumped by these constraints in two-threshold solutions. To this end, the degree to which false-positive and false-negative rates are constrained should reflect the relative "cost" associated with each type of misdiagnosis. Furthermore, the false-negative rate should be further constrained if the proportion of true-positives in the testing population is high. However, if the proportion of true-positives is low, then the false-positive rate should be constrained instead. In the case of acute concussion assessment, our efficiency analysis demonstrated that balanced constraints (i.e., γ^{fp} and γ^{fn} are both low and nearly equal) can be robust for different proportions of true-positives in the testing population.

Given the similarities between TTP and MTTP, similar principles can be applied in choosing parameters for MTTP. However, choosing parameters for OTP can be challenging. In our analysis of OTP, we found that the parameters $\gamma^{fp}_{k,j}$ and $\gamma^{fn}_{k,j}$ must satisfy certain conditions in order for OTP to be feasible. To this end, when using the sensitivity-based formulation of OTP (i.e., the objective function is given by (8a)), one can satisfy these conditions by setting $\gamma^{fn}_{k,j} = 1$ for all $j > k$ and focus on choosing appropriate values for $\gamma^{fp}_{k,j}$ for all $j \leq k$. In our analysis, we found that the performance of different parameter choices were sensitive to the underlying quality of the risk estimation model. For example, when the risk estimation model was poor, choosing high values of $\gamma^{fp}_{k,j}$ (e.g., 0.11-0.13) was imperative to attaining high accuracy and low mean squared error — regardless of sample size. On the other hand, parameter sets which had low values for $\gamma^{fp}_{k,j}$ (e.g., 0.01-0.05) achieved high accuracy at low mean squared error when the risk estimation models were high quality. Given these guiding principles, one can construct a set of viable parameters and then use methods such as cross-validation to evaluate the performance of different parameter choices.

Which data-driven solution method should be used? In our simulation analysis, we analyzed the performance of each solution method under different sample sizes and the quality of the underlying risk estimation model. If the available sample size is small and the underlying risk estimation model is poor, then TTP*-DR should be applied if ensuring feasibility is important to

the application. Additionally, TTP*-DR can be useful when the available sample of data is suspected to be quite different from the distribution generated by the population. However, as sample sizes get larger, the computational burden of TTP*-DR increases to the point where it may not be practical to implement. To this end, TTP*-Q can be executed much more quickly at large sample sizes and performs similarly to TTP*-DR. Hence, TTP*-Q should be used once the sample size is large enough and the underlying quality of the risk estimation model is reasonably accurate.

How does this methodology compare to existing methods? For acute concussion assessment, combining two-threshold solutions with risk estimation models allows for both greater diagnostic accuracy and lower likelihood of misdiagnoses. In particular, given the same risk estimation model, one-threshold solutions are adept at providing higher sensitivity and specificity but at the cost of greater misclassification rates compared to two-threshold solutions. Compared to common clinical practice such as normative value comparison, combining risk estimation models with data-driven thresholds can drastically improve diagnostic accuracy. Therefore, this modeling framework should be transformed into a clinical decision aid to facilitate its implementation in practice.

In our extensions of TTP to multi-class classification, we found that OTP outperforms a common method for determining decision thresholds in ordinal classification. Specifically, OTP attained higher accuracy and lower mean squared error than a commonly used equidistant threshold method. This improvement in performance can be owed to the fact that OTP has greater flexibility in optimizing thresholds given that it does not require thresholds to be equidistant. However, this improved performance comes at a cost of requiring more parameters. To this end, we have identified some principles to guide the choice of parameters for OTP.

How does this work impact diagnosis decisions in clinical practice? It is natural for clinicians to have varying levels of certainty in their diagnosis decisions based on their initial evaluation of a patient. For instance, consider concussion diagnosis, for which no "gold standard" objective test currently exists, common concussion symptoms are sometimes slow to evolve and not necessarily specific to concussion, and clinical presentation of concussion may vary largely from one patient to the next (Kutcher and Giza, 2014). While the clinical exam remains the standard in this field, our research can provide valuable decision support to clinical judgment by quantifying this uncertainty with real data. Our modeling framework provides a more objective and data-driven way to

guide risk-based diagnosis decisions by identifying "easy" cases which can be diagnosed immediately and "hard" cases which may require further evaluation before diagnosing. We believe that our modeling framework permits clinicians to blend their expertise with quantitative evidence; while it does not recommend a next step for patients who are deferred, it has the flexibility to incorporate clinical judgment when necessary. Directing the expertise of clinicians to those patients who most need it can have a great impact in improving patient care while reducing unnecessary workload for clinicians. To facilitate the use of TTP and its extensions in clinical practice, the models developed in this research may initially undergo a pilot in clinic. After sufficient validation, it may eventually be incorporated into a mobile application or integrated with electronic health records. For example, MTTP can be paired with electronic health records to alert clinicians when patients may be at relatively high risk for developing any number of new conditions. By supplementing clinical decision-making, this research has the potential to improve diagnosis decisions, patient care, and health outcomes.

7.2 Limitations and Future Work

This research can be extended to address limitations in our current modeling framework. First, this work only determines two thresholds but some risk classifications may fall into many more categories. Consider the diagnosis of pulmonary hypertension, where patients may be classified into 1 of 5 groups (Galie et al., 2009). In this case, 2 thresholds may not be enough. While we have shown that TTP can be extended or modified to fit multi-label and ordinal classification frameworks, the extension to other types of multi-class classification (e.g., one-vs-all classification) is not so straightforward due to the potential for conflicting results.

The impact of time has been absent in our model formulations and analyses. However, in time-critical applications, it is important to consider whether deferring decisions is still a plausible action to take. Therefore, our modeling framework can be extended to consider the impact that this additional factor would have on our results.

Throughout this work, we assume a fixed risk estimation model. However, one may wish to jointly optimize a risk estimation model and its accompanying decision thresholds. Particularly, when the same data is used to formulate the risk estimation model and the thresholds, one may need to optimize the proportion of data used for the risk estimation model and the threshold problem.

In practice, differences between sample data and the true population distribution can act as barriers to implementing risk estimation models in practice. While our distributionally robust formulation is useful for dealing with such situations, one may obtain decision thresholds which are too conservative to be practically useful. Future research can investigate the utility of different ambiguity sets for such scenarios.

Finally, this model was motivated by medical diagnosis problems, but may be applied to other application domains where the determination of a dichotomous outcome is critical. Examples of such applications include bankruptcy prediction or natural disaster forecasting.

As medical data continues to become more readily available, methods which can incorporate these data to supplement medical decisions become increasingly relevant. While several extensions can be made to our work, the framework we have developed in this research provides a promising baseline for applying and understanding data-driven risk estimates in diagnosis decisions. Through future implementation and validation in clinical practice, it is our hope that this modeling framework ultimately leads to an improved quality of healthcare delivery.

8 Acknowledgments

The authors thank two anonymous referees and an associate editor, whose comments have improved the paper. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1256260.

This publication was made possible, in part, with support from the Grand Alliance Concussion Assessment, Research, and Education (CARE) Consortium, funded, in part, by the National Collegiate Athletic Association (NCAA) and the Department of Defense (DOD). The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office. This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Psychological Health and Traumatic Brain Injury Program under Award No. W81XWH-14-2-0151. Opinions, interpretations, conclusions and recommendations are those of the author(s) and are not necessarily endorsed by the Department of Defense (DHP funds).

The CARE Consortium Investigators are listed alphabetically by institution: April Marie (Reed)

Hoy, MS, ATC (Azusa Pacific University), Joseph B. Hazzard Jr., EdD, ATC (Bloomsburg University), Kelly A. Louise, PhD (California Lutheran University), Justus D. Ortega, PhD (Humboldt State University), Nicholas Port, PhD (Indiana University), Margot Putukian, MD (Princeton University), Gerald McGinty, DPT and Jonathan C. Jackson, PhD (United States Air Force Academy), Patrick G. O'Donnell, MHA (United States Coast Guard Academy), Kenneth L. Cameron, PhD, MPC, ATC (United States Military Academy), Christopher Giza, MD (University of California Los Angeles), Holly J. Benjamin, MD (University of Chicago), Thomas Buckley, EdD, ATC and Thomas W. Kaminski, PhD, ATC (University of Delaware), James R. Clugston, MD, MS (University of Florida), Julianne D. Schmidt, PhD, ATC (University of Georgia), Louis A. Feigenbaum, DPT, ATC (University of Miami), Steve P. Broglio, PhD, ATC (University of Michigan), Jason P. Mihalik, PhD, ATC (University of North Carolina), Jessica Dysart Miles, PhD, ATC (University of North Georgia), Scott Anderson, ATC (University of Oklahoma), Christina L. Master, MD (University of Pennsylvania), Anthony P. Kontos, PhD and Micky Collins, PhD (University of Pittsburgh), Jeffrey J. Bazarian, MD, MPH (University of Rochester), Sara P.D. Chrisman, MD, MPH (University of Washington), Alison Brooks, MD, MPH (University of Wisconsin), Steven Rowson, PhD (Virginia Tech), Christopher M. Miles, MD (Wake Forest University),

References

Ahsen, M. E., Ayvaci, M. U. S., and Raghunathan, S. (2019). When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis. *Information Systems Research*, 30(1):97–116.

Alagoz, O., Ayer, T., and Erenay, F. S. (2011). Operations Research Models for Cancer Screening. In *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Altman, D. G., Vergouwe, Y., Royston, P., and Moons, K. G. M. (2009). Prognosis and prognostic research: validating a prognostic model. *British Medical Journal*, 338(may28 1):b605–b605.

American Diabetes Society (2016). 2. Classification and Diagnosis of Diabetes. *Diabetes Care*, 39(Supplement 1):S13–S22.

Andelic, N., Hammergren, N., Bautz-Holter, E., Sveen, U., Brunborg, C., and Røe, C. (2009).

Functional outcome and health-related quality of life 10 years after moderate-to-severe traumatic brain injury. *Acta Neurologica Scandinavica*, 120(1):16–23.

Anderson, K. M. (1991). A nonproportional hazards Weibull accelerated failure time regression model. *Biometrics*, 47(1):281–8.

Asken, B. M., Bauer, R. M., Guskiewicz, K. M., McCrea, M. A., Schmidt, J. D., Giza, C. C., Snyder, A. R., Houck, Z. M., Kontos, A. P., McAllister, T. W., Broglio, S. P., Clugston, J. R., Anderson, S., Bazarian, J., Brooks, A., Buckley, T., Chrisman, S., Collins, M., DiFiori, J., Duma, S., Dykhuizen, B., Eckner, J. T., Feigenbaum, L., Hoy, A., Kelly, L., Langford, T. D., Lintner, L., McGinty, G., Mihalik, J., Miles, C., Ortega, J., Port, N., Putukian, M., Rowson, S., and Svoboda, S. (2018). Immediate Removal From Activity After Sport-Related Concussion Is Associated With Shorter Clinical Recovery and Less Severe Symptoms in Collegiate Student-Athletes. *The American Journal of Sports Medicine*, page 036354651875798.

Ayer, T., Alagoz, O., and Stout, N. K. (2012). OR Forum—A POMDP Approach to Personalize Mammography Screening Decisions. *Operations Research*, 60(5):1019–1034.

Ayer, T., Alagoz, O., Stout, N. K., and Burnside, E. S. (2016). Heterogeneity in Women's Adherence and Its Role in Optimal Breast Cancer Screening Policies. *Management Science*, 62(5):1339–1362.

Ayvaci, M. U. S., Ahsen, M. E., Raghunathan, S., and Gharibi, Z. (2017). Timing the Use of Breast Cancer Risk Information in Biopsy Decision-Making. *Production and Operations Management*, 26(7):1333–1358.

Ayvaci, M. U. S., Alagoz, O., and Burnside, E. S. (2012). The Effect of Budgetary Restrictions on Breast Cancer Diagnostic Decisions. *Manufacturing & Service Operations Management*, 14(4):600–617.

Barnett, C. L., Tomlins, S. A., Underwood, D. J., Wei, J. T., Morgan, T. M., Montie, J. E., and Denton, B. T. (2017). Two-Stage Biomarker Protocols for Improving the Precision of Early Detection of Prostate Cancer. *Medical Decision Making*, 37(7):815–826.

Bayati, M., Bhaskar, S., and Montanari, A. (2018). Statistical analysis of a low cost method for multiple disease prediction. *Statistical Methods in Medical Research*, 27(8):2312–2328.

Bertsimas, D., Silberholz, J., and Trikalinos, T. (2016). Optimal healthcare decision making under multiple mathematical models: application in prostate cancer screening. *Health Care Management Science*, pages 1–14.

Broglio, S. P., Ferrara, M. S., Sopiarz, K., and Kelly, M. S. (2008). Reliable Change of the Sensory Organization Test. *Clinical Journal of Sport Medicine*, 18(2):148–154.

Broglio, S. P., McCrea, M., McAllister, T., Harezlak, J., Katz, B., Hack, D., and Hainline, B. (2017). A National Study on the Effects of Concussion in Collegiate Athletes and US Military Service Academy Members: The NCAA–DoD Concussion Assessment, Research and Education (CARE) Consortium Structure and Methods. *Sports Medicine*, 47(7):1437–1451.

Chin, E. Y., Nelson, L. D., Barr, W. B., McCrory, P., and McCrea, M. A. (2016). Reliability and validity of the sport concussion assessment tool-3 (SCAT3) in high school and collegiate athletes. *American Journal of Sports Medicine*, 44(9):2276–2285.

Collins, M. W., Kontos, A. P., Reynolds, E., Murawski, C. D., and Fu, F. H. (2014). A comprehensive, targeted approach to the clinical care of athletes following sport-related concussion. *Knee Surgery, Sports Traumatology, Arthroscopy*, 22(2):235–246.

Daníelsson, J. and de Vries, C. G. (1997). Tail index and quantile estimation with very high frequency data. *Journal of Empirical Finance*, 4(2-3):241–257.

Degeling, K., Koffijberg, H., and IJzerman, M. J. (2017). A systematic review and checklist presenting the main challenges for health economic modeling in personalized medicine: towards implementing patient-level models. *Expert Review of Pharmacoeconomics and Outcomes Research*, 17(1):17–25.

Deneef, P. and Kent, D. L. (1993). Using Treatment-tradeoff Preferences to Select Diagnostic Strategies. *Medical Decision Making*, 13(2):126–132.

Deo, S., Rajaram, K., Rath, S., Karmarkar, U. S., and Goetz, M. B. (2015). Planning for HIV Screening, Testing, and Care at the Veterans Health Administration. *Operations Research*, 63(2):287–304.

Deo, S. and Sohoni, M. (2015). Optimal Decentralization of Early Infant Diagnosis of HIV in Resource-Limited Settings. *Manufacturing & Service Operations Management*, 17(2):191–207.

Ebell, M. (2010). AHRQ White Paper: Use of Clinical Decision Rules for Point-of-Care Decision Support. *Medical Decision Making*, 30(6):712–721.

El-Amine, H., Bish, E. K., and Bish, D. R. (2018). Robust Postdonation Blood Screening Under Prevalence Rate Uncertainty. *Operations Research*, 66(1):1–17.

Erenay, F. S., Alagoz, O., and Said, A. (2014). Optimizing Colonoscopy Screening for Colorectal Cancer Prevention and Surveillance. *Manufacturing & Service Operations Management*, 16(3):381–400.

Felder, S. and Mayrhofer, T. (2014). Risk Preferences: Consequences for Test and Treatment Thresholds and Optimal Cutoffs. *Medical Decision Making*, 34(1):33–41.

Galie, N., Hoeper, M. M., Humbert, M., Torbicki, A., Vachiery, J.-L., Barbera, J. A., Beghetti, M., Corris, P., Gaine, S., Gibbs, J. S., and Others (2009). Guidelines for the diagnosis and treatment of pulmonary hypertension. *European Heart Journal*, 30(20):2493–2537.

Garcia, G.-G. P., Broglio, S. P., Lavieri, M. S., McCrea, M., and McAllister, T. (2018). Quantifying the Value of Multidimensional Assessment Models for Acute Concussion: An Analysis of Data from the NCAA-DoD Care Consortium. *Sports Medicine*, 48(7):1739–1749.

Glasziou, P. and Hilden, J. (1986). Threshold Analysis of Decision Tables. *Medical Decision Making*, 6(3):161–168.

Godbole, S. and Sarawagi, S. (2004). Discriminative Methods for Multi-labeled Classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3056, pages 22–30.

Greiner, M., Pfeiffer, D., and Smith, R. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*, 45(1-2):23–41.

Greiner, M., Sohr, D., and Göbel, P. (1995). A modified ROC analysis for the selection of cut-off

values and the definition of intermediate results of serodiagnostic tests. *Journal of Immunological Methods*, 185(1):123–132.

Güneş, E. D., Örmeci, E. L., and Kunduzcu, D. (2015). Preventing and diagnosing colorectal cancer with a limited colonoscopy resource. *Production and Operations Management*, 24(1):1–20.

Guskiewicz, K. M., Marshall, S. W., Bailes, J., McCrea, M., Cantu, R. C., Randolph, C., and Jordan, B. D. (2005). Association between Recurrent Concussion and Late-Life Cognitive Impairment in Retired Professional Football Players. *Neurosurgery*, 57(4):719–726.

Guskiewicz, K. M., Marshall, S. W., Bailes, J., McCrea, M. A., Harding, H. P., Matthews, A., Register-Mihalik, J. K., and Cantu, R. C. (2007). Recurrent Concussion and Risk of Depression in Retired Professional Football Players. *Medicine & Science in Sports & Exercise*, 39(6):903–909.

Hänninen, T., Tuominen, M., Parkkari, J., Vartiainen, M., Öhman, J., Iverson, G. L., and Luoto, T. M. (2016). Sport concussion assessment tool – 3rd edition – normative reference values for professional ice hockey players. *Journal of Science and Medicine in Sport*, 19(8):636–641.

Harrell, F. E. and Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, 69(3):635–640.

Hartz, A., McKinney, W. P., Centor, R., Krieg, A., Simms, G., and Henck, S. (1986). Stochastic Thresholds. *Medical Decision Making*, 6(3):145–148.

Helm, J. E., Lavieri, M. S., Van Oyen, M. P., Stein, J. D., and Musch, D. C. (2015). Dynamic Forecasting and Control Algorithms of Glaucoma Progression for Clinician Decision Support. *Operations Research*, 63(5):979–999.

Humphreys, I., Wood, Phillips, C., and Macey (2013). The costs of traumatic brain injury: a literature review. *ClinicoEconomics and Outcomes Research*, 5(1):281.

Jacobson, S. H., Yu, G., and Jokela, J. A. (2016). A double-risk monitoring and movement restriction policy for Ebola entry screening at airports in the United States. *Preventive Medicine*, 88:33–38.

Jónasson, J. O., Deo, S., and Gallien, J. (2017). Improving HIV Early Infant Diagnosis Supply

Chains in Sub-Saharan Africa: Models and Application to Mozambique. *Operations Research*, 65(6):1479–1493.

Jund, J., Rabilloud, M., Wallon, M., and Ecochard, R. (2005). Methods to Estimate the Optimal Threshold for Normally or Log-Normally Distributed Biological Tests. *Medical Decision Making*, 25(4):406–415.

Kerr, Z. Y., Evenson, K. R., Rosamond, W. D., Mihalik, J. P., Guskiewicz, K. M., and Marshall, S. W. (2014). Association between concussion and mental health in former collegiate athletes. *Injury Epidemiology*, 1(1):28.

Kerr, Z. Y., Marshall, S. W., Harding, H. P., and Guskiewicz, K. M. (2012). Nine-Year Risk of Depression Diagnosis Increases With Increasing Self-Reported Concussions in Retired Professional Football Players. *The American Journal of Sports Medicine*, 40(10):2206–2212.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17.

Kutcher, J. S. and Giza, C. C. (2014). Sports Concussion Diagnosis and Management. *Continuum*, 20(6):1552–1569.

Lau, B. C., Kontos, A. P., Collins, M. W., Mucha, A., and Lovell, M. R. (2011). Which On-field Signs/Symptoms Predict Protracted Recovery From Sport-Related Concussion Among High School Football Players? *The American Journal of Sports Medicine*, 39(11):2311–2318.

Lee, E., Lavieri, M. S., Volk, M. L., and Xu, Y. (2015). Applying reinforcement learning techniques to detect hepatocellular carcinoma under limited screening capacity. *Health Care Management Science*, 18(3):363–375.

Li, Y., Zhu, M., Klein, R., and Kong, N. (2014). Using a partially observable Markov chain model to assess colonoscopy screening strategies – A cohort study. *European Journal of Operational Research*, 238(1):313–326.

Maillart, L. M., Ivy, J. S., Ransom, S., and Diehl, K. (2008). Assessing Dynamic Breast Cancer Screening Policies. *Operations Research*, 56(6):1411–1427.

Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1995). Breast Cancer Diagnosis and Prognosis Via Linear Programming. *Operations Research*, 43(4):570–577.

Maruta, J., Lumba-Brown, A., and Ghajar, J. (2018). Concussion Subtype Identification With the Rivermead Post-concussion Symptoms Questionnaire. *Frontiers in Neurology*, 9(December):1–7.

McCrory, P., Meeuwisse, W., Dvorak, J., Aubry, M., Bailes, J., Broglio, S., Cantu, R. C., Cassidy, D., Echemendia, R. J., Castellani, R. J., Davis, G. A., Ellenbogen, R., Emery, C., Engebretsen, L., Feddermann-Demont, N., Giza, C. C., Guskiewicz, K. M., Herring, S., Iverson, G. L., Johnston, K. M., Kissick, J., Kutcher, J., Leddy, J. J., Maddocks, D., Makdissi, M., Manley, G. T., McCrea, M., Meehan, W. P., Nagahiro, S., Patricios, J., Putukian, M., Schneider, K. J., Sills, A., Tator, C. H., Turner, M., and Vos, P. E. (2017). Consensus statement on concussion in sport—the 5 th international conference on concussion in sport held in Berlin, October 2016. *British Journal of Sports Medicine*, pages bjsports–2017–097699.

McDonald, W. I., Compston, A., Edan, G., Goodkin, D., Hartung, H.-P., Lublin, F. D., McFarland, H. F., Paty, D. W., Polman, C. H., Reingold, S. C., Sandberg-Wollheim, M., Sibley, W., Thompson, A., Van Den Noort, S., Weinshenker, B. Y., and Wolinsky, J. S. (2001). Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis. *Annals of Neurology*, 50(1):121–127.

McGregor, M. and Caro, J. J. (2006). QALYs. *PharmacoEconomics*, 24(10):947–952.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., and Others (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):263–269.

McLay, L. A., Foufoulides, C., and Merrick, J. R. W. (2010). Using simulation-optimization to construct screening strategies for cervical cancer. *Health Care Management Science*, 13(4):294–318.

Mohajerin Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using

the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166.

Moons, K. G., Stijnen, T., Michel, B. C., Büller, H. R., Van Es, G.-A., Grobbee, D. E., and Habbema, J. D. F. (1997). Application of Treatment Thresholds to Diagnostic-test Evaluation. *Medical Decision Making*, 17(4):447–454.

Moons, K. G. M., Altman, D. G., Vergouwe, Y., and Royston, P. (2009). Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *British Medical Journal*, 338(7709):1487–1490.

Nease, R. F., Owens, D. K., and Sox, H. C. (1989). Threshold Analysis Using Diagnostic Tests with Multiple Results. *Medical Decision Making*, 9(2):91–103.

Nord, E., Daniels, N., and Kamlet, M. (2009). QALYs: Some Challenges. *Value in Health*, 12(SUPPL. 1):S10–S15.

Odetola, F. O., Bruski, L., Zayas-Caban, G., and Lavieri, M. (2016). An innovative framework to improve efficiency of interhospital transfer of children in respiratory failure. *Annals of the American Thoracic Society*, 13(5):671–677.

Pauker, S. G. and Kassirer, J. P. (1975). Therapeutic Decision Making: A Cost-Benefit Analysis. *New England Journal of Medicine*, 293(5):229–234.

Pauker, S. G. and Kassirer, J. P. (1980). The Threshold Approach to Clinical Decision Making. *New England Journal of Medicine*, 302(20):1109–1117.

Peck, J. S., Benneyan, J. C., Nightingale, D. J., and Gaehde, S. A. (2012). Predicting Emergency Department Inpatient Admissions to Improve Same-day Patient Flow. *Academic Emergency Medicine*, 19(9):E1045–E1054.

Perk, J., De Backer, G., Gohlke, H., Graham, I., Reiner, Ž., Verschuren, W. M. M., Albus, C., Benlian, P., Boysen, G., Cifkova, R., and Others (2012). European Guidelines on Cardiovascular Disease Prevention in Clinical Practice (Version 2012). *International Journal of Behavioral Medicine*, 19(4):403–488.

Pierskalla, W. P. and Brailer, D. J. (1994). Operations Research and The Public Sector. *Handbooks in Operations Research and Management Science*, 6:469–505.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

Roozenbeek, B., Lingsma, H. F., Perel, P., Edwards, P., Roberts, I., Murray, G. D., Maas, A. I., and Steyerberg, E. W. (2011). The added value of ordinal analysis in clinical trials: an example in traumatic brain injury. *Critical Care*, 15(3):R127.

Schell, G. J., Lavieri, M. S., Helm, J. E., Liu, X., Musch, D. C., Van Oyen, M. P., and Stein, J. D. (2014). Using filtered forecasting techniques to determine personalized monitoring schedules for patients with open-angle glaucoma. *Ophthalmology*, 121(8):1539–1546.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). *Lectures on Stochastic Programming*. Society for Industrial and Applied Mathematics.

Shapiro, D. E. (1999). The interpretation of diagnostic tests. *Statistical Methods In Medical Research*, 8(2):113–34.

Sheppard, J. W. and Kaufman, M. A. (2005). A Bayesian approach to diagnosis and prognosis using built-in test. *IEEE Transactions on Instrumentation and Measurement*, 54(3):1003–1018.

Si, B., Yakushev, I., and Li, J. (2017). A sequential tree-based classifier for personalized biomarker testing of Alzheimer's disease risk. *IISE Transactions on Healthcare Systems Engineering*, 7(4):248–260.

Somoza, E. and Mossman, D. (1992). Comparing and Optimizing Diagnostic Tests. *Medical Decision Making*, 12(3):179–188.

Tejada, J. J., Ivy, J. S., Wilson, J. R., Ballan, M. J., Diehl, K. M., and Yankaskas, B. C. (2015). Combined DES/SD model of breast cancer screening for older women, I: Natural-history simulation. *IIE Transactions*, 47(6):600–619.

Teng, Y., Kong, N., and Tu, W. (2015). Optimizing strategies for population-based chlamydia

infection screening among young women: an age-structured system dynamics approach. *BMC Public Health*, 15(1):639.

Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., Lallas, A., Lapins, J., Longo, C., Malvehy, J., Marchetti, M. A., Marghoob, A., Menzies, S., Oakley, A., Paoli, J., Puig, S., Rinner, C., Rosendahl, C., Scope, A., Sinz, C., Soyer, H. P., Thomas, L., Zalaudek, I., and Kittler, H. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The Lancet Oncology*, 20(7):938–947.

van Giessen, A., de Wit, G. A., Moons, K. G., Dorresteijn, J. A., and Koffijberg, H. (2018). An alternative approach identified optimal risk thresholds for treatment indication: an illustration in coronary heart disease. *Journal of Clinical Epidemiology*, 94:122–131.

Vermont, J., Bosson, J., François, P., Robert, C., Rueff, A., and Demongeot, J. (1991). Strategies for graphical threshold determination. *Computer Methods and Programs in Biomedicine*, 35(2):141–150.

Wang, T. J., Massaro, J. M., Levy, D., Vasan, R. S., Wolf, P. A., D'Agostino, R. B., Larson, M. G., Kannel, W. B., and Benjamin, E. J. (2003). A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the Framingham Heart Study. *Journal of the American Medical Association*, 290(8):1049–56.

Weise, K., Hübel, K., Rose, E., Schläger, M., Schrammel, D., Täschner, M., and Michel, R. (2006). Bayesian decision threshold, detection limit and confidence limits in ionising-radiation measurement. *Radiation Protection Dosimetry*, 121(1):52–63.

Xue, Y., Klabjan, D., and Luo, Y. (2019). Predicting ICU readmission using grouped physiological and medication trends. *Artificial Intelligence in Medicine*, 95(August):27–37.

Yang, Y., Goldhaber-Fiebert, J. D., and Wein, L. M. (2013). Analyzing Screening Policies for Childhood Obesity. *Management Science*, 59(4):782–795.

Yao, Y. (2010). Three-way decisions with probabilistic rough sets. *Information Sciences*, 180(3):341–353.

Yao, Y. and Zhou, B. (2016). Two Bayesian approaches to rough sets. *European Journal of Operational Research*, 251(3):904–917.

Zhang, J., Denton, B. T., Balasubramanian, H., Shah, N. D., and Inman, B. A. (2012). Optimization of Prostate Biopsy Referral Decisions. *Manufacturing & Service Operations Management*, 14(4):529–547.

Zhang, M. L., Li, Y. K., Liu, X. Y., and Geng, X. (2018). Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202.

Zhang, M.-L. and Zhang, K. (2010). Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*, page 999, New York, New York, USA. ACM Press.

Zhu, Y. and Fang, J. (2016). Logistic Regression–Based Trichotomous Classification Tree and Its Application in Medical Diagnosis. *Medical Decision Making*, 36(8):973–989.

Zimmer, A., Marcinak, J., Hibyan, S., and Webbe, F. (2015). Normative Values of Major SCAT2 and SCAT3 Components for a College Athlete Population. *Applied Neuropsychology: Adult*, 22(2):132–140.

Zufferey, D., Hofer, T., Hennebert, J., Schumacher, M., Ingold, R., and Bromuri, S. (2015). Performance comparison of multi-label learning algorithms on clinical data for chronic diseases. *Computers in Biology and Medicine*, 65:34–43.