New Developments Of Minimum Clinically Important Difference: Theory and Methodology

BY
TU XU
B.S., EAST CHINA NORMAL UNIVERSITY, SHANGHAI, CHINA, 2005
M.S., EAST CHINA NORMAL UNIVERSITY, SHANGHAI, CHINA, 2008
M.S., OHIO UNIVERSITY, ATHENS, OH, 2009
M.S., UNIVERSITY OF ILLINOIS AT CHICAGO, CHICAGO, IL, 2010
of the
University of Illinois at Chicago

THESIS
Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathematics
in the Graduate College of the
University of Illinois at Chicago, 2013
Chicago, Illinois

Defense Committee:

Samad Hedayat, Chair and Advisor, MSCS
Junhui Wang, Advisor, MSCS
Huayun Chen, Division of Epidemiology and Biostatistics
Ryan Martin, MSCS
Jing Wang, MSCS

Dedicated to

my parents and my wife Xuexue Qiao

ACKNOWLEDGMENT

I would like to take this opportunity to express my gratefulness to all who have warmly supported my study and research during my time at University of Illinois at Chicago.

My special thanks are for my advisors Professor Samad Hedayat and Professor Junhui Wang, whose guidance, mentoring, and unsurpassed support on my academic and personal life made my graduate study at UIC a precious and rewarding experience. This dissertation could not have been accomplished without their continuous help.

I also want to thank Dr. Lawrence Lin and Dr. Xin Fang for their insightful and patient guidance during my stays at Baxter Healthcare Corporation and U.S. Food and Drug Administration. Their continuous encouragement and enlightening discussion made my dissertation research enjoyable and exciting. My other sincere thanks go to Professor Dibyen Majumdar, Professor Jing Wang, and Professor Jie Yang for their instruction on my coursework without which the work could not be possible.

It is my honor to have Professor Huayun Chen, Professor Ryan Martin, and Professor Jing Wang serving as members of my dissertation defense committee. I extremely appreciate their extraordinary support and assistance.

Last, but not the least, I want to thank all my family members, for their love and support.

The research work and writing of this dissertation are supported by The United States National Science Foundation (NSF) Grants DMS-0904125 and DMS-1306394 of Professor Samad

ACKNOWLEDGMENT (Continued)

Hedayat. The contents are solely the responsibility of the author and do not necessarily represent the official view of NSF.

T.X.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DCA	Difference Convex Algorithm
FDA	Food and Drug Administration
MCID	Minimum Clinically Important Difference
NPV	Negative Predictive Value
PPV	Positive Predictive Value
PRO	Patient Reported Outcome
RKHS	Reproducing Kernel Hilbert Space
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine

SUMMARY

In clinical trials, minimum clinically important difference (MCID) has attracted increasing interest as an important supportive clinical and statistical inference tool. Many estimation methods have been developed based on various intuitions, while little theoretical justification has been established. In this dissertation, a new estimation framework of MCID using both diagnostic measurements and patient-reported outcomes (PROs) is proposed. It first provides a precise definition of population-based MCID so that estimating such a MCID can be formulated as a large margin classification problem. The framework is then extended to personalized MCID to allow individualized thresholding value for patients whose clinical profiles may affect their PRO responses. More importantly, it is shown that the proposed estimation framework is asymptotically consistent, and a finite-sample upper bound is established for its prediction accuracy compared against the ideal MCID. The advantage of our proposed method is also demonstrated in a variety of simulated experiments as well as real applications.

The dissertation is organized as follows. In the first part of Chapter 1, the background and existing determination of MCID are introduced. Margin-based methods and reproducing kernel Hilbert space are important tools for our proposed MCID estimation, and a literature review on them is included as the second part of Chapter 1. Chapter 2 presents a general framework for the population-based MCID and its estimation algorithm and asymptotic properties are studied. Chapter 3 extends the framework to the personalized MCID, and discusses in details the appropriate large margin loss as well as the efficient non-convex optimization technique.

The asymptotic properties of our proposed method for estimating the personalized MCID are also established. Chapter 4 presents numerical experiments that demonstrate the performance of our proposed method in simulated examples, and Chapter 5 applies our proposed method to two benchmark datasets as well as two phase-3 clinical trial datasets. Chapter 6 contains some conclusions and discussion on future research topics, especially the research on the Youden index. Selected technical proofs are given in an appendix.

CHAPTER 1

INTRODUCTION

1.1 Background

In clinical trials for drugs or medical devices, statistical significance is widely used to infer the effectiveness of drugs or medical devices. However, there has been growing recognition that statistical significance could be misleading when evaluating treatment effect (Jacobson et al., 1984; Jacobson and Truax, 1991).

In many trials, the statistical significance of the treatment effect may have little to do with its clinical significance. It is known that statistical significance only infers the existence of treatment effect, regardless of the effect size. Further, the statistical significance could result from a small sample variability or a huge sample size, and thus provides little information about the clinical meaningfulness of the treatment (Jacobson and Truax, 1991). For instance, in a paired t-test, no matter how small the size of a treatment effect d is, statistical significance could always be declared when the sample size n is large enough, such as $n = O(1/d^2)$.

Furthermore, the statistical significance for the treatment group compared to the placebo group ignores the possible heterogeneity among individuals. For instance, in a pain reduction study, a statistically significant reduction is concluded for a test treatment while many individual patients in the treatment group actually report little improvement regarding the pain reduction (Younger et al., 2009).

Clinical significance is desired in practice as it provides a better assessment of the clinically meaningful improvement. It is often based on the patients' reports in a community according to certain external standards (Jacobson and Truax, 1991). One common approach is to collect patient-reported outcomes (PROs) (FDA, 2009), such as their satisfaction of a treatment. Some earlier practice suggested to replace the statistical significance test by analyzing the PROs only, which is problematic due to the subjective bias in the PROs or unreliability of a poorly designed questionnaire. The concept of minimum clinically important difference (MCID) is proposed to overcome the aforementioned shortcomings by incorporating both certainty of effective treatment and patients' satisfactions.

1.2 The concept of MCID

The MCID, as discussed in Jaeschke et al. (1989), was intuitively defined as a thresholding value in post-treatment change, and a patient is considered experiencing a clinically meaningful improvement if her/his change exceeds the MCID. The concept of MCID provides objective reference for clinicians and health policy makers regarding the effectiveness of the treatment, and has quickly gained popularity among practitioners. In November 2012, FDA hosted a special conference on the MCID for orthopaedic devices; c.f. `http://www.fda.gov/MedicalDevices/NewsEvents/Workshops/Conferences/ucm327292.htm` for more details.

1.2.1 Existing methods for determining MCID

Although the importance of MCID has been recognized, only a few ad-hoc approaches have been proposed for its estimation. Copay et al. (2007) gave a comprehensive review on these

methods, and classified them into two major types: anchor-based approaches and distribution-based approaches.

Anchor-based approaches

The main idea of anchor-based approach is to compare the change in PRO with other external criteria (anchor) such as a rating assessment from physical therapists or a change on diagnostic measurement. The MCID is determined based on the association between the PRO and the anchor. Among the existing anchor-based approaches, within-patients score change, between-patients score change, social comparison approach, and sensitivity and specificity-based approach are four popular variations.

Within-patients score change estimates MCID as the mean change in PRO scores of a group of patients who are pre-selected as markers. The selected patients are those who have exhibited small change, such as patients whose scores are $\pm 1, 2, 3$ on a 15-point scale in Jaeschke et al. (1989) or only those with scores $\pm 2, 3$ in Juniper et al. (1994). Between-patients score change defines MCID as the difference in the score change of two adjacent levels of a scale, such as "better" and "unchanged" patients (Hägg et al., 2003). Social comparison approach estimates the MCID using the score change of patients who rate themselves in a better situation when compared with others. For all these three variations, the arbitrariness comes in when selecting patients for estimating the MCID.

In statistics literature, sensitivity and specificity-based approaches have been widely applied in the research on MCID. Sensitivity and specificity are defined as the proportion of true positives and true negatives, respectively. The MCID is typically defined as the efficacy size

with equal sensitivity and specificity among researchers. Similarly, Bennett (1985), Leisenring and Alonzo (2000) discussed the connection between the MCID and positive predictive value (PPV) and negative predictive value (NPV). Shiu and Gatsonis (2008) even defined MCID as the maximizer of the sum of PPV and NPV based on the argument that the sum reflects a distance from ideal situation without further elaboration, where the ideal situation refers to a perfect match between PRO and predicted response based on diagnostic measurements.

Distribution-based approaches

Distribution-based approaches are methods that compare the change in PRO scores to certain measure of its variability such as the standard error of measurement (SEM). The SEM refers to the variation in the patients' PRO scores, and Wyrwich et al. (1999) defined MCID equal to SEM. Copay et al. (2007) pointed out that distribution-based approaches do not really address the clinical significance and ignore the purpose of MCID.

For these existing estimations of MCID, little theoretical justification has been developed. Different methods lead to different estimations of MCID and no agreement has been reached regarding the suitability of estimating MCID.

The next section is devoted to a literature review on margin-based methods and its estimation in reproducing kernel Hilbert space, which are important tools for our proposed methods.

1.3 Margin-based methods and reproducing kernel Hilbert space

1.3.1 Margin-based methods

Margin-based approaches are wildly used for classification problems and have gained great successes in real applications. The general idea is to seek a classifier that minimizes the pre-specified loss function. To evaluate the learning accuracy of a loss function, generalization error, $\frac{1}{2}E(1-\mathrm{sign}(Yf(X)))$ for binary classification or $P(Y \neq \mathrm{argmax}_j f_j(X))$ for multi-category classification, is commonly used.

Definition *The Bayes decision rule* $f^*(x)$ *for classification is the classification rule that yields minimal generalization error.*

Lin (2002) proved that the Bayes decision rule for binary classification is $\mathrm{sign}(p(x) - 1/2)$, where $p(x) = P(Y = 1|X = x)$. Similarly, for multi-category classification, the rule is $\mathrm{argmax}_{j=1,\ldots,k} p_j(x)$, where k is the number of classes and $p_j(x) = P(Y = j|X = x)$. In the situation of the weighted classification, Wang et al. (2008) showed that $\mathrm{sign}(p(x) - \pi)$ is the Bayes decision rule when the weight $1 - \pi$ is imposed on the class $Y = 1$. Wu et al. (2010) proved that the rule for the multi-category case is $\mathrm{argmax}_{j=1,\ldots,k} \pi_j p_j(x)$, where π_j is the weight function and $\sum_{j=1}^{k} \pi_j = 1$.

Support vector machine (SVM) (Cortes and Vapnik, 1995; Vapnik, 1996) is one of the most commonly used tools for classification problems. It aims to construct an optimal hyperplane that creates the largest margin between two classes. In the non-separable case, the hinge loss function $L_{SVM}(y, f(x)) = (1 - yf(x))_+$ is adopted to maximize the sum of residual distance. Lin (2002) showed that the minimizer of $E(L_{SVM}(Y, f(X)))$ targets directly on $\mathrm{sign}(p(x) - 1/2)$.

Lee et. al (2004) extended the binary SVM to multi-category support vector machine (MSVM). The proposed loss function is $L_{MSVM}(y, f(x)) = (W(y) \cdot (f(x) - y)_+)$, where y and $W(y)$ are two k-dimensional vectors with 1 in the l-th coordinate and $-1/(k-1)$ elsewhere and 0 in the l-th coordinate and 1 elsewhere, receptively if $l = \mathrm{argmax}_{j=1,\ldots,k} p_j(x)$. $y - f(x)$ is a k-dimensional vector $(f_1(x) - y_1, \ldots, f_k(x) - y_k)$. They proved that the minimizer of $E(L_{MSVM}(Y, f(X)))$ targets on the Bayes decision rule, that is, $\mathrm{argmax}_{j=1,\ldots,k} f_j^*(x) = \mathrm{argmax}_{j=1,\ldots,k} p_j(x)$. Liu and Yuan (2011) proposed the reinforced multi-category hinge loss functions $L_{RMSVM}(y, f(x)) = \gamma[(k-1) - f_y(x)]_+ + (1-\gamma)\sum_{j \neq y}[1 + f_j(x)]_+)$, which includes MSVM as a special case by setting $\gamma = 0$ and connects with one-versus-rest approach (Weston, 1999) when $\gamma = 1/2$.

Distance weighted discrimination (DWD) is another margin-based method for binary classification (Marron et al., 2007). In contrast to SVM that aims to maximize the sum of residual distance, DWD seeks classifiers that minimize the sum of inverse residual distance. Qiao et al. (2010) showed that the classifier generated by DWD targets on the Bayes decision rule. Marron et al., (2007) and Qiao et al. (2010) demonstrated DWD's advantage over SVM under high dimension low sample size setup through theoretical and numerical analyses. DWD was generalized to multi-category DWD (MDWD) by Huang et al. (2013).

Shen et al. (2003) proposed the ψ-loss $L_\psi(y, f(x)) = \min((1 - yf(x))_+, 1)$ to seek a classifier that minimize the generalization error for binary classification. It was showed by theoretical and numerical analyses that the ψ-learning not only targets on the Bayes decision rule, but

also can outperform SVM regarding generalization properties. Liu and Shen (2006) extended the ψ-learning to multi-category classification and similar properties were presented.

Both SVM and ψ-learning target estimating the Bayes decision rule is $f^*(x) = \mathtt{sign}(p(x) - 1/2)$ while the conditional probability $p(x)$ itself is also interesting. Wahba (2002) and Liu et al. (2011) compared the hard classification which focuses on predicting the class label and the soft classification whose primary goal is to estimate $p(x)$. Zhu and Hastie (2005) proposed the import vector machine where logistic loss $L_{log}(y, f(x)) = \log(1 + e^{-yf(x)})$ is employed. Direct derivation yields that $f(x) = \log(p(x)/(1 - p(x)))$ minimizes $E(L_{log}(Y, f(X)))$, and therefore $\mathtt{sign}(f(x))$ targets on the Bayes classification rule. Import vector machine also provides an estimate for $p(x) = e^{f(x)}/(1 + e^{f(x)})$. Multi-category logistic loss was also generalized in their paper.

Since the distribution of (X, Y) is unknown, the classifier is estimated through minimizing the empirical version of a pre-specified loss function $(1/n)\sum_{i=1}^{n} L(y_i, f(x_i))$. However, estimation using only empirical version suffers the problem of over-fitting. The penalty (regularization) term $J(f)$ is commonly employed to address this issue. Some typical regularization models are ridge regression, LASSO (Tibshirani, 1996), adaptive LASSO (Zou, 2006), SCAD (Fan and Li, 2001) and Elastic Net (Zou and Hastie, 2005). Therefore, the classifier is obtained by searching $f(x)$ in a candidate functional space $\mathcal{F}$ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \frac{\lambda}{2} J(f), \tag{1.1}$$

where $L(\cdot)$ is a pre-specified loss function, λ is a tuning parameter.

The selection of tuning parameter is also crucial to the model performance. For instance, the selection consistency of lasso regression depends on the convergence rate of the tuning parameter (Zhao and Yu, 2006). In literature, many criteria have been proposed for selecting tuning parameter, such as AIC (Akaike, 1974), BIC (Schwarz, 1978), cross-validation (Stone, 1974; Geisser, 1975), SRM (Vapnik, 1996) or stability selection (Meinshausen and Bühlmann, 2010; Wang, 2010). AIC and BIC are well known criteria that aim to optimize the in-sample error while cross-validation focus on the optimization of extra-sample error. Cross-validation is one of the most commonly used methods for tuning parameter selection in classification problems. In more details, K-folder cross-validation refers to the method that splits available data into K roughly equal-sized parts and uses $K - 1$ parts of them for model fitting and the remaining part for prediction. When K is set as the number of available data points, it is also known as leave-one-out cross-validation. Craven and Wahba (1979) proposed generalized cross-validation (GCV) as a convenient approximation for leave-one-out cross-validation. SRM is a selection criterion built on the theory of VC dimension. Stability selection motivates form the idea that a proper selection of tuning parameter should yield good estimates with high probability from different samples.

1.3.2 Reproducing kernel Hilbert space

In practice, the candidate functional space $\mathcal{F}$ is often set as a reproducing kernel Hilbert spaces (RKHS) $\mathcal{H}_K$ generated by a pre-specified reproducing kernel $K(\cdot, \cdot)$. The RKHS includes the entire family of smoothing splines (Wahba, 1990) and therefore is widely used for estimat-

ing complex decision function $f(x)$. Some popular choices of reproducing kernel in practice are linear kernel $K(x_1, x_2) = x_1^T x_2$ and Gaussian kernel $K(x_1, x_2) = exp(-\|x_1 - x_2\|^2/2\sigma^2)$. More importantly, the representer theorem (Wahba, 1990) shows that the solution $f \in \mathcal{H}_K$ to Equation 1.1 is of the form $\hat{f}(x) = b + \sum_{i=1}^{n} w_i K(x_i, x)$. That is to say, although it may require an infinite expansion of basis kernel $K(x, \cdot)$ to represent all functions in $\mathcal{H}_K$, the solution to Equation 1.1 is always in an additive form of finite basis kernels. Therefore, $J(f) = w^T K w$ with $w = (w_1, \cdots, w_n)^T$ and $K = (K(z_i, z_j))_{i,j=1}^{n}$ is a proper penalty function to enforce the smoothness of f (Wahba, 1990). Then the formulation involving RHKS is

$$\min_{b,w} \frac{1}{n} \sum_{i=1}^{n} L\Big(y_i(b + \sum_{i=1}^{n} w_i K(x_i, x))\Big) + \frac{\lambda}{2} w^T K w.$$

Depending on the choice of loss functions, many methods are employed to solve for b, w such as quadratic programming (SVM, Hastie et al., 2009), DCA (ψ-learning, Shen et al., 2003; Liu and Shen, 2006) or Newton-Raphson method (Zhu and Hastie, 2005).

1.4 A new method for estimating MCID

Copay et al. (2007) suggested that the most useful concept of MCID should incorporate both certainty of effective treatment and patients' satisfaction which motivates a new framework for estimating MCID in Chapter 2. More importantly, personalized MCID is proposed in Chapter 3 to capture the effect of various factors such as age, gender, or baseline diagnostic measurement on MCID by allowing MCID to vary based on each patient's clinical profile. The estimation scheme is developed in large margin classification framework and DCA is employed. Asymptotic

results are also established. The effectiveness of our proposed methods are demonstrated in simulated experiments as well as real applications in Chapter 4 and Chapter 5.

CHAPTER 2

A NEW FRAMEWORK FOR MINIMUM CLINICALLY IMPORTANT DIFFERENCE

2.1 Defining MCID

In this dissertation, the MCID is formally defined as the thresholding value in post-treatment change such that the probability of disagreement between the estimated satisfaction based on the MCID and the PROs is minimized.

Suppose that a patient's diagnostic measurement $X \in \mathcal{R}^1$ is continuously connected, and the patient-reported outcome (PRO) $Y \in \{-1, 1\}$, where $Y = 1$ denotes a clinically meaningful treatment reported by the patient and $Y = -1$ otherwise. Let $f(x, y)$, $f_y(x)$ and $f(x)$ be the joint density of (X, Y), the conditional density of X given $Y = y$, and the marginal density of X, respectively. The MCID is defined as the thresholding value c^* such that $\mathrm{sign}(X - c^*)$ agrees with Y as much as possible, where $\mathrm{sign}(u) = 1$ if $u \geq 0$ and -1 otherwise. Mathematically, c^* is defined as a solution of

$$\min_c \; P(Y \neq \mathrm{sign}(X - c)) = \min_c \; \frac{1}{2} E\left(1 - Y \; \mathrm{sign}(X - c)\right), \tag{2.1}$$

where $P(\cdot)$ and $E(\cdot)$ are taken with respect to both X and Y.

Lemma 1 *Assume that* $p(x) = P(Y = 1|X = x)$ *is continuous and increasing in* x*, then the MCID* c^* *satisfies*

$$p(c^*) = P(Y = 1|X = c^*) = \frac{1}{2}. \tag{2.2}$$

Furthermore, if $p(x)$ *is strictly increasing in* x*, then* c^* *is unique.*

Proof Note that c^* is a solution of

$$\min_c \frac{1}{2}E(1 - Y\ sign(X - c)) = \min_c \frac{1}{2}E_X(1 - E(Y|X)\ sign(X - c)),$$

where E_X represents the expectation with respect to X. It then suffices to find c^* to maximize $E(Y|X = x)\ sign(x - c)$ for any given x. Therefore, c^* must satisfy that

$$sign(x - c) = sign(E(Y|X = x)) = sign(2p(x) - 1), \tag{2.3}$$

for any x. We now show contradiction when $p(c^*) \neq 1/2$. Without loss of generality, assume $p(c^*) > 1/2$. Since $p(x)$ is continuous and increasing in x, there must exist $\tilde{c}$ such that $p(\tilde{c}) = 1/2$ and $\tilde{c} < c^*$. This leads to the contradiction to (Equation 2.3) since

$$0 > sign(\tilde{c} - c^*) = sign(2p(\tilde{c}) - 1) = 1.$$

Therefore, c^* must satisfies $p(c^*) = \frac{1}{2}$. Furtheremore, when $p(x)$ is continuous and strictly increasing, the uniqueness follows from the fact that $p(c^*) = \frac{1}{2}$ has a unique solution. □

Lemma 1 describes the ideal MCID when $p(x)$ is known, which is analogous to the Bayes rule in classification (Lin 2002; Hastie et al., 2009). Note that it is reasonable to assume that $p(x)$ is increasing in x since patients with better diagnostic measurements are expected to be more likely to give positive responses. If $p(x)$ is only non-decreasing, the equation in (Equation 2.2) may have multiple roots and a conservative choice is to set c^* as the largest root. Furthermore, the continuity assumption of $p(x)$ can be relaxed to semi-continuity, and then the equation in (Equation 2.2) may have no root at all. In such scenarios, it could be proved similarly as Lemma 1 that $c^* = \mathtt{argmin}_c\{p(c) \geq 1/2\}$.

It is known that the quality of PROs is largely affected by patients' subjectivity (Frost et al., 2007). The proposed definition of MCID (Equation 2.1) accounts for such subjectivity in $p(x)$, which can be interpreted as the probability of patient's telling the truth. In Fang (2011), the subjectivity is explicitly modeled by a parameter Q and assumes that $p(x) = Q$ when $x \geq c^*$ and $p(x) = 1 - Q$ otherwise, where $Q > \frac{1}{2}$ measures how trustworthy the PROs are. It is clear that this assumed model is a special case of the semi-continuous $p(x)$. More importantly, the optimal MCID (Equation 2.2) is less affected by the subjectivity in the PROs, as it relies on $p(x)$ only when x is in the neighborhood of c^*. This is similar to the Bayes rule in classification, where the optimal classification only relies on whether $p(x) \geq 1/2$ or not (Lin, 2002).

Based on (Equation 2.3), the MCID c^* can also be interpreted as the thresholding test outcome for which a patient's own judgment is completely random. If the test outcome is higher than the MCID, a patient would be more likely to think that the test is clinically meaningful, and vise versa.

In addition, the MCID has an interesting connection with the median lethal dose in toxicology research. The median lethal dose refers to the smallest dose required to kill half of the animals that receive it after a specified test duration. To describe the interaction between dosage and mortality rate, the logistic dose-response curve is popularly used (Williams, 1986; Alho and Valtonen, 1995; Kelly, 2001) . It assumes that the mortality rate is expected to strictly increase with dose, which coincides with our assumption in Lemma 1.

2.2 Estimating MCID

In practice, the conditional distribution $p(x)$ is unknown, and thus the MCID needs to be estimated based on the available i.i.d. training sample $(x_i, y_i)_{i=1}^n$. Naturally, the expectation in (Equation 2.1) can be approximated by its empirical version, and the estimated MCID $\hat{c}$ is defined as a solution of

$$\min_c \ \frac{1}{2n} \sum_{i=1}^{n} (1 - y_i \ sign(x_i - c)) . \tag{2.4}$$

Note that (Equation 2.4) is a simple 1-dimensional optimization problem, and the objective function remains the same for $x_{(i)} \leq c < x_{(i+1)}$, where $x_{(i)}$ is the i-th order statistic. Therefore, an exhaustive grid search scheme can be implemented, and the global minimizer $\hat{c}$ is simply the x_i that yields the smallest objective function value.

Definition *The estimated MCID* $\hat{c}$ *is a consistent estimate of* c^* *if* $\hat{c} \xrightarrow{P} c^*$ *as* $n \to +\infty$.

Theorem 1 *The estimated MCID* $\hat{c}$ *in (Equation 2.4) is a consistent estimate of* c^* *if* $p(x)$ *is continuous and strictly increasing in* x*. Further, if there exist positive constants* α_1, $\gamma_1 < 2/\alpha_1 + 4/\alpha_1^2$, a_1 *and* a_2*, such that for sufficiently small* $\xi > 0$*,*

$$P(|p(X) - p(c^*)| \leq \xi) \leq a_1 \xi^{\alpha_1}, \tag{2.5}$$

$$\sup_{|x-c^*|\leq \xi} |p(x) - p(c^*)| \leq a_2 \xi^{\gamma_1}, \tag{2.6}$$

then $|\hat{c} - c^*| = O_p\Big((n \log^{-2} n)^{-1/(2(1+2/\alpha_1)-\alpha_1\gamma_1)}\Big)$.

Theorem 1 establishes the asymptotic convergence rate of $|\hat{c} - c^*|$, and the finite sample bound for $|\hat{c} - c^*|$ can also be found as in (Equation .5) of Appendix. In Theorem 1, (Equation 2.5) is similar to the low noise assumption (Polonik, 1995; Bartlett et al., 2003; Tsybakov, 2004) that describes the behavior of X in the neighborhood of c^*, and (Equation 2.6) is implied by a Hölder continuity condition on $p(x)$.

For illustration, if X is uniformly distributed on $[a, b]$ and (Equation 2.6) is met with γ_1, then (Equation 2.5) can be verified with $\alpha_1 = 1/\gamma_1$ for sufficiently small ξ, and thus Theorem 1 implies that $|\hat{c} - c^*| = O_p\Big((n \log^{-2} n)^{-1/(1+4\gamma_1)}\Big)$. It leads to a fast convergence rate when $p(x)$ has a steep derivative at c^* with γ_1 close to 0, and a rate $(n \log^{-2} n)^{-1/3}$ when $p(x)$ is Hölder continuous at c^* with order $\gamma_1 = 1/2$.

2.3 Weighted MCID

In many clinical studies, it is a common practice to be conservative when predicting whether the test outcome is clinically meaningful. It is then less desirable to predict positive for an

unsatisfied patient than negative for a satisfied patient. To accommodate the unbalanced severity, the weighted MCID can be introduced with the weights reflecting the severity of the disagreements. Specifically, the weighted MCID c_w^* is defined as a solution of

$$\min_c \ \frac{1}{2}E\Big(w(Y)\,(1 - Y\ sign(X-c))\Big), \tag{2.7}$$

where $w(1) = w$ and $w(-1) = 1 - w$. Similarly as in Lemma 1, it can be shown that

$$p(c_w^*) = P(Y = 1|X = c_w^*) = 1 - w, \tag{2.8}$$

where an appropriate choice of $w < 1/2$ leads to a conservative estimation.

The weighted MCID has another useful interpretation in the context of hypothesis testing. In particular, we denote the type-I error and type-II error as $R_0(c) = P(X - c > 0|Y = -1)$ and $R_1(c) = P(X - c < 0|Y = 1)$, respectively. Then it is natural to find c_α^* to solve

$$\min_c R_1(c) \ \text{ subject to } \ R_0(c) \le \alpha, \tag{2.9}$$

where α is the significance level as in the standard hypothesis testing setup. This formulation is closely related with the Neyman-Pearson classification as discussed in Scott and Nowak (2005) and Rigollet and Tong (2011). More interestingly, Lemma 2 establishes a one-to-one correspondence between the weighted MCID c_w^* in (Equation 2.8) and the solution c_α^* in (Equation 2.9).

Lemma 2 *Assume that* $p(x)$ *is continuous and strictly increasing in* x*, then for any* α*, there exists a unique* w *such that* $c_\alpha^* = c_w^*$*, and vice versa.*

Proof Since X is continuously supported in $\mathcal{R}$, it follows immediately that $R_0(c)$ and $R_1(c)$ are continuous in $c \in \mathcal{R}$, and $R_0(C)$ and $R_1(c)$ are decreasing and increasing with respect to c, respectively. Therefore, for any α, the solution c_α^* to (Equation 2.9) satisfies $R_0(c_\alpha^*) = P_{X|Y=-1}(X > c_\alpha^*) = \alpha$.

Let $w = 1 - p(c_\alpha^*)$, then $c_\alpha^* = c_w^*$ by (Equation 2.8). The uniqueness follows from the strict monotonicity of $p(x)$ and Lemma 1. On the other hand, for any w, since $p(x)$ is strictly increasing, there is a unique solution c_w^* to (Equation 2.8). Let $\alpha = P_{X|Y=-1}(X > c_w^*)$, then $c_\alpha^* = c_w^*$. $\square$

If $p(x)$ is only non-decreasing, there may exist multiple roots to the equation in (Equation 2.8) for certain w, whereas the solution c_α^* to (Equation 2.9) is still uniquely defined for any given α. In fact, it can be showed that c_α^* and α are one-to-one correspondent as long as $p(x)$ stays away from 0 and 1.

In this chapter, a new framework for defining as well as estimating MCID is proposed. Especially, the proposed method is formulated as a large margin classification framework and the asymptotic properties are established. The weighted MCID and its connection to Neyman-Pearson classification are also studied. The MCID defined in this chapter is population-based, whereas the difference among subgroups of patients should also be considered. In the next chapter, personalized MCID is proposed to address this issue.

CHAPTER 3

PERSONALIZED MCID

In many clinical trials, it is commonly believed that patients' report could be influenced by various factors such as their expectation of treatment (Wise, 2004). For instance, in a shoulder pain reduction study, healthy people demonstrate a higher threshold than those with chronic conditions due to their expectation of complete recovery. It is expected that the clinical profile of each patient can largely affect her judgment on the clinically meaningful blood loss reduction. This chapter extends the estimation framework to personalized MCID by allowing the MCID to vary according to each patient's clinical profiles.

3.1 Formulation

Suppose that a patient's diagnostic measurement X and PRO Y are defined the same as those in Chapter 2 and $Z = (Z_1, \ldots, Z_p)$ denotes patients' clinical profiles, such as age, gender or the baseline diagnostic measurement. Then the personalized MCID $c^*(z)$ is defined as a solution of

$$\min_c P(Y \neq \mathrm{sign}(X - c(Z))) = \min_c \frac{1}{2} E\left(1 - Y\ \mathrm{sign}(X - c(Z))\right), \tag{3.1}$$

where $P(\cdot)$ and $E(\cdot)$ are taken with respect to X, Y and Z. Similarly as in (Equation 2.2), we can show that $c^*(z)$ satisfies

$$p_z(c^*(z)) = P(Y = 1 | X = c^*(z), Z = z) = \frac{1}{2}, \tag{3.2}$$

where $p_z(x) = P(Y = 1|X = x, Z = z)$ is assumed to be a continuous and strictly increasing function in x for any value of z. If only semi-continuity is assumed, the MCID can be defined as $c^*(z) = \mathrm{argmin}_c\{c : p_z(c) \geq \frac{1}{2}\}$.

Although the formulation in (Equation 3.1) is similar as in (Equation 2.1) with population-based c^*, the difficulty arises in the estimation part. Since the empirical version of (Equation 3.1)

$$\min_c \frac{1}{2n} \sum_{i=1}^{n} \Big(1 - y_i \ \mathrm{sign}(x_i - c(z_i))\Big), \tag{3.3}$$

involves the 0-1 loss $L_{01}(u) = \frac{1}{2}(1-\mathrm{sign}(u))$ and needs to be optimized with respect to functional $c(z)$, it can no longer be solved by the exhaustive grid search or any other efficient optimization techniques. Therefore, a surrogate loss function needs to be introduced to replace the 0-1 loss and facilitate the estimation. The idea of using a surrogate loss to simplify computation has been widely studied in machine learning literature. As for the properness of surrogate loss functions, Fisher consistency is often used as a minimal condition (Lin, 2002).

Definition *A surrogate loss function* $L(u)$ *is Fisher consistent in estimating* $c^*(z)$ *if* $c^*(z) = \mathrm{argmin}_c E(L(Y \ \mathrm{sign}(X - c(Z))))$.

Popularly used surrogate loss functions $L(u)$ include the hinge loss $L(u) = (1-u)_+$ (Vapnik, 1996), the logistic loss $L(u) = \log(1 + \exp(-u))$ (Zhu and Hastie, 2005), and the ψ-loss $L(u) = \min((1 - u)_+, 1)$ (Shen et al., 2003). Unfortunately, the hinge loss, logistic loss and ψ-loss are not generally Fisher consistent in estimating $c^*(z)$, and counter examples can be easily constructed.

For instances, assume that X is uniformly distributed on $[a, b]$, $p(x)$ is continuous and strictly increasing, and $\min\{c^* - a, b - c^*\} > 1$. By (Equation 2.2), c^* is the unique MCID. On the other hand, the minimizer of the hinge loss must satisfy that

$$\int_a^{c^*+1} p(x)dx - \int_{c^*-1}^{b} (1 - p(x))dx = 0,$$

the minimizer of the logistic loss must satisfy that

$$\int_a^b \left[p(x) - \frac{1}{1 + e^{c^*-x}}\right] dx = 0,$$

and the minimizer of the ψ-loss must satisfy that

$$\int_{c^*-1}^{c^*+1} \left[p(x) - \frac{1}{2}\right] dx = 0.$$

These equalities do not hold in general. For instance, let $p''(x) = 0$, when $x \geq c^*$ and $p''(x) > 0$ otherwise, then minimizers for all three losses are strictly higher than c^*.

In this dissertation, we propose a novel surrogate loss, ψ_δ-loss, which is defined as

$$L_\delta(u) = \min\left(\frac{1}{\delta}(\delta - u)_+, 1\right). \tag{3.4}$$

The ψ_δ-loss (ψ_1-loss in Figure 1) extends the ψ-loss by introducing a new parameter δ that controls the difference between the surrogate loss and the 0-1 loss. More importantly, Lemma 3

shows that the ψ_δ-loss is asymptotically Fisher consistent in estimating $c^*(z)$ when δ converges to 0.

Lemma 3 *For any given z, if $f_z(x) = f(x|Z = z)$ is continuous and $p_z(x)$ is strictly increasing in x, then $E\Big(L_\delta(Y(X-c))|Z = z\Big)$ converges to $E\Big(L_{01}(Y(X-c))|Z = z\Big)$ as $\delta \to 0$ uniformly over a compact set $\mathcal{D}_z$ containing $c^*(z)$ and*

$$\mathrm{argmin}_c E\Big(L_\delta(Y(X - c(z)))|Z = z\Big) \longrightarrow c^*(z).$$

Proof For any given z, since $L_\delta(u) = L_{01}(u) + \delta^{-1}(\delta - u)I(0 \le u \le \delta)$, we have

$$\begin{aligned} E\Big(L_\delta(Y(X-c))|Z = z\Big) &= E\Big(L_{01}(Y(X-c))|Z = z\Big) && (3.5)\\ &\quad + E\Big(\frac{\delta - Y(X-c)}{\delta} I(0 \le Y(X-c) \le \delta)|Z = z\Big). && (3.6)\end{aligned}$$

Note that $E(\frac{\delta - Y(X-c)}{\delta} I(0 \le Y(X-c) \le \delta)|Z = z)$ is decreasing in δ, and approaches 0 when $\delta \to 0$. Furthermore, $E(L_{01}(Y(X-c))|Z = z) - E(L_{01}(Y(X-c^*(z)))|Z = z) = \int_{c^*(z)}^{c}(2p_z(x) - 1)f_z(x)dx$, which is increasing in c when $c > c^*(z)$. Therefore, there exist $\delta_u(z) > 0$ and $c_u(z)$ such that

$$\int_{c^*(z)}^{c_u(z)} (2p_z(x) - 1)f_z(x)dx \ge E\Big(\frac{\delta_u(z) - Y(X-c)}{\delta_u(z)} I(0 \le Y(X-c) \le \delta_u)|Z = z\Big).$$

This implies that for any $\delta < \delta_u(z)$, $\mathrm{argmin}_c E\Big(L_\delta(Y(X-c))|Z = z\Big) \le c_u(z)$. Similarly, there exist $\delta_l(z)$ and $c_l(z)$ such that for any $\delta < \delta_l(z)$, $\mathrm{argmin}_c E\Big(L_\delta(Y(X-c))|Z = z\Big) \ge c_l(z)$.

Therefore, for any $\delta < \min\{\delta_l(z), \delta_u(z)\}$, $\mathrm{argmin}_c E\Big(L_\delta(Y(X-c))|Z=z\Big)$ must lie in a compact set $\mathcal{D}(z)$ around $c^*(z)$.

The second term on the right hand side of (Equation 3.6) is bounded below by 0 and above by $P\Big(|X-c| \le \delta|Z=z\Big)$ and is decreasing in δ. Therefore, by Dini's theorem, $P\Big(|X-c| \le \delta|Z=z\Big)$ converges to 0 uniformly over $\mathcal{D}(z)$ as $\delta \to 0$. It further implies that $E\Big(L_\delta(Y(X-c))|Z=z\Big)$ converges to $E\Big(L_{01}(Y(X-c))|Z=z\Big)$ uniformly over $\mathcal{D}(z)$ as $\delta \to 0$. This, together with the fact that $E\Big(L_{01}(Y(X-c))|Z=z\Big)$ is convex in c, implies that

$$\mathrm{argmin}_c E\Big(L_\delta(Y(X-c))|Z=z\Big) \longrightarrow \mathrm{argmin}_c E\Big(L_{01}(Y(X-c))|Z=z\Big) = c^*(z),$$

when δ converges to zero. □

Note that the assumption that $f(x)$ is continuous can be relaxed. For example, if $f(x)$ is semi-continuous and then the convergence property can be shown by similar proof.

With the ψ_δ-loss, the proposed estimation formulation for the personalized MCID $\hat{c}(z)$ is a solution of

$$\min_{c\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} L_\delta(y_i(x_i - c(z_i))) + \lambda J(c), \tag{3.7}$$

where λ is a tuning parameter, $J(c)$ is a penalty term to avoid overfitting, and $\mathcal{F}$ is a candidate functional space. Here we set $\mathcal{F}$ as a reproducing kernel Hilbert spaces (RKHS; Wahba, 1990), and the final estimation formulation becomes

$$\min_{c\in\mathcal{H}_K} \frac{1}{n}\sum_{i=1}^{n} L_\delta(y_i(x_i - c(z_i))) + \frac{\lambda}{2}\|c\|^2_{\mathcal{H}_K}, \tag{3.8}$$

where $\mathcal{H}_K$ is the RKHS induced by some pre-specified kernel function $K(\cdot,\cdot)$, and $J(c) = \frac{1}{2}\|c\|^2_{\mathcal{H}_K}$ is the associated RKHS norm of $c(z)$. It follows from the representer theorem (Wahba, 1990) that the solution to (Equation 3.8) is of the form $\hat{c}(z) = b + \sum_{i=1}^n w_i K(z_i, z)$, and thus $\|c\|^2_{\mathcal{H}_K} = w^T K w$ with $w = (w_1, \cdots, w_n)^T$ and $K = (K(z_i, z_j))_{i,j=1}^n$.

3.2 Non-convex optimization

The loss function in (Equation 3.7) is non-convex, and thus we employ the difference convex algorithm (DCA; An and Tao, 1997) to tackle the non-convex optimization. The key idea of the DCA is to decompose the non-convex cost function into the difference of two convex functions, and then construct a sequence of subproblems by approximating the second convex function with its affine minorization function.

In particular, the ψ_δ-loss is decomposed as

$$L_\delta(u) = \min\left(\frac{1}{\delta}(\delta - u)_+, 1\right) = \frac{1}{\delta}(\delta - u)_+ - \frac{1}{\delta}(-u)_+.$$

Then the cost function in (Equation 3.7) can be decomposed as $s(w) = s_1(w) - s_2(w)$, where

$$\begin{aligned}
s(w) &= \frac{1}{n}\sum_{i=1}^n L_\delta(y_i(x_i - c(z_i))) + \frac{\lambda}{2}\|c\|^2_{\mathcal{H}_K}, \\
s_1(w) &= \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{\delta}\left(\delta - y_i(x_i - c(z_i))\right)_+\right) + \frac{\lambda}{2}\|c\|^2_{\mathcal{H}_K}, \\
s_2(w) &= \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{\delta}\left(-y_i(x_i - c(z_i))\right)_+\right),
\end{aligned}$$

and w is the coefficient vector for the RKHS representation of $c(z)$.

Next, the DCA constructs a sequence of decreasing upper envelop of $s(w)$ by approximating $s_2(w)$ with its affine minorization function,

$$s_2(w^{(k)}) + \langle w - w^{(k)}, \nabla s_2(w^{(k)})\rangle,$$

where $w^{(k)}$ is the estimated w at the k-th iteration, and $\nabla s_2(w^{(k)})$ is the subgradient of $s_2(w)$ at $w^{(k)}$. The updated $w^{(k+1)}$ is then obtained by solving

$$w^{(k+1)} = \mathtt{argmin}_w \; s_1(w) - s_2(w^{(k)}) - \langle w - w^{(k)}, \nabla s_2(w^{(k)})\rangle. \tag{3.9}$$

Since $s_2(w)$ is a nonconcave function,

$$s_1(w) - s_2(w) \leq s_1(w) - s_2(w^{(k)}) - \langle w - w^{(k)}, \bigtriangledown s_2(w^{(k)})\rangle,$$

which means these subproblems form a sequence of non-increasing upper approximations to the original minimization problem. The updating scheme is iterated until convergence. Although the DCA cannot guarantee global optimum, it delivers a superior numerical performance as demonstrated in the extensive simulation study in Liu et al. (2005).

For illustration, when the kernel function is set as linear, that is $c(z) = w^T z + b$ and denote $\tilde{w} = (w, b)$, the subgradient of $s_2(\tilde{w})$ is expressed as

$$\nabla s_2(\tilde{w}) = \frac{\partial \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\delta}\left(-y_i(x_i - c(z_i))\right)_+\right)}{\partial \tilde{w}} = \frac{1}{n\delta}\sum_{j\in D}(y_j z_{j_1}, \ldots, y_j z_{j_p}, y_j),$$

where $D = \{j : y_j(x_j - c(z_j)) < 0\}$ and p is the dimension of z.

Since the w-unrelated terms do not effect the solution of (Equation 3.9), then $(k+1)$th subproblem is defined as minimizing

$$s_1(\tilde{w}) - \langle \tilde{w}, \nabla s_2(\tilde{w}^{(k)}) \rangle = s_1(\tilde{w}) - \frac{1}{n\delta} \sum_{j \in D} y_j(\tilde{w}^T z_j + b_j).$$

Note that the last expression does not contain $w^{(k)}$, however, $w^{(k)}$ is applied when we derive D in each iteration.

3.3 Asymptotic theory

This section quantifies the asymptotic behavior of $\hat{c}(z)$ in estimating the personalized MCID. The estimation accuracy of $\hat{c}(z)$ is measured by

$$e(\hat{c}, c^*) = E\Big(L_{01}(Y(X - \hat{c}(Z))) - L_{01}(Y(X - c^*(Z)))\Big).$$

Denote $e_{\delta_n}(\hat{c}, c^*) = E\Big(L_{\delta_n}(Y(X - \hat{c}(Z))) - L_{\delta_n}(Y(X - c^*(Z)))\Big)$ with $\delta_n > 0$, where the parameters δ and λ are rewritten as δ_n and λ_n to denote their dependency on n. We make the following four technical assumptions.

Assumption A. For some positive sequence $s_n \to 0$ as $n \to \infty$, there exists $c_0(z) \in \mathcal{F}$, such that for sufficiently small δ_n, $e_{\delta_n}(c_0, c^*) \leq s_n$. That is, $\inf_{\{c \in \mathcal{F}\}} e_{\delta_n}(c, c^*) \leq s_n$.

Assumption A describes the approximation error of $\mathcal{F}$ in approximating $c^*(z)$.

Assumption B. There exist constants $0 < \alpha_2 < +\infty$ and $a_3 > 0$ such that for any given z, $P(|p_z(X) - p_z(c^*(z))| \leq \xi) \leq a_3 \xi^{\alpha_2}$ for sufficiently small ξ.

Assumption B is the low noise assumption that describes the distribution of the diagnostic outcome X in the neighborhood of $c^*(z)$.

Assumption C. There exist constants $0 < \gamma_2 < +\infty$ and $a_4 > 0$ such that for any given z, $\sup_{|x-c^*(z)|\leq\xi} |p_z(x) - p_z(c^*(z))| \leq a_4\xi^{\gamma_2}$ for sufficiently small ξ.

Assumption C is a Hölder condition that describes the behavior of $p_z(x)$ around $c^*(z)$.

Before specifying Assumption D, we first define the metric entropy for any given set. For a given class $\mathcal{B}$ of subsets of S and any $\epsilon > 0$, $\{(G_1^l, G_1^u, \cdots, G_m^l, G_m^u)\}$ forms an ϵ-bracketing set of $\mathcal{B}$ if for any $G \in \mathcal{B}$ there is a j such that $G_j^l \subset G \subset G_j^u$ and $\max_{\{1\leq j\leq m\}} d(G_j^u, G_j^l) \leq \epsilon$, where $d(\cdot,\cdot)$ is a distance for any two subsets in S defined as $d(G_1, G_2) = Pr(G_1 \Delta G_2)$ and $G_1 \Delta G_2 = (G_1 \backslash G_2) \bigcup (G_2 \backslash G_1)$. Then the metric entropy $H(\epsilon, \mathcal{B})$ of $\mathcal{B}$ is defined as the logarithm of the cardinality of the ϵ-bracketing set of $\mathcal{B}$ of the smallest size. More details and examples could be found in Section 19.2 of Van der Vaart (1998). Let

$$\mathcal{G}(k) = \{G_c = \{(x,z) : x - c(z) \geq 0\}, c \in \mathcal{F}, J(c) \leq k\} \subset \mathcal{G}(\mathcal{F}) = \{G_c = \{(x,z) : x - c(z) \geq 0\}, c \in \mathcal{F}, J(c) < +\infty$$

Assumption D. For positive constants a_5, a_6 and a_7, there exists some $\epsilon_n > 0$ such that

$$\sup_{\{k\geq 1\}} \phi(\epsilon_n, k) \leq a_5 n^{1/2},$$

where $\phi(\epsilon_n, k) = \int_{a_7 L}^{(8a_6)^{1/2} L^{\alpha/2(\alpha+\gamma)}} H^{1/2}(u^2/2, \mathcal{G}(k)) du/L$ and $L = L(\epsilon_n, C, k) = \min(\epsilon_n^2 + \lambda_n J_0(k/2 - 1), 1)$.

Theorem 2 *Suppose that Assumptions A-D are met. For the estimated personalized MCID* $\hat{c}(z)$*, there exists a constant* $a_8 > 0$ *such that*

$$Pr(e(\hat{c}, c^*) \geq \beta_n^2) \leq 3.5 \exp\left(-a_8 n(\lambda_n J(c_0))^{\frac{\alpha_2+2}{\alpha_2+1}}\right),$$

provided that $\beta_n^2 \geq 4\lambda_n \max(J(c_0), 1)$*, where* $\beta_n^2 = \min(\max(\epsilon_n^2, 2s_n + 2a_3 a_4^{\alpha_2} \delta_n^{\alpha_2 \gamma_2}), 1)$ *and* δ_n *is a sufficiently small sequence that goes to 0.*

Corollary 1 *Under the assumptions of Theorem 2,*

$$e(\hat{c}, c^*) = O_p(\beta_n^2), E|e(\hat{c}, c^*)| = O(\beta_n^2),$$

provided that $n(\lambda_n J(c_0))^{\frac{\alpha_2+2}{\alpha_2+1}}$ *is bounded away from 0. In addiction, if* $f(c^*(z))$ *is lower bounded away from 0, then*

$$|\hat{c}(Z) - c^*(Z)| = O_p(\beta_n^{\frac{2\alpha_2}{\alpha_2+2}}).$$

Proof The first part follows immediately after Theorem 2, and we only prove the second part. For any given $z \in \mathcal{D}_Z$, Assumptions B and C are similar to (Equation 2.5) and (Equation 2.6) in Theorem 1 and yield that for any sufficiently small $\xi > 0$, there exists a constant $C > 0$ such that

$$E\Big(L_{01}(Y(X - (c^*(z) \pm \xi))|Z = z\Big) - E\Big(L_{01}(Y(X - c^*(z)))|Z = z\Big) \geq C\xi^{1+2/\alpha_2},$$

and $E\Big(L_{01}(Y(X-(c^*(z)\pm\xi))|Z=z\Big) - E\Big(L_{01}(Y(X-c^*(z)))|Z=z\Big)$ is monotonically increasing with ξ. Therefore,

$$\begin{aligned} e(\hat{c},c^*) &\geq E\Big(\{L_{01}(Y(X-\hat{c}(Z))) - L_{01}(Y(X-c^*(z)))\}I(|\hat{c}(Z)-c^*(Z)|\geq\xi)\Big) \\ &\geq E\Big(\{L_{01}(Y(X-(c^*(Z)+\xi sign(\hat{c}(Z)-c^*(Z)))) \\ &\qquad -L_{01}(Y(X-c^*(z)))\}I(|\hat{c}(Z)-c^*(Z)|\geq\xi)\Big) \\ &\geq C\xi^{1+2/\alpha_2}E(I(|\hat{c}(Z)-c^*(Z)|\geq\xi)) \\ &= C\xi^{1+2/\alpha_2}P(|\hat{c}(Z)-c^*(Z)|\geq\xi), \end{aligned}$$

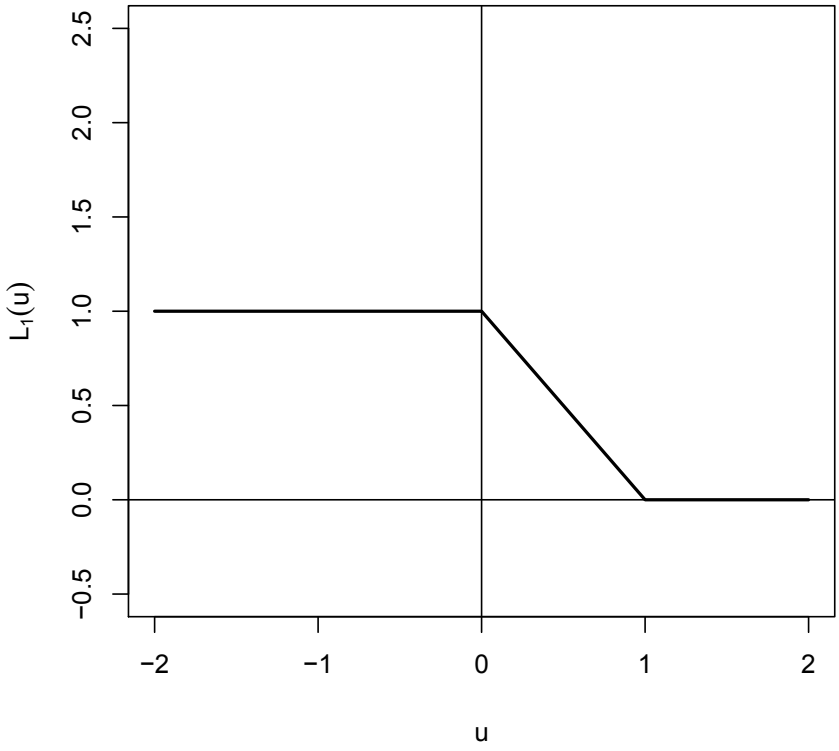
which, together with the first part, implies $P(|\hat{c}(Z)-c^*(Z)|\geq\xi) = O_p(\beta_n^2\xi^{-(1+2/\alpha_2)})$. The desired result follows immediately. □

For illustration, consider an example with a linear personalized MCID $c^*(Z)=Z$, where Z follows $Unif(a,b)$. For any given z, X is generated from $Unif(z-1,z+1)$ and $p_z(x)=P_z(X\leq x)$. We consider a class of linear function $\mathcal{F}=\{z\in(a,b):c(z)=w\cdot z+b, w\in\mathcal{R}\}$, then Assumption A is satisfied with $s_n=1/n$. Assumption B and Assumption C are also satisfied with $\alpha_2=1$ and $\gamma_2=1$. For Assumption D, there is $H(\epsilon,\mathcal{G}(k))\leq O(\log(1/\epsilon))$ since for any $G_c\in\mathcal{G}(k)$, $G^u=G_{c-\epsilon/2}$ and $G^l=G_{c+\epsilon/2}$ could be a bracket of G_c. Therefore, an ϵ-bracketing with size $O((1/\epsilon)^2)$ could be constructed by connecting grid points of $D_{c(a)}$ and $D_{c(b)}$, which are the ranges of $c(a)$ and $c(b)$ for $c\in\mathcal{G}(k)$. Applying the same technique in Shen et al. (2003), there is $\sup_{k\geq1}\leq C(\log(1/\epsilon_n))^{1/2}/\epsilon_n$ for some constant C, which yields a rate $\epsilon_n=(\log n/n)^{1/2}$ and $e(\hat{c},c^*)=O_p(\log n/n)$, when $\beta_n^2=O(\log n/n)$.

Theorem 2 and Corollary 1 develop upper bounds for the misclassification error induced by the estimated $\hat{c}(z)$. The convergence rate β_n^2 in Corollary 1 depends on the value of δ_n, ϵ_n^2, s_n and λ_n. In particular, when $\delta_n = O(s_n^{1/(\alpha_2\gamma_2)})$, the convergence rate becomes $O_p(\min(\max(\epsilon_n^2, 2s_n), 1))$. More importantly, Corollary 1 also establishes asymptotic convergence rate for the estimation of the personalized MCID $c^*(z)$. Such a result cannot be established for the standard classification function $g(x, z)$ due to its lack of explicit estimation of $c^*(z)$.

In this chapter, the personalized MCID is proposed which allows the MCID to vary according to each patient's clinical profiles. A novel surrogate loss ψ_δ-loss is employed to overcome the difficulty in estimation and the estimation scheme is also developed. The effectiveness of our proposed method will be demonstrated in the following two chapters.

Figure 1. Plot of ψ_1-loss.



CHAPTER 4

SIMULATION

This chapter examines the proposed estimation methods for estimating MCID using simulated examples. Two scenarios are considered. Scenario I focuses on the population-based MCID for all patients, and scenario II focuses on the personalized MCID that varies among patients and relies on each patient clinical profile. To assess the estimation performance, we report the estimated MCID as well as the misclassification error (MCE) based on the testing set, which is defined as

$$\mathsf{MCE}(\hat{c}) = \frac{1}{n_{test}} \sum_{i \in \text{testing set}} I(y_i \neq \mathsf{sign}(x_i - \hat{c}(z_i))),$$

where n_{test} denotes the size of the testing set, and $\hat{c}(z_i) = \hat{c}$ for the population-based MCID.

4.1 Scenario I: population-based MCID

Two simulated examples are examined.

Example 1 A random sample $\{(X_i, Y_i); i = 1, \cdots, n + 2000\}$ is generated as follows. First, X_i is generated from $\mathsf{Unif}(-1, 1)$ and then Y_i is generated from $\mathsf{Bern}((x_i + 1)/2)$. Next, a sample of size n is randomly selected for training and the remaining 2000 samples are allocated for testing.

Example 2 A random sample $\{(X_i, Y_i); i = 1, \ldots, n + 2000\}$ is generated as follows. First, X_i is generated from the mixture of two Gaussian distributions $0.7\mathsf{N}(-1, 1) + 0.3\mathsf{N}(1, 1)$ and

then Y_i is generated from $\mathrm{Bern}(F(x_i))$, where $F(x_i) = \Pr(X \leq x_i)$. Next, a sample of size n is randomly selected for training and the remaining 2000 samples are allocated for testing.

In both examples, the training sizes are set as $n = 250$, 500 and 1000. Both examples are replicated 100 times. The averaged performance measures of our proposed method and Shiu and Gatsonis (2008) are reported in Table I. In addition, the ideal MCID's and their corresponding misclassification errors are used as baseline for the comparison in Table I.

In both examples, our proposed method yields accurate MCID estimates that are very close to the ideal MCID's. The resulting MCE's are also close to the MCE's produced by using the ideal MCID's. The performance of the method by Shiu and Gatsonis appears to be less competitive. Even with a large sample size $n = 1000$, their estimated MCID's are still considerably different from the ideal MCID's.

4.2 Scenario II: personalized MCID

For personalized MCID, the MCE by using our proposed method with linear and Gaussian kernels are examined. The linear kernel is defined as $K(z_1, z_2) = z_1^T z_2$, and the Gaussian kernel is defined as $K(z_1, z_2) = e^{-\|z_1 - z_2\|^2/2\sigma^2}$, where the scale parameter σ^2 is set as the median of pairwise Euclidean distances within the training set. To optimize the performance of our proposed method, a grid search by 5-fold cross validation is employed to select the tuning parameter λ. The grid for all examples is set as $\{10^{(s-31)/10}; s = 1, \cdots, 61\}$. For illustration, three simulated examples are examined with $\delta = 0.1$.

Example 3. A random sample $\{(X_i, Y_i, Z_i); i = 1, \cdots, n\}$ is generated as follows. First, Z_i's are independently generated from $N_2(\mu, I_2)$ with $\mu = (0, 0)^T$. Second, X_i's are independently

generated from $N(b+w^T z_i, 1)$, where $b = 0$ and $w = (1,2)^T$. Next, the response Y_i is generated from $Bern(F(x_i))$, where $F(x_i) = Pr(X_i \leq x_i)$.

Example 4. A random sample $\{(X_i, Y_i, Z_i); i = 1, \cdots, n\}$ is generated as follows. First, Z_i's are independently generated from $N_2(\mu, I_2)$ with $\mu = (0,0)^T$. Second, X_i's are independently generated from $N(b + w^T z_i - w^T z_i^2, 1)$, where $b = 0$ and $w = (1,2)^T$. Next, the response Y_i is generated from $Bern(F(x_i))$, where $F(x_i) = Pr(X_i \leq x_i)$.

Example 5. A random sample $\{(X_i, Y_i, Z_i); i = 1, \cdots, n\}$ is generated as follows. First, Z_i's are independently generated from $N_3(\mu, I_3))$ with $\mu = (0,0,0)^T$. Second, X_i's are independently generated from $N(b + \cos(w^T z_i), 1)$, where $b = 0$ and $w = (1, 1.5, 2)^T$. Next, the response Y_i is generated from $Bern(F(x_i))$, where $F(x_i) = Pr(X_i \leq x_i)$.

For each example, the training sizes are set as 100, 250, 500 and testing size is set as 2000. All examples are replicated 50 times, and the averaged test errors are reported in Table II.

Our proposed method delivers satisfactory performance in estimating the personalized MCID in all three examples. In addition, the linear kernel yields slightly better performance than the Gaussian kernel in Example 3 as the true boundary is linear and the estimated MCID with linear kernel $\hat{c}_L$ and with Gaussian kernel $\hat{c}_G$ are displayed in Figure 2. The linear kernel is outperformed by the Gaussian kernel in the other two examples with nonlinear boundaries. Therefore, the Gaussian kernel would be suggested if no prior knowledge about the boundary is available.

For estimating the personalized MCID, the choice of δ may impact the performance of our proposed method. By Theorem 2, large δ leads to less accurate prediction while computational instability may occur when small δ is used for the estimation.

For illustration, we conducted a sensitivity analysis on the values of δ in a random replication of Example 1 with training size 250. The estimated coefficients and prediction error as functions of δ are displayed in Figure 3. It is evident that when δ is too large, the estimation of $\mathsf{c}(z)$ moves away from the truth and yields a larger error rate. When δ is close to 0, the error rate and estimation of $\mathsf{c}(z)$ are relatively stable. Therefore, we recommend to set δ as 0.1 for simplicity.

TABLE I

SIMULATION I. AVERAGED MCID AND THE MISCLASSIFICATION ERROR (MCE) AND THEIR STANDARD ERRORS (IN PARENTHESES) FOR OUR METHOD (OUR) AND THE METHOD BY SHIU AND GATSONIS (SG) BASED ON 100 REPLICATIONS. THE IDEAL PERFORMANCE IS INCLUDED AS THE BASELINE FOR COMPARISON.

		n=250	n=500	n=1000	Ideal
		Example 1			
MCID	OUR	0.055(0.0116)	-0.021(0.0058)	0.004(0.0032)	0.000
	SG	0.078(0.0387)	-0.065(0.0290)	-0.080(0.0222)	
MCE	OUR	0.260(0.0010)	0.255(0.0005)	0.253(0.0003)	0.250
	SG	0.344(0.0045)	0.355(0.0033)	0.374(0.0024)	
		Example 2			
MCID	OUR	-0.563(0.0187)	-0.496(0.0095)	-0.497(0.0056)	-0.514
	SG	-0.436(0.0827)	-0.286(0.0676)	-0.370(0.0526)	
MCE	OUR	0.257(0.0009)	0.253(0.0005)	0.252(0.0003)	0.250
	SG	0.338(0.0043)	0.361(0.0033)	0.374(0.0024)	

TABLE II

SIMULATION II. ESTIMATED MEANS AND STANDARD DEVIATIONS (IN PARENTHESES) OF THE MISCLASSIFICATION ERROR BY USING OUR PROPOSED METHOD WITH LINEAR AND GAUSSIAN KERNELS BASED ON 50 REPLICATIONS.

	n=100	n=250	n=500	Ideal
		Example 1		
Linear	0.256(0.0119)	0.254(0.0112)	0.250(0.0108)	0.250
Gaussian	0.280(0.0177)	0.270(0.0146)	0.259(0.0130)	
		Example 2		
Linear	0.412(0.0146)	0.408(0.0140)	0.408(0.0095)	0.250
Gaussian	0.290(0.0169)	0.274(0.0133)	0.260(0.0118)	
		Example 3		
Linear	0.315(0.0132)	0.313(0.0129)	0.318(0.0103)	0.250
Gaussian	0.323(0.0182)	0.308(0.0122)	0.293(0.0109)	

Figure 2. The estimated MCID with linear kernel $\hat{c}_L$ and with Gaussian kernel $\hat{c}_G$ in a randomly selected replication of Example 3 when $n = 250$ and $Z_2 = 0$.

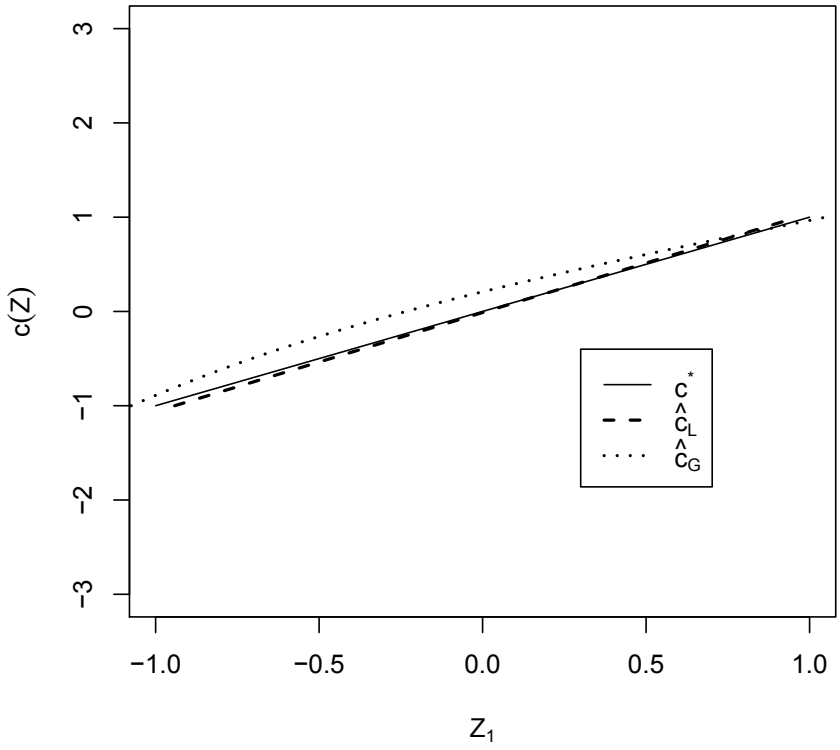
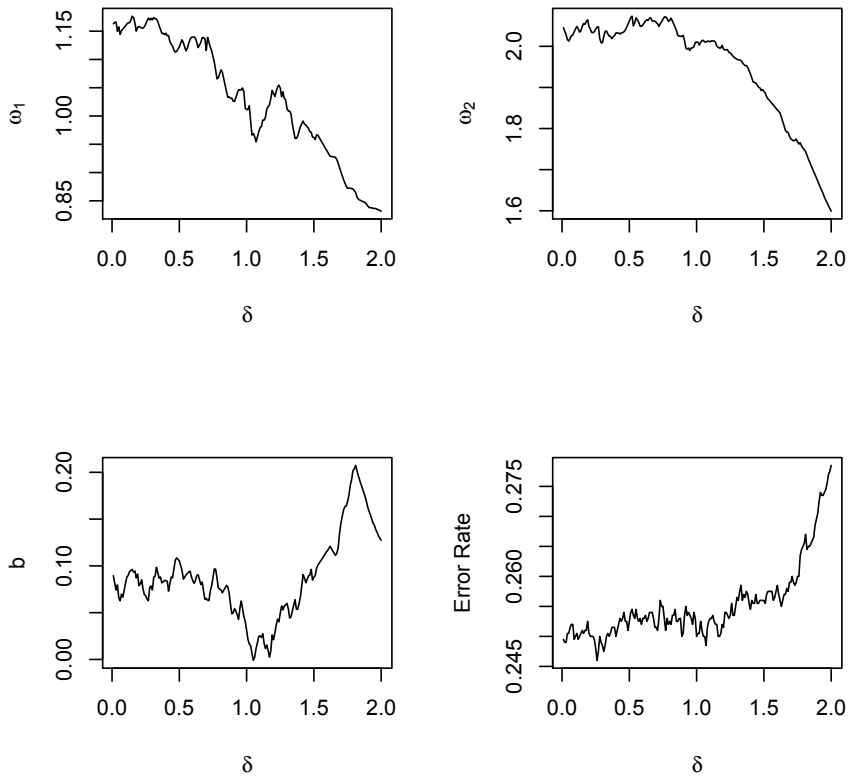


Figure 3. Sensitivity analysis of δ in a randomly selected replication of Example 3 with $n = 250$.



CHAPTER 5

REAL APPLICATIONS

In this chapter, our proposed method is applied to two benchmark datasets and two phase-3 clinical trial datasets. The benchmark datasets are the breast cancer Wisconsin (diagnostic) dataset (WDBC) and Parkinsons disease dataset (PD) which are publicly available at the University of California Irvine Machine Learning Repository (`http://archive.ics.uci.edu/ml/`). The phase-3 clinical trial datasets are a woman heavy menstrual blood loss dataset (WHMBL) and a hot flush dataset (Hot Flush).

5.1 Benchmark examples

The WDBC dataset contains 10 continuous variables collected from 569 patients and one binary response indicating whether the cancer tumor is malignant or benign. For each variable, mean, standard error and mean of the three largest values (worst) are computed. Therefore, the dataset consists of 30 covariates for each patient. The 569 patients are randomly split into a training set of 280 patients and a test set of 289 patients. The PD dataset consists of 22 continuous covariates which characterize 195 voice recordings from 31 persons. Among the 31 persons, 23 of them have Parkinson's disease and the others are healthy. The 195 voice recordings are randomly split into the training set of 100 recordings and the test set of 95 recordings.

For these two benchmark datasets, there is no diagnostic measurement. For illustration, pseudo diagnostic measurements are generated as follows. For WDBC data, we randomly selected three covariates and fit a linear classifier, and the resulting classification function values are used as the pseudo diagnostic measurements. For PD data, we also randomly selected three covariates and fit a nonlinear classifier with the Gaussian kernel, and the resulting classification function values were used as the pseudo diagnostic measurements.

For simplicity, our proposed method with $\delta = 0.1$ is employed. The tuning parameter λ is selected as in Section 4.2. Each example is replicated 50 times, and the averaged MCE using the method by Shiu and Gatsonis method, the population-based MCID, and the personalized MCID with the linear and Gaussian kernels are summarized in Table III.

In both examples, our personalized MCID outperforms the population-based MCID and the method by Shiu and Gatsonis. This supports the application of personalized MCID in clinical trials in order to provide more accurate estimates of patients' satisfaction.

5.2 WHMBL and hot flush clinical trials

The WHMBL clinical trial aims to develop a treatment for reducing the amount of blood loss during a menstrual cycle in excessive bleeding women. The primary efficacy variable is the change from baseline in blood loss volume. The blood loss of each patient is measured per menstrual cycle and the PROs are collected based on a questionnaire answered by each patient at a post-treatment visit. The WHMBL trial dataset consists of 481 patients administered either placebo or active doses.

Patient profile contains the information of age, body mass index (BMI), alcohol (Yes/No), tobacco (Yes/No) and baseline value of blood loss. The 481 patients were randomly split into a training set of 240 patients and a test set of 241 patients.

The hot flush clinical trial aims to develop a treatment for reducing hot flush in women due to menopause. The hot flush clinical trial dataset consists of 1684 patients administered either placebo or active doses.

Patient profile contains the information for age, BMI, race and baseline hot flushes. 300 patients were selected randomly to form the training set and the remaining 1384 patients were used as the testing set.

Here, $\delta = 0.1$ is used for simplicity and the tuning parameter λ is selected as in Section 4.2. Each example is replicated 50 times, and Table IV summarizes the averaged performance measures of the method by Shiu and Gatsonis, the population-based MCID, and the personalized MCID with the linear and Gaussian kernels.

In both scenarios, our proposed method delivers competitive performance in comparison with the method by Shiu and Gatsonis. In WHMBL trial, the method by Shiu and Gatsonis yields a negative MCID which is clinically misleading. It is also interesting to notice that for the WHMBL trial, personalized MCID yields larger MCE when compared with population-based MCID. It could be due to the homogeneity among the enrolled patients. For the hot flush trial, patients' satisfaction on treatment effect is more accurately estimated when the clinical profiles are included. A closer investigation of the fitted classification function implies that patients' satisfaction is highly affected by the baseline hot flushes. This is reasonable as patients with

higher baseline hot flushes tend to expect better treatment effect. In addition, when the actual change is replaced by the proportions of change in hot flush trial, the improvement of personalized MCID over population-based MCID diminishes. Therefore, in this trial personalized MCID in absolute change improves the agreement and the improvement becomes smaller when using percent change from baseline.

TABLE III

BENCHMARK EXAMPLES. ESTIMATED MEANS AND STANDARD DEVIATIONS (IN PARENTHESES) OF THE MISCLASSIFICATION ERROR (MCE) BY USING THE METHOD BY SHIU AND GATSONIS (SG), THE POPULATION-BASED MCID (OUR), THE PERSONALIZED MCID WITH LINEAR KERNEL (OUR_{L}) AND GAUSSIAN KERNEL (OUR_{G}) BASED ON 50 REPLICATIONS.

	SG	OUR	OUR_{L}	OUR_{G}
WDBC	0.140(0.0013)	0.129(0.0010)	0.038(0.0180)	0.053(0.0210)
PD	0.224(0.0035)	0.172(0.0036)	0.170(0.0357)	0.147(0.0421)

TABLE IV

REAL APPLICATIONS. AVERAGED MCID AND MISCLASSIFICATION ERROR (MCE) AND THEIR STANDARD ERRORS(IN PARENTHESIS) BY USING THE METHOD BY SHIU AND GATSONIS (SG), THE POPULATION-BASED MCID (OUR), THE PERSONALIZED MCID WITH LINEAR KERNEL (OUR_L) AND GAUSSIAN KERNEL (OUR_G) BASED ON 50 REPLICATIONS.

	SG	OUR	OUR_L	OUR_G
		WHMBL		
MCID	-45.004(3.3011)	20.610(0.4905)	-	-
MCE	0.436(0.0016)	0.358(0.0014)	0.365(0.0186)	0.376(0.0185)
		Hot Flush		
MCID	5.426(0.4453)	6.060(0.0229)	-	-
MCE	0.399(0.0049)	0.282(0.0005)	0.260(0.0054)	0.268(0.0031)

CHAPTER 6

CONCLUSION AND FUTURE RESEARCH

6.1 Conclusion

In this dissertation, a general framework for defining and estimating population-based and personalized MCID's is proposed. The concept of MCID has attracted much attention in clinical trials, while little statistical work has been done for appropriately modeling MCID. Our proposed method unifies both population-based and personalized MCID's into a large margin classification framework, and delivers superior estimation performance in both simulated examples and real applications to benchmark datasets and two phase-3 clinical trials. More importantly, the asymptotic properties of our proposed method are established for both population-based and personalized MCID's.

6.2 Future research

6.2.1 The Youden index and optimal cut-point

New and sophisticated biomarkers have been popularly developed for identifying patients with or without certain disease and widely applied for early detection and prevention of chronic and acute diseases. Therefore, it is of great importance to evaluate identification effectiveness of biomarkers. The Youden index (Youden, 1950) is commonly used to measure biomarker's overall diagnostic effectiveness (Schisterman et al., 2005).

Formally, the Youden index is defined as the maximum vertical distance between the receiver operating characteristics (ROC) curve and the chance line (Figure 4). Suppose that every observation contains a continuously supported diagnostic measurement X, and a binary disease status $Y \in \{-1, 1\}$, where $Y = 1$ denotes a positive status and $Y = -1$ otherwise. A cut point c is introduced so that positive status is predicted if $X \geq c$ and negative otherwise. Mathematically, the Youden index is defined as

$$J = \max\{sen(c) + spe(c) - 1\},$$

where $sen(c) = Pr(X \geq c|Y = 1)$ is the sensitivity and $spe(c) = Pr(X < c|Y = -1)$ is the specificity. The optimal cut-point c^* is the point c that yields J,

$$argmax_c\{Pr(X > c|Y = 1) + Pr(X \leq c|Y = -1) - 1\}. \tag{6.1}$$

Direct derivation yields that (Equation 6.1) is equivalent to

$$\max_c E\left(w(Y)(1 + Ysign(X - c))\right), \tag{6.2}$$

where $w(1) = 1/\pi$ and $w(-1) = 1/(1 - \pi)$ with $\pi = Pr(Y = 1)$, $sign(u) = 1$ if $u \geq 0$ and -1 otherwise. Moreover, its solution is the same as that of

$$\max_c E_X\left(\frac{sign(X - c)}{\pi(1 - \pi)}(p(X) - \pi)\right). \tag{6.3}$$

Lemma 4 *Assume that* $p(x) = Pr(Y = 1|X = x)$ *is increasing in* x*, then the solution of (Equation 6.2) satisfies that* $p(c^*) = \pi$.

Lemma 4 follows immediately after (Equation 6.3) and describes the optimal-cut when $p(x)$ is known, which is analogous to MCID in (Equation 2.2). Note that it is reasonable to assume that $p(x)$ is increasing in x since risk of disease is expected to increase with an effective biomarker level. Schisterman et al. (2005) showed that the optimal-cut occurs at the intersection between probability density functions of cases ($Y = 1$) and controls ($Y = -1$), that is $f(c^*|Y = 1) = f(c^*|Y = -1)$. Direct derivation yields that

$$\frac{p(c^*)}{Pr(Y = 1)} = \frac{1 - p(c^*)}{Pr(Y = -1)},$$

which is equivalent to the conclusion in Lemma 4.

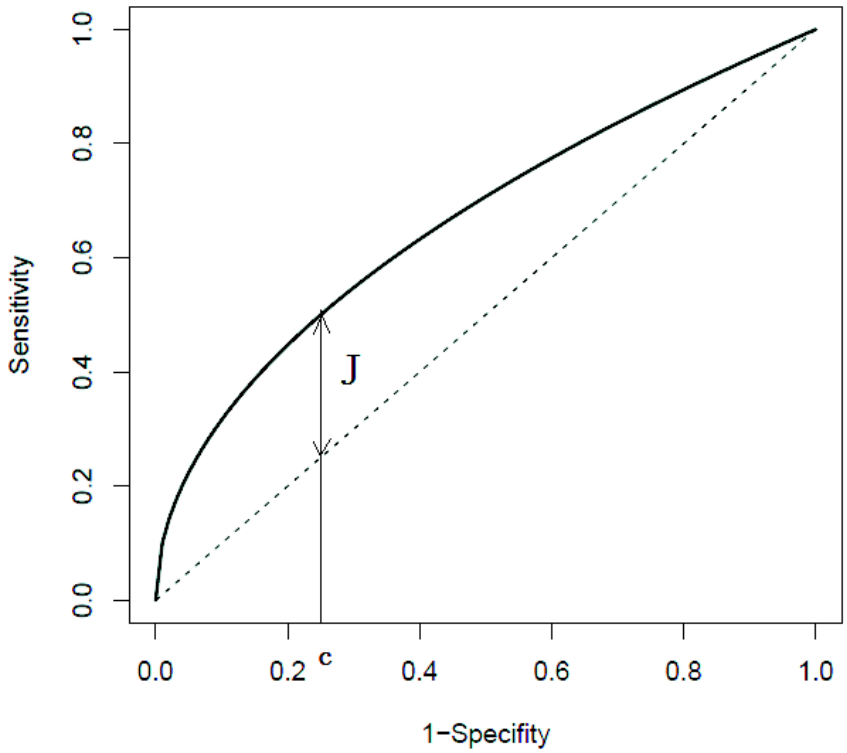
We can formulate the estimation of optimal cut-point as a large margin classification problem similarly as for the estimation of MCID. It is also interesting to extend the framework to personalized optimal cut-point which allows individualized cut-point value for each patient since their clinical profiles may indicate their disease status differently. The estimation scheme and asymptotic results will be developed to justify the effectiveness of the proposed method.

6.2.2 Others

Further investigation could focus on the potential issues when applying the proposed methods to various real-life clinical trials. Some interesting further extensions are listed as follows.

- The proposed MCID can be used to claim the success of treatment. Certain type of statistical hypothesis test involving the estimated MCID could be adopted. For instance, Copay et al. (2007) discussed one approach that compares the proportions of patients who reach the MCID between treatment group and placebo group.
- The ordinal type of PRO scores (Jaeschke et al., 1989; Juniper et al., 1994) are commonly used in clinical trails. It is interesting to extend our proposed method to this scenario. An intuitive approach is to dichotomize the PRO score based on certain external criterion or guidance.
- In a situation that only few covariates included in patients' clinical profiles Z affect the patients' judgment, variable selection in Z is very important since it would further help clinical assessment.
- Bayesian type approach could be applied to estimate population-based and personalized MCID.
- Other nonparametric regression models could be employed for estimating personalized MCID.
- The asymptotic properties established in this dissertation provide upper bounds for the convergence rates of population-based and personalized MCID. Future exploration on the optimal rates of convergence would be interesting.

Figure 4. Receiver Operating Characteristic (ROC) curve with the Youden index (J) and optimal cut-point (c) displayed.



APPENDICES

Proof of Theorem 1. We first show that $\hat{c} \xrightarrow{p} c^*$. Let $F_y(x) = P(X \le x, Y = y)$, then

$$\begin{aligned} \frac{1}{2}E\left(1 - Y\mathrm{sign}(X - c)\right) &= P(X \le c, Y = 1) + P(X > c, Y = -1) \\ &= F_1(c) + P(Y = -1) - F_{-1}(c). \end{aligned}$$

By strong law of large number, $\frac{1}{n}\sum_{i=1}^{n} I(Y_i = -1) \xrightarrow{a.s.} P(Y = -1)$. Further, it follows from Theorem 19.1 of Van der Vaart (1998) that

$$\begin{aligned} F_{1,n}(c) &= \frac{1}{n}\sum_{i=1}^{n} I(X_i \le c, Y_i = 1) \xrightarrow{a.s.} F_1(c), \\ F_{-1,n}(c) &= \frac{1}{n}\sum_{i=1}^{n} I(X_i \le c, Y_i = -1) \xrightarrow{a.s.} F_{-1}(c), \end{aligned}$$

uniformly over c. Therefore,

$$\frac{1}{2n}\sum_{i=1}^{n}\left(1 - y_i\mathrm{sign}(x_i - c)\right) \xrightarrow{a.s.} \frac{1}{2}E\left(1 - Y\mathrm{sign}(X - c)\right)$$

uniformly over c. Also by Lemma 1, $\frac{1}{2}E\left(1 - Y\mathrm{sign}(X - c)\right)$ has a unique minimizer c^* when $p(x)$ is continuous and strictly increasing in x. The desired asymptotic consistency follows immediately after Theorem 5.7 of Van der Vaart (1998).

Next, we establish the convergence rate of $|\hat{c} - c^*|$ by using Theorem 5.52 of Van der Vaart (1998). We just need to verify the necessary assumptions. Note that c^* is the minimizer of $\frac{1}{2}E(1 - ysign(x - c))$. Without loss of generality, for any $c > c^*$, direct deviation yields that

$$\begin{aligned} E(m_c(X,Y) - m_{c^*}(X,Y)) &= P(c^* \le X < c, Y = 1) - P(c^* \le X < c, Y = -1) \\ &= \int_{c^*}^{c} p(x)f(x)dx - \int_{c^*}^{c} (1 - p(x))f(x)dx \\ &= \int_{c^*}^{c} (2p(x) - 1)f(x)dx, \end{aligned}$$

where $m_c(x, y) = \frac{1}{2}(1 - ysign(x - c))$.

Since $f(x)$ is continuous at c^*, it can be shown that $P(c^* \le X \le c^* + \xi) \ge a_9\xi$ for sufficient small $\xi > 0$, where $a_9 = f(c^*)/2 > 0$. Furthermore, $p(c^* + \xi) - p(c^*) > (a_9/a_1)^{1/\alpha_1}(\xi)^{2/\alpha_1}$, since otherwise there exists $0 < \xi_1 < 1$ such that $p(c^* + \xi_1) - p(c^*) \le (a_9/a_1)^{1/\alpha_1}(\xi_1)^{2/\alpha_1}$, and by assumption (Equation 2.5)

$$a_9\xi_1 \le P(c^* \le X \le c^* + \xi_1) \le P\Big(|p(X) - p(c^*)| \le (a_9/a_1)^{1/\alpha_1}(\xi_1)^{2/\alpha_1}\Big) \le a_9(\xi_1)^2,$$

which leads to a contradiction to the fact that $\xi_1 < 1$.

Since $p(x)$ in continuous in x, there exists $0 < \xi_2 < \xi$ such that $p(c^* + \xi_2) - p(c^*) = (a_9/a_1)^{1/\alpha_1}(\xi)^{2/\alpha_1}$, and then

$$E(m_{c^*+\xi}(X,Y) - m_{c^*}(X,Y))$$

$$
\begin{aligned}
&= \int_{c^*}^{c^*+\xi} (2p(x)-1)f(x)dx > \int_{c^*+\xi_2}^{c^*+\xi} (2p(x)-1)f(x)dx > (a_9/a_1)^{1/\alpha_1}(\xi)^{2/\alpha_1} \int_{c^*+\xi_2}^{c^*+\xi} f(x)dx \\
&= (a_9/a_1)^{1/\alpha_1}(\xi)^{2/\alpha_1}(P(c^* \le X \le c^*+\xi) - P(c^* \le X \le c^*+\xi_2)) \\
&\ge (a_9/a_1)^{1/\alpha_1}(\xi)^{2/\alpha_1}\Big(P(c^* \le X \le c^*+\xi) - P(|p(X)-p(c^*)| \le (a_9/a_1)^{1/\alpha_1}(\xi)^{2/\alpha_1})\Big) \\
&\ge a_9^{1+1/\alpha_1} a_1^{-1/\alpha_1} \xi^{2/\alpha_1}(\xi-\xi^2).
\end{aligned}
$$

It can be shown similarly that

$$
E(m_{c^*-\xi}(X,Y) - m_{c^*}(X,Y)) \ge a_9^{1+1/\alpha_1} a_1^{-1/\alpha_1} \xi^{2/\alpha_1}(\xi-\xi^2).
$$

Therefore, there exists constant $a_{10} > 0$ such that for sufficiently small $\xi > 0$,

$$
\sup_{|c-c^*|<\xi} E(m_c(X,Y) - m_{c^*}(X,Y)) \ge a_{10}\xi^{1+2/\alpha_1}. \tag{.4}
$$

Furthermore, denote $\mathcal{F}_m = \{m_c(x,y) - m_{c^*}(x,y) : x \in \mathcal{R},\ y \in \{-1,+1\}\}$. Consider the grid $-\infty = t_1 < t_1 < \cdots < t_k = +\infty$ with $t_{\lceil k/2 \rceil} = c^*$ and $P(x < t_i) - P(x \le t_{i-1}) < \epsilon$ for each t_i. Note that

$$
m_c(x,y) - m_{c^*}(x,y) = \begin{cases} I(c^* \le x < c, y = -1) - I(c^* \le x < c, y = 1), & \text{if } c > c^*, \\ I(c < x \le c^*, y = 1) - I(c < x \le c^*, y = -1), & \text{if } c \le c^*. \end{cases}
$$

Then the functional brackets $[1_{[c^*,t_i]}(x), 1_{[c^*,t_{i+1})}(x)]$ for $i > \lceil k/2 \rceil$ and $[1_{[t_i,c^*]}(x), 1_{(t_{i-1},c^*]}(x)]$ for $i \leq \lceil k/2 \rceil$ forms $L_1(P)$ brackets of size ϵ for $\mathcal{F}_m$ with cardinality $k < 2/\epsilon$. Thus the bracketing number $N_{[\,]}(\epsilon, \mathcal{F}_m, L_2(P)) = O(\epsilon^{-2})$ and then the bracketing integral

$$J_{[\,]}(\eta, \mathcal{F}_m, L_2(P)) = \int_0^\eta \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}_m, L_2(P))}d\epsilon \leq a_{11}\eta \log \eta,$$

for some constant $a_{11} > 0$.

Also $g(x) = I(c^* - \xi \leq x \leq c^* + \xi)$ is an envelop function of $m_c - m_{c^*}$ with $|c - c^*| < \xi$, and then assumptions (Equation 2.5) and (Equation 2.6) imply that

$$\|g\|_{P,2} = (P(|X - c^*| \leq \xi))^{1/2} \leq (P(|p(X) - p(c^*)| \leq a_2\xi^{\gamma_1}))^{1/2} \leq (a_1 a_2^{\alpha_1})^{1/2}\xi^{\alpha_1\gamma_1/2}.$$

By Corollary 19.35 of Van der Vaart (1998),

$$\begin{aligned} E^* \sup_{|c-c^*|<\xi} |G_n(m_c - m_{c^*})| &\leq J_{[\,]}(\|g\|_{P,2}, \mathcal{F}_m, L_2(P)) \leq J_{[\,]}((a_1 a_2^{\alpha_1})^{1/2}\xi^{\alpha_1\gamma_1}, \mathcal{F}_m, L_2(P)) \\ &\leq \frac{1}{2}a_{11}\alpha_1\gamma_1(a_1 a_2^{\alpha_1})^{1/2}\xi^{\alpha_1\gamma_1/2}\log \xi. \end{aligned}$$

Thereupon, denote $A = 1 + 2/\alpha_1 - \alpha_1\gamma_1/2$, it follows from Theorem 5.52 of Van der Vaart (1998) that for a sufficiently large integer M,

$$\begin{aligned} P^*\Big(|\hat{c} - c^*| \geq 2^M(n(\log n)^{-2})^{-1/(2A)}\Big) &\leq \frac{2^{1+2/\alpha_1}}{a_{10}}a_{11}\alpha_1\gamma_1(a_1 a_2^{\alpha_1})^{1/2}\sum_{j\geq M}\Big(\frac{2^{-jA}}{A} - \frac{2^{-jA/2}}{\log n}\Big) \\ &\leq \frac{2^{1+2/\alpha_1}}{a_{10}}a_{11}\alpha_1\gamma_1(a_1 a_2^{\alpha_1})^{1/2}(\log n)^{-1}\frac{2^{-MA/2}}{1 - 2^{-A/2}}. \quad (.5) \end{aligned}$$

Before delving into the proof of Theorem 2, we first define the L_2- metric entropy with bracketing for a function class $\mathcal{F}$. For any $\epsilon > 0$, $\{(l_1^l, l_1^u), \cdots, (l_m^l, l_m^u)\}$ forms an ϵ-bracketing of $\mathcal{F}$, if for any $c \in \mathcal{F}$, there is a j, such that $l_j^l \le l(c, \cdot) \le l_j^u$ and $\max_{\{1 \le j \le m\}} \|l_j^l - l_j^u\|_2 \le \epsilon$, where $\|\cdot\|_2$ is the L_2-norm. Then the L_2-metric entropy of $\mathcal{F}$ with bracketing $H_B(\epsilon, \mathcal{F})$ is defined as a logarithm of the cardinality of the ϵ-bracketing of $\mathcal{F}$ of the smallest size.

Proof of Theorem 2. First we introduce some notations to be used in the proof. Let $\tilde{l}_{\delta_n}(c, D_i) = L_{\delta_n}(y_i(x_i - c(z_i))) + \lambda J(c)$, where $D_i = (x_i, y_i, z_i)$. Similarly, denote $\tilde{l}(c, D_i) = L_{01}(y_i(x_i - c(z_i))) + \lambda J(c)$. Then the scaled empirical process $E_n(\tilde{l}(c, D) - \tilde{l}_{\delta_n}(c_0, D))$ is defined as

$$E_n(\tilde{l}(c, D) - \tilde{l}_{\delta_n}(c_0, D)) = \frac{1}{n} \sum_{i=1}^{n} \Big(\tilde{l}(c, D_i) - \tilde{l}_{\delta_n}(c_0, D_i) - E(\tilde{l}(c, D_i) - \tilde{l}_{\delta_n}(c_0, D_i)) \Big).$$

Since $L_{\delta_n}(y_i(x_i - c(z_i))) \ge L_{01}(y_i(x_i - c(z_i)))$ for any $\delta_n > 0$, we have

$$\tilde{l}_{\delta_n}(c_0, D_i) - \tilde{l}(c, D_i) \ge \tilde{l}_{\delta_n}(c_0, D_i) - \tilde{l}_{\delta_n}(c, D_i).$$

Furthermore, by Assumptions A-C,

$$\begin{aligned} e(c_0, c^*) &= EL_{01}(Y(X - c_0(Z))) - EL_{01}(Y(X - c^*(Z))) \\ &\le e_{\delta_n}(c_0, c^*) + P(|X - c^*(z)| \le \delta_n | Z = z) \\ &\le e_{\delta_n}(c_0, c^*) + P\Big(|p_z(X) - p_z(c^*(z))| \le a_4 \delta_n^{\gamma_2} \Big) \\ &\le s_n + a_3 a_4^{\alpha_2} \delta_n^{\alpha_2 \gamma_2} \le \beta_n^2/2. \end{aligned}$$

Let $\hat{c} = \mathrm{argmin}_{c\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \tilde{l}_{\delta_n}(c, D_i)$ be the estimated personalized MCID, then

$$\begin{aligned}\{e(\hat{c}, c^*) \geq \beta_n^2\} &\subset \left\{\sup_{\{e(c,c^*)\geq\beta_n^2\}} \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{l}_{\delta_n}(c_0, D_i) - \tilde{l}_{\delta_n}(c, D_i)\right) \geq 0\right\} \\ &\subset \left\{\sup_{\{e(c,c^*)\geq\beta_n^2\}} \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{l}_{\delta_n}(c_0, D_i) - \tilde{l}(c, D_i)\right) \geq 0\right\}.\end{aligned}$$

It immediately implies that

$$P(e(\hat{c}, c^*) \geq \beta_n^2) \leq P^*\left(\sup_{\{e(c,c^*)\geq\beta_n^2\}} \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{l}_{\delta_n}(c_0, D_i) - \tilde{l}(c, D_i)\right) \geq 0\right) \triangleq I,$$

where P^* denotes the outer probability measure.

Next, we derive some preliminary results for bounding I. Note that the functional space $\{c \in \mathcal{F} : e(c, c^*) \geq \beta_n^2\}$ can be partitioned as

$$A_{ij} = \{c \in \mathcal{F} : 2^{i-1}\beta_n^2 \leq e(c, c^*) < 2^i\beta_n^2, 2^{j-1}\max(J(c_0), 1) \leq J(c) < 2^j \max(J(c_0), 1)\};$$

$$A_{i0} = \{c \in \mathcal{F} : 2^{i-1}\beta_n^2 \leq e(c, c^*) < 2^i\beta_n^2, J(c) < \max(J(c_0), 1)\},$$

for $i = 1, 2, \cdots$ and $j = 1, 2, \cdots$. Then we need to establish some inequalities on the first and second moments of $\tilde{l}(c, D) - \tilde{l}_{\delta_n}(c_0, D)$ for $c \in A_{ij}$.

For the first moment, note that for any $c \in \mathcal{F}$,

$$E(L_{01}(c, D) - L_{\delta_n}(c_0, D)) = E(L_{01}(c, D) - L_{01}(c^*, D)) + E(L_{01}(c^*, D) - L_{\delta_n}(c_0, D))$$

$$\begin{aligned} &\geq e(c, c^*) + e_{\delta_n}(c^*, c_0) - a_3 a_4^{\alpha_2} \delta_n^{\alpha_2 \gamma_2} \\ &\geq e(c, c^*) - s_n - a_3 a_4^{\alpha_2} \delta_n^{\alpha_2 \gamma_2} \geq e(c, c^*) - \beta_n^2/2. \end{aligned}$$

Then with the assumption that $\lambda \max(J(c_0), 1) \leq \beta_n^2/4$,

$$\inf_{A_{ij}} E(\tilde{l}(c, D) - \tilde{l}_{\delta_n}(c_0, D)) \geq 2^{i-2}\beta_n^2 + (2^{j-1} - 1)\lambda J(c_0) = M(i, j), \tag{.6}$$

$$\inf_{A_{i0}} E(\tilde{l}(c, D) - \tilde{l}_{\delta_n}(c_0, D)) \geq (2^{i-1} - 3/4)\beta_n^2 \geq 2^{i-3}\beta_n^2 = M(i, 0). \tag{.7}$$

For the second moment, it follows from Assumptions B and C that for any $c \in \mathcal{F}$,

$$\begin{aligned} e(c, c^*) &= E|p_Z(X) - 1/2|\,|sign(X - c^*(Z)) - sign(X - c(Z))| \\ &\geq \xi E\,|sign(X - c^*(Z)) - sign(X - c(Z))|\, I(|p_Z(X) - 1/2| \geq \xi) \\ &\geq \xi\,(E|sign(X - c^*(Z)) - sign(X - c(Z))| - 2a_3\xi^{\alpha_2}) \\ &\geq 2^{-1-2/\alpha_2} a_3^{-1/\alpha_2}\,(E|sign(X - c^*(Z)) - sign(X - c(Z))|)^{(1+\alpha_2)/\alpha_2} \\ &= 2^{-1-2/\alpha_2} a_3^{-1/\alpha_2}\,(E|L_{01}(c^*, D) - L_{01}(c, D)|)^{(1+\alpha_2)/\alpha_2}\,, \end{aligned}$$

with a choice of $\xi = (E|sign(X - c^*(Z)) - sign(X - c(Z))|/4a_6)^{1/\alpha_2}$. Now we are ready to establish an upper bound for the second moment. Note that for any d, $L_{01}(c, D) \leq L_{\delta_n}(c, D)$,

then $E(|L_{01}(c_0, D) - L_{\delta_n}(c_0, D)|) = E(L_{\delta_n}(c_0, D) - L_{01}(c_0, D)) = E(L_{\delta_n}(c_0, D) - L_{\delta_n}(c^*, D) + L_{\delta_n}(c^*, D) - L_{01}(c_0, D)) \leq e_{\delta_n}(c_0, c^*) + a_3 a_4^{\alpha_2} \delta_n^{\alpha_2 \gamma_2}$. Therefore, by the triangular inequality,

$$
\begin{aligned}
& E\left(l(c, D) - l_{\delta_n}(c_0, D)\right)^2 \leq E(|l(c, D) - l_{\delta_n}(c_0, D)|) \\
\leq\ & E|l(c^*, D) - l(c, D)| + E|l(c^*, D) - l(c_0, D)| + E|l(c_0, D) - l_{\delta_n}(c_0, D)| \\
\leq\ & E|l(c^*, D) - l(c, D)| + E|l(c^*, D) - l(c_0, D)| + e_{\delta_n}(c_0, c^*) + a_3 a_4^{\alpha_2} \delta_n^{\alpha_2 \gamma_2} \\
\leq\ & 2^{1+2/\alpha_2} a_3^{1/\alpha_2} (e(c, c^*)^{\alpha_2/(1+\alpha_2)} + e(c_0, c^*)^{\alpha_2/(1+\alpha_2)}) + e_{\delta_n}(c_0, c^*) + a_3 a_4^{\alpha_2} \delta_n^{\alpha_2 \gamma_2} \\
\leq\ & a_6 (e(c, c^*))^{\alpha_2/(1+\alpha_2)},
\end{aligned}
$$

where $a_6 = 2^{2+2/\alpha_2} a_3^{1/\alpha_2} + 1$, and the last inequality is due to the fact that $e(c, c^*) \geq \beta_n^2 \geq e_{\delta_n}(c_0, c^*) + a_3 a_4^{\alpha_2} \delta_n^{\alpha_2 \gamma_2} \geq e(c_0, c^*)$ for any $c \in A_{ij}$. Consequently,

$$
\sup_{A_{ij}} E\left(l(c, D) - l_{\delta_n}(c_0, D)\right)^2 \leq v^2(i, j) \triangleq 8 a_6 M(i, j)^{\alpha_2/(1+\alpha_2)},
$$

where $i = 1, 2, \cdots$ and $j = 0, 1, 2, \cdots$.

Now we are ready to establish the upper bound of I. Using (Equation .6) and (Equation .7), we have

$$
\begin{aligned}
I \ \leq\ & \sum_{i,j} P^* \left(\sup_{A_{ij}} E_n(l_{\delta_n}(c_0, D) - l(c, D)) \geq M(i, j) \right) \\
& + \sum_{i} P^* \left(\sup_{A_{i0}} E_n(l_{\delta_n}(c_0, D) - l(c, D)) \geq M(i, 0) \right) \triangleq I_1 + I_2.
\end{aligned}
$$

Then we bound I_1 and I_2 separately by using Theorem 3 of Shen and Wong (1994), and we just need to verify the conditions (4.5)-(4.7) therein.To compute the metric entropy in (4.7), applying the same technique as in Shen et al. (2003) yields that $H_B(\epsilon, \mathcal{F}(2^j)) \leq H(\epsilon^2/2, \mathcal{G}(2^j))$ for any $\epsilon > 0$ and $j = 0, 1, \cdots$, where $\mathcal{F}(2^j) = \{l(c, d) - l_{\delta_n}(c, d) : c \in \mathcal{F}, J(c) \leq 2^j\}$. Since $\int_{a_7 M(i,j)}^{v(i,j)} H^{1/2}(u^2/2, \mathcal{G}(2^j))du/M(i, j)$ is non-increasing in i and $M(i, j)$, we have

$$\begin{aligned} & \int_{a_7 M(i,j)}^{v(i,j)} H^{1/2}(u^2/2, \mathcal{G}(2^j))du/M(i, j) \\ \leq & \int_{a_7 M(1,j)}^{(8a_6)^{1/2} M(1,j)^{\alpha_2/2(\alpha_2+1)}} H^{1/2}(u^2/2, \mathcal{G}(2^j))du/M(1, j) \leq \phi(\epsilon_n, 2^j), \end{aligned}$$

where $a_7 = 1/64$. Simply let $\epsilon = 1/2$, then Assumption D implies (4.7). Furthermore, (4.5) and (4.6) are satisfied with the above choice of $\epsilon, M(i, j), v(i, j)$ and $T = 1$. In more details, (4.7) implies (4.5) and $M(i, j)/v^2(i, j) \leq 1/8$ implies (4.6).

Then Theorem 3 of Shen and Wong (1994) with $M = n^{1/2} M(i, j), v = v^2(i, j), \epsilon = 1/2$ and $T = 1$ implies that

$$\begin{aligned} I_1 &\leq \sum_{j=1}^{+\infty} \sum_{i=1}^{+\infty} 3 \exp\left(-\frac{nM(i, j)^2}{4(4v^2(i, j) + M(i, j)/3)}\right) \\ &\leq \sum_{j=1}^{+\infty} \sum_{i=1}^{+\infty} 3 \exp\left(-a_8 n M(i, j)^{\frac{\alpha_2+2}{\alpha_2+1}}\right) \\ &\leq \sum_{j=1}^{+\infty} \sum_{i=1}^{+\infty} 3 \exp\left(-a_8 n [2^{i-2}\beta_n^2 + (2^{j-1} - 1)\lambda J(c_0)]^{\frac{\alpha_2+2}{\alpha_2+1}}\right) \\ &\leq \sum_{j=1}^{+\infty} \sum_{i=1}^{+\infty} 3 \exp\left(-a_8 n [(2^{i-2}\beta_n^2)^{\frac{\alpha_2+2}{\alpha_2+1}} + (2^{j-1} - 1)\lambda J(c_0)^{\frac{\alpha_2+2}{\alpha_2+1}}]\right) \end{aligned}$$

$$\leq \frac{3\exp\left(-a_8 n(\lambda J(c_0))^{\frac{\alpha_2+2}{\alpha_2+1}}\right)}{\left(1-\exp(-a_8 n(\lambda J(c_0))^{\frac{\alpha_2+2}{\alpha_2+1}})\right)^2},$$

where a_8 is a positive constant. I_2 can be bounded similarly, and thus

$$I \leq \frac{6\exp\left(-a_8 n(\lambda J(c_0))^{\frac{\alpha_2+2}{\alpha_2+1}}\right)}{\left(1-\exp(-a_8 n(\lambda J(c_0))^{\frac{\alpha_2+2}{\alpha_2+1}})\right)^2},$$

which implies that $I^{1/2} \leq (2.5 + I^{1/2})\exp\left(-a_8 n(\lambda J(c_0))^{\frac{\alpha_2+2}{\alpha_2+1}}\right)$. With $I \leq I^{1/2} \leq 1$, the desired result follows immediately. □

CITED LITERATURE

1. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Center for Devices and Radiological Health: Guidance for industry. Patient-report outcome measures: use in medical product development to support labeling claims, 2009. U.S. Department of Health and Human Services, Rockville, MD.

2. Akaike, H.: A new look at the statistical model idenification. IEEE Transactions on Automatic Control, 19:716–723, 1974.

3. Alho, J. and Valtonen, E.: Interval estimation of inverse dose-response. Biometrics, 51:491–501, 1995.

4. An, L. and Tao, P.: Solving a class of linearly constrained indefinite quadratic problems by d.c. algorithms. Journal of Global Optimization, 11:253–285, 1997.

5. Bartlett, P., Jordan, M., and McAuliffe, J.: Large margin classifiers: convex loss, low noise, and convergence rates. In proceeding of: Advances in Neural Information Processing Systems 16, NIPS, 2003.

6. Bennett, B.: On tests for equality of predictive values for t diagnostic procedures. Statistics in Medicine, 4:535–539, 1985.

7. Copay, A., Subach, B., Glassman, S., Polly, J. D., and Schuler, T.: Understanding the minimum clinically important difference: a review of concepts and methods. The Spine Journal, 7:541–546, 2007.

8. Cortes, C. and Vapnik, V.: Support-vector nteworks. Machine Learning, 20:273–297, 1995.

9. Craven, P. and Wahba, G.: Smoothing noisy data with spline functions. Numerical Mathematics, 31:377–403, 1979.

10. Fan, J. and Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of American Statistical Association, 96:1348–1360, 2001.

11. Fang, X.: A new statistical method for estimating clinically meaningful threshold. Joint Statistical Meetings Proceedings, 2011.

12. Frost, M., Reeve, B., Liepa, A., Stauffer, J., Hays, R., and Group, M. P.-R. O. C. M.: What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health, 10(2S):S94–S105, 2007.

13. Geisser, S.: A predictive sample reuse method with applications. Journal of the American Statistical Association, 70:320–328, 1975.

14. Huang, H., Liu, Y., Du, Y., Perou, C., Hayes, N., Todd, M., and Marron, J.: Multiclass distance weighted discrimination. Journal of Computational and Graphical Statistics, to appear.

15. Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning. 2nd Edition, Springer, 2009.

16. Jacobson, N., Follette, W., and Revenstorf, D.: Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. Behavior Therapy, 15:336–352, 1984.

17. Jacobson, N. and Truax, P.: Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology, 59:12–19, 1991.

18. Jaeschke, R., Singer, J., and Guyatt, H.: Measurement of health status: ascertaining the minimal clinically important difference. Controlled Clinical Trails, 10:407–415, 1989.

19. Juniper, F., Guyatt, H., Willan, A., and Griffith, L.: Determining a minimal important change in a disease-specific quality of life questionnaire. Journal of Clinical Epidemiology, 47:81–87, 1994.

20. Kelly, G.: The median lethal dose-design and estimation. The Statistician, 50:41–50, 2001.

21. Kimeldorf, G. and Wahba, G.: Some results on tchebycheffian spline functions. Journal of Mathematical Analysis and Applications, 33:82–95, 1971.

22. Lee, Y., Lin, Y., and Wahba, G.: Multicategory support vector machines, theory, and application to the classification of micoarray data and satellite radiance data. Journal of the American Statistical Association, 99:67–81, 2004.

23. Leisenring, W., Alonzo, T., and Pepe, M.: Comparisons of predictive values of binary medical diagnostic tests for paired designs. Biometrics, 56:345–351, 2000.

24. Leisenring, W., Alonzo, T., and Pepe, M.: Support vector machines and the bayes rule in classification. Data Mining and Knowledge Discovery, 6:259–275, 2002.

25. Liu, S., Shen, X., and Wong, W.: Computational development of $\psi-$ learning. Proceedings of the SIAM International Conference on Data Mining, Newport, CA, pages 1–12, 2005.

26. Liu, S., Shen, X., and Wong, W.: Multicategory ψ-learning. Journal of the American Statistical Association, pages 500–509, 2006.

27. Liu, Y. and Yuan, M.: Reinforced multicategory support vector machines. Journal of Computational and Graphical Statistics, page 901919, 2011.

28. Marron, J., Todd, M., and Ahn, J.: Distance-weighted discrimination. Journal of the American Statistical Association, page 12671271, 2007.

29. Meinshausen, N. and Bühlmann, P.: Stability selection. Journal of the Royal Statistical Society, Series B, page 414473, 2010.

30. Qiao, X., Zhang, H., Liu, Y., Todd, M., and Marron, J.: Asymptotic properties of distance-weighted discrimination. Journal of the American Statistical Association, page 401414, 2007.

31. Polonik, W.: Measuring mass concentrations and estimating density contour clusters - an excess mass approach. The Annals of Statistics, pages 855–881, 1995.

32. Rigollet, P. and Tong, X.: Neyman-pearson classification, convexity and stochastic constraints. Journal of Machine Learning Research, 12:2831–2855, 2011.

33. Scott, C. and Nowak, R.: A neyman-pearson approach to statistical learning. IEEE Transactions on Information Theory, 51:3806–3819, 2005.

34. Scott, C.: Performance measures for neyman-pearson classification. IEEE Transactions on Information Theory, 53:2852–2863, 2007.

35. Shen, X. and Wong, W.: Convergence rate of sieve estimates. The Annals of Statistics, 22:580–615, 1994.

36. Shen, X., Tseng, G., Zhang, X., and Wong, W.: On ψ-learning. Journal of the American Statistical Association, 98:724–734, 2003.

37. Shiu, S. and Gatsonis, C.: The predictive receiver operating characteristic curve for the joint assessment of the positive and negative predictive values. Philosophical Transactions of The Royal Society A, 366:2313–2333, 2008.

38. Schisterman, E., Perkins, N., Liu, A., and Bondell, H.: Optimal cut-point and its corresponding youden index to discriminate individuals using pooled blood samples. Epidemiology, 16:73–81, 2005.

39. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics, 6:461–464, 1978.

40. Stone, M.: Cross-validatory choice and the assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society, B, 36:111–133, 1975.

41. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58:267–288, 1996.

42. Tsybakov, A.: Optimal aggregation of classifiers in statistical learning. The Annals of Statistics, 32:135–166, 2004.

43. VanderVaart, A.: Asymptotic statistics. New York, Cambridge University Press, 1998.

44. Vapnik, V.: The nature of statistical learning theory. New York, Springer, 1996.

45. Wahba, G.: Spline models for observational data. Philadelphia, SIAM, 1990.

46. Wahba, G.: Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. Cambridge, MA, MIT Press, 1999.

47. Wahba, G.: Soft and hard classification by reproducing kernel hilbert space methods. In Proceedings of the National Academy of Sciences, pages 16524–16530, 2002.

48. Wang, J., Shen, X., and Liu, Y.: Probability estimation for large-margin classifiers. Biometrika, pages 149–167, 2008.

49. Wang, J.: Consistent selection of the number of clusters via cross validation. Biometrika, pages 893–904, 2010.

50. Weston, J.: Extensions to the support vector method. Ph.D. thesis, Royal Holloway University of London, 1999.

51. Williams, D.: Interval estimation of the median lethal dose. Biometrics, 42:641–645, 1986.

52. Wise, E.: Methods for analyzing psychotherapy outcomes: a review of clinical significance, reliable change, and recommendations for future directions. Journal of Personality Assessment, 82:50–59, 2004.

53. Wu, Y., Zhang, H., and Liu, Y.: Robust model-free multiclass probability estimation. Journal of the American Statistical Association, 105:424436, 2010.

54. Wyrwich, W., Nienaber, A., Tierney, M., and Wolinsky, F.: Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-realted quality of life. Medical Care, 37:469–478, 1999.

55. Youden, W.: An index for rating diagnostic tests. Cancer, 3:32–25, 1950.

56. Younger, J., McCue, R., and Mackey, S.: Pain outcomes: a brief review of instruments and techniques. Current Pain and Headache Reports, 13:39–43, 2009.

57. Zhao, P. and Yu, B.: On model selection consistency of lasso. Journal of Machine Learning Research, 7:2541–2563, 2006.

58. Zhou, D.: The covering number in learning theory. Journal of Complexity, 18:739–767, 2002.

59. Zhu, J. and Hastie, T.: Kernel logistic regression and the import vector machine. Journal of Computational and Graphical Statistics, 14:185–205, 2005.

60. Zou, H. and Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67:301–320, 2005.

VITA

Education

- Ph.D. in Statistics, University of Illinois at Chicago, 2013
- M.S. in Statistics, University of Illinois at Chicago, 2010
- M.S. in Mathematics, Ohio University, 2009
- M.S. in Operations Research, East China Normal University, 2008
- B.S. in Mathematics, East China Normal University, 2005

Publication

- An efficient model-free estimation of multiclass conditional probability, *Journal of Statistical Planning and Inference*, **143**, 2079-2088, 2013, with *Wang, J.*.
- Comparing logistic regression, support vector machines and permanental classification methods in predicting hypertension, to appear in *BMC Proceedings*, with *Huang, H. and Yang, J.*.
- On minimum clinically important difference, submitted for publication, with *Hedayat, A. S., Wang, J., and Fang, X.*.

Working Experience

- *ORISE Fellow*, U.S. Food and Drug Administration, Silver Spring, MD, 2013.
- *Co-op*, Baxter Healthcare Corporation, Round Lake, IL, 2011-2012.

Membership

- American Statistical Association (ASA)
- Institute of Mathematical Statistics (IMS)
- Society for Industrial and Applied Mathematics (SIAM)