HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2015 July 31.

Published in final edited form as:

Biometrics. 2014 September ; 70(3): 489–499. doi:10.1111/biom.12179.

Evaluating Marker-Guided Treatment Selection Strategies

Roland A. Matsouaka[#1,*], **Junlong Li**[#2,**], and **Tianxi Cai**[2,***]

[1]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA

[2]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA

[#] These authors contributed equally to this work.

Summary

A potential venue to improve healthcare efficiency is to effectively tailor individualized treatment strategies by incorporating patient level predictor information such as environmental exposure, biological, and genetic marker measurements. Many useful statistical methods for deriving individualized treatment rules (ITR) have become available in recent years. Prior to adopting any ITR in clinical practice, it is crucial to evaluate its value in improving patient outcomes. Existing methods for quantifying such values mainly consider either a single marker or semi-parametric methods that are subject to bias under model misspecification. In this paper, we consider a general setting with multiple markers and propose a two-step robust method to derive ITRs and evaluate their values. We also propose procedures for comparing different ITRs, which can be used to quantify the incremental value of new markers in improving treatment selection. While working models are used in step I to approximate optimal ITRs, we add a layer of calibration to guard against model misspecification and further assess the value of the ITR non-parametrically, which ensures the validity of the inference. To account for the sampling variability of the estimated rules and their corresponding values, we propose a resampling procedure to provide valid confidence intervals for the value functions as well as for the incremental value of new markers for treatment selection. Our proposals are examined through extensive simulation studies and illustrated with the data from a clinical trial that studies the effects of two drug combinations on HIV-1 infected patients.

Keywords

Biomarker-analysis Design; Counterfactual Outcome; Personalized Medicine; Perturbation-resampling; Predictive Biomarkers; Subgroup Analysis

[*]rmatsoua@hsph.harvard.edu. [**]juli@hsph.harvard.edu. [***]tcai@hsph.harvard.edu.

SUPPLEMENTARY MATERIAL Web Appendices referenced in Sections 1 and 3 as well as the R code used to implement our simulations are available with this paper at the Biometrics website on Wiley Online Library.

1. Introduction

The standard analyses of randomized clinical trial evaluate the treatment effect based on the overall treatment difference on the entire study population. However, such overall the treatment effect assessment may not be adequate when a new treatment benefits patients differentially depending on each patient's characteristics. A treatment deemed effective on average does not guarantee that it will be beneficial to all future patients. Conversely, a negative finding on the average the treatment effect does not imply that the new treatment is entirely futile since the effectiveness of a treatment on a small subgroup of patients may be hidden by its inactivity on a large (heterogeneous) patient population (Rothwell, 1995; Rothwell et al., 2005; Kent and Hayward, 2007). When the treatment effect varies across subpopulations, it would be desirable to develop *individualized treatment rules* (ITR) according to individual patients' baseline characteristics. Assigning treatments to achieve optimal patient outcomes may substantially improve healthcare efficiency (Baker et al., 2012).

Statistical methods for developing optimal ITRs have received much attention in recent years. Traditional methods based on ad hoc subgroup analyses or searching for markertreatment interactions, while useful, may not be efficient or valid due to the curse of dimensionality and multiple comparisons. More systematic approaches to deriving ITR have been recently proposed. With a single baseline marker, semi- and non-parametric procedures have been proposed to identify a subgroup of patients who would benefit from the new treatment (e.g., Song and Pepe, 2004; Bonetti and Gelber, 2000, 2004). With multiple baseline markers, a wide range of procedures have been proposed to derive ITRs that combines information across all markers (e.g., Qian and Murphy, 2011; Imai and Strauss, 2011; Foster et al., 2011; Cai et al., 2011; Zhao et al., 2012; Zhang et al., 2012a; Zhao et al., 2013).

As strategies for deriving ITRs become increasingly available, it is important to examine the net benefit of assigning treatment according to an ITR prior to recommending its wide spread use. Most current research focuses on developing ITRs with relative little attention given to making robust inference about such estimated ITRs and their value in improving population outcomes. Although a few methods have been proposed to quantify such values, these methods consider either a single marker or semi-parametric methods that are subject to bias under model misspecification (Song and Pepe, 2004; Song and Zhou, 2009; Janes et al., 2011; Huang et al., 2012, e.g). Zhang et al. (2012a) propose a robust approach to overcome model misspecification by restricting the ITR in a parametric class and estimate the ITR parameters by maximizing an empirical value function associated with the ITR. The direct maximization of the non-smooth empirical value function could suffer substantial variability in the estimated ITR parameters. As we show Section 3.2 and Web Appendix B, even for a univariate X with ITR given by $I(X \geq c)$, direct maximization gives an estimate of c with a cubic convergence rate. When there are multiple markers, direct maximization of an empirical value function with respect to all unknown parameters involved in the ITR, such as those proposed in Zhang et al. (2012b), could be computationally prohibitive and unstable. Here, we consider a general setting with multiple markers and adopt a two-step method to derive a class of ITRs and make inference about the value of such ITRs. We also

propose procedures for comparing different ITRs, which can be used to quantify the *incremental value* (IncV) of new markers in improving treatment selection. Such IncV assessment is particularly important if a marker used in the ITR is expensive and/or invasive.

The remainder of this paper is organized as follows. We describe in Section 2, the general framework for quantifying the value of ITRs and deriving ITRs that attain maximal values. We also provide some simple results demonstrating that a two-step procedure could potentially lead to an ITR that is optimal (i) among all ITRs based on a set of predictors X when the fitted models in the first step are nearly correct; and (ii) within a smaller class of ITRs when the models are misspecified. In Section 3, we provide the estimation and procedures for making inference about the proposed two-step ITR as well as its value function. In Section 4, we evaluate the finite sample performance of our proposed methods through a series of simulation studies. We apply the proposed method, in Section 5, to a data set from a clinical trial (ACTG 320) conducted by the AIDS Clinical Trial Group as a further illustration. In Section 6, we provide some concluding remarks and further discussion.

2. Quantifying the Value of ITR and Optimizing ITR

2.1 Notations and Settings

Let Y be the response variable and $Y^{(j)}$ denote the potential outcome (Rubin, 1974; Rubin et al., 1978; Holland, 1986; Robins, 1986) of a patient if assigned to treatment $G = j$, where $j = 1$ refers to the experimental treatment and $j = 0$ to the standard treatment. The potential outcome (also referred to as counterfactual outcome) $Y^{(j)}$ is defined as the value of the outcome had the treatment been set to $G = j$ by external intervention. Both Y and $Y^{(j)}$ are related via the *consistency assumption* $Y = GY^{(1)} + (1 – G)Y^{(0)}$. We assume the standard *stable unit treatment value assumption* i.e. each subject's potential response to a treatment does not depend on the treatment assignment mechanism, the treatments received by other patients or their potential responses to treatments (Rubin, 1980, 1986).Without loss of generality, we assume that a larger value of Y is more beneficial.

Let $\mathscr{I}(\mathbf{X})$ denote a binary ITR as a function of a baseline covariate vector $\mathbf{X}$ with $\mathscr{I}(\mathbf{X}) = j$ indicating assignment to treatment j. Our goal is to identify an optimal $\mathscr{I}(\cdot)$ that maximizes patients' outcomes. When the treatment selection is optimized for all patients, the resulting population average outcome is also optimized. Thus, an optimal ITR is also expected to maximize a population average *value function*. A sensible choice of the value function is the *cost-adjusted population average outcome* associated with $\mathscr{I}$:

$$\mathbb{V}_{\mathscr{I}} = E\left\{\mathscr{I}(\mathbf{X})\left(Y^{(1)} - \xi\right) + (1 - \mathscr{I}(\mathbf{X}))\, Y^{(0)}\right\},$$

where ξ is a pre-specified incremental financial and/or medical cost associated with taking the new treatment as compared to the standard treatment. It is not difficult to see that the optimal $\mathscr{I}(\cdot)$ maximizing $\mathbb{V}_{\mathscr{I}}$ is the Bayes rule

$$\mathscr{I}_{Bayes}(\mathbf{X}) = \underset{\mathscr{I}}{argmax}\, \mathbb{V}_{\mathscr{I}} = I\{D(\mathbf{X}) \geq \xi\}, \quad (2.1)$$

where $I(.)$ is the indicator function and $D(\mathbf{X}) = E(Y^{(1)}|\mathbf{X} - E(Y^{(0)}|\mathbf{X})$. With a given dataset, the optimal ITR $\mathscr{I}_{Bayes}(\mathbf{X})$ can be approximated by estimating the conditional treatment effect function $D(\mathbf{X})$ or by direct optimization of $\mathbb{V}_{\mathscr{I}}$ within a class of $\mathscr{I}$.

Throughout, we use notation $D(\beta,\mathbf{X})$ to denote a model based approximation to the true conditional mean difference $D(\mathbf{X})$ with a model parameter value β; let $\hat{\beta}$ denote its estimate from the data and $\bar{\beta}$ denote its limiting value. In addition, we let $\bar{\Delta}(s) = E\left\{Y^{(1)} - Y^{(0)} | D\left(\bar{\beta}, \mathbf{X}\right) = s\right\}$ denote a calibrated estimate of treatment difference given $D\left(\bar{\beta}, \mathbf{X}\right) = s$. We next describe the pros and cons of various approaches.

2.2 Various Approaches to Approximating $\mathscr{I}_{Bayes}$

When $\mathbf{X} = X$ is univariate, we can approximate $\mathscr{I}_{Bayes}(\cdot)$ by estimating $D(X)$ via kernel smoothing. Note that if $D(X)$ is an increasing function in X, the optimal rule $\mathscr{I}_{Bayes}(X)$ must take the form $I(X \geq c°)$, where $c° = \text{argmax}_c\, \mathbb{V}(c)$ and $\mathbb{V}(c) = E\left\{I(X \geq c)\left(Y^{(1)} - \xi\right) + I(X < c)\, Y^{(0)}\right\}$. Evaluating the utility of a single marker based on $\mathbb{V}(c)$ has been previously considered in Song and Pepe (2004) and Song and Zhou (2009). However, when $D(X)$ is not monotone in X, the optimal ITR may not take the form of $I(X \geq c)$ and $\mathbb{V}(c) < \mathbb{V}_{\mathscr{I}_{Bayes}}$ for any c if there exists $x_1 > x_2$ such that $D(x_1) = D(x_2) = \xi$ and $P(x_1 > X > x_2) > 0$.

With multivariate $\mathbf{X}$, using fully non-parametric methods and incorporating non-linear functional spaces for approximating $\mathscr{I}_{Bayes}(\cdot)$ (Foster et al., 2011; Zhao et al., 2012) could be extremely valuable, especially when $D(\mathbf{X})$ takes a complex form. However, these methods are subject to curse of dimensionality and pose challenges in making inference about the resulting ITR and its associated value function. On the other hand, if $D(\mathbf{X})$ is estimated by imposing parametric or semi-parametric models on $E(Y^{(j)}|\mathbf{X})$, the plug-in estimate of $\mathscr{I}_{Bayes}(\mathbf{X})$ may lead to a much lower population average outcome compared to that of the true $\mathscr{I}_{Bayes}(\mathbf{X})$ (Qian and Murphy, 2011). One may reduce model misspecification by including non-linear bases and selecting important variables via regularized estimation (Qian and Murphy, 2011; Imai and Strauss, 2011). However, it remains challenging to efficiently choose non-linear basis functions to achieve an optimal bias and variance trade-off.

We seek to overcome model misspecification by following a two-step principal previously proposed in Cai et al. (2011) and Zhao et al. (2013): (I) obtain a parametric or semiparametric model based estimate of $D(\mathbf{X})$, denoted by $D\left(\hat{\beta};\mathbf{X}\right)$, where $\hat{\beta}$ is the estimated model parameters; and (II) non-parametrically estimate treatment effect

parameters to account for possible model misspecification in step I. Both of these existing methods, while related to using working models for treatment selection, do not address the question of how to derive an optimal ITR or how to make inference about its associated value, which is the focus of this paper. Cai et al. (2011) uses $D\left(\hat{\beta};\mathbf{X}\right)$ as a univariate score to create subgroups, $\Omega_s=\left\{\mathbf{X}:D\left(\hat{\beta};\mathbf{X}\right)=s\right\}$ and provides inference procedures for $E\{Y^{(1)}-Y^{(0)}\mid \mathbf{X}\in\Omega_s\}$. Zhao et al. (2013) proposes a non-parametric estimator of $E\left\{Y^{(1)}-Y^{(0)}|D\left(\hat{\beta},\mathbf{X}\right)\geq c\right\}$ for a range of c in step II and hence only relate to ITRs of the form $D\left(\hat{\beta},\mathbf{X}\right)\geq c$. Although the methods given in Zhao et al (2013) can be used to compare the potential of two scores in guiding treatment selection, their measures do not have a clear clinical interpretation and cannot be used to quantify the performance of a single score.

In this paper, we propose a two-step approach to construct a calibrated ITR,

$$\mathscr{I}_{calib}\left(\mathbf{X}\right)=I\left\{\bar{\Delta}\left(D\left(\bar{\beta},\mathbf{X}\right)\right)\geq\xi\right\},$$

and evaluate its value $\mathbb{V}_{\mathscr{I}_{calib}}$. It follows from (2.1) that $\mathscr{I}_{calib}\left(\mathbf{X}\right)$ is the optimal ITR based on the univariate score $D\left(\bar{\beta},\mathbf{X}\right)$. We next detail some pros and cons of using $\mathscr{I}_{calib}$ for treatment selection under various conditions.

- When the working models in the first step are *nearly correct* such that $D(\mathbf{X})$ is an increasing function of $D\left(\bar{\beta},\mathbf{X}\right)$, then $\bar{\Delta}\left(D\left(\bar{\beta},\mathbf{X}\right)\right)=D\left(\mathbf{X}\right)$. Hence, the two-step procedure leads to the optimal ITR with $\mathbb{V}_{\mathscr{I}_{calib}}=\mathbb{V}_{\mathscr{I}_{Bayes}}$.
- Under more severe model misspecification when $P\left[I\left\{D\left(\mathbf{X}\right)>\xi\right\}\neq I\left\{\bar{\Delta}\left(D\left(\bar{\beta},\mathbf{X}\right)\right)>\xi\right\}\right]>0$, $\mathscr{I}_{calib}$ will be sub-optimal relative to $\mathscr{I}_{Bayes}$ with $\mathbb{V}_{\mathscr{I}_{calib}}<\mathbb{V}_{\mathscr{I}_{Bayes}}$. However, when $\mathscr{I}_{Bayes}$ and $\mathscr{I}_{calib}$ are replaced with their respective estimates $\hat{\mathscr{I}}_{Bayes}$ and $\hat{\mathscr{I}}_{calib}$ obtained in finite sample, applying $\hat{\mathscr{I}}_{Bayes}$ to a future population may not yield a value function higher than that of $\hat{\mathscr{I}}_{calib}$ if the sampling variability associated with $\hat{\mathscr{I}}_{Bayes}$ is much larger than that of $\hat{\mathscr{I}}_{calib}$.
- When the model misspecification is not severe with $\bar{\Delta}\left(s\right)$ being increasing in s, $\mathscr{I}_{calib}\left(\mathbf{X}\right)$ takes the form $I_{c^\circ}=I\left(D\left(\bar{\beta},\mathbf{X}\right)>c^\circ\right)$, where $c^\circ=\text{argmax}_c\,\mathbb{V}_{I_c}$ with $\mathbb{V}_{I_c}=E\left\{I\left(D\left(\bar{\beta},\mathbf{X}\right)\geq c\right)\left(Y^{(1)}-\xi\right)+I\left(D\left(\bar{\beta},\mathbf{X}\right)<c\right)Y^{(0)}\right\}$.

- When $\bar{\Delta}(s)$ is not monotone and there exists $d_1 > d_2$ such that $\bar{\Delta}(d_1) = \bar{\Delta}(d_2) = \xi$ and $P\left(d_1 > D\left(\bar{\beta}, \mathbf{X}\right) > d_2\right) > 0$, we have $\mathbb{V}_{I_{c^\circ}} < \mathbb{V}_{\mathscr{I}_{calib}} < \mathbb{V}_{\mathscr{I}_{Bayes}}$. In other words, under certain model misspecifications such as missing a quadratic effect, assigning treatment according to $I\left(D\left(\bar{\beta}, \mathbf{X}\right) \geq c\right)$ as in Zhao et al. (2013) will be sub-optimal when compared to $\mathscr{I}_{calib}$.

These findings suggest the benefit in approximating $\mathscr{I}_{Bayes}$ with $\mathscr{I}_{calib}$ and motivate us to provide a robust estimate of the ITR as follows:

1. posit working models to approximate $D(\mathbf{X})$ using a model based score $D\left(\bar{\beta}, \mathbf{X}\right)$;
2. non-parametrically estimate $\bar{\Delta}(s)$ as $\hat{\Delta}(s)$ using observed responses along with $D\left(\bar{\beta}, \mathbf{X}\right)$;
3. for a future patient with $\mathbf{X} = \mathbf{x}$, the treatment assignment is determined using $\hat{\mathscr{I}}_{calib}(\mathbf{X}) = I\left[\hat{\Delta}\left\{D\left(\hat{\beta}, \mathbf{X}\right)\right\} \geq \xi\right]$, which is an estimator of $\mathscr{I}_{calib}$ with $P\left\{\hat{\mathscr{I}}_{calib}(\mathbf{X}) \neq \mathscr{I}_{calib}(\mathbf{X})\right\} \to 0$;
4. evaluate the performance of $\hat{\mathscr{I}}_{calib}(\cdot)$ based on the observed data by estimating $\mathbb{V}_{\mathscr{I}_{calib}}$.

We next detail our proposed estimators for $\mathscr{I}_{calib}$ and its associated value $\mathbb{V}_{\mathscr{I}_{calib}}$ along with their asymptotic properties.

3. Estimation and Inference Procedures for $\mathscr{I}_{calib}$ and $\mathbb{V}_{\mathscr{I}_{calib}}$

We assume that data available for analysis are from a randomized clinical trial with study participants randomly assigned to either a standard treatment ($G = 0$) or an experimental treatment ($G = 1$). Our data consist of n random vectors *****
$\mathscr{D} = \{(Y_i, \mathbf{X}_i, G_i), i = 1, \ldots, n\}$, where we assume that $\{(Y_i, \mathbf{X}_i) : G_i = j\}$ are $n_j = \sum_{i=1}^{n} I(G_i = j)$ independent and identically distributed random vectors, for $j = 0, 1$. Furthermore, we assume that the ratio n_1/n converges to a constant $\pi_1 \in (0, 1)$ as $n \to \infty$.

3.1 Estimation of $\mathscr{I}_{calib}$ and $\mathbb{V}_{\mathscr{I}_{calib}}$

We first approximate $D(\mathbf{X})$ by imposing parametric or semi-parametric working models as $E(Y_i \mid \mathbf{X}_i = \mathbf{X}, G_i = j) = \mu_j(\beta_j; \mathbf{X})$ with β being the unknown model parameter for group j and μ_j being a known link function. Without loss of generality, we suppose that β_j is estimated by $\hat{\beta}_j$, the solution to $\hat{\mathbb{U}}_j(\beta) = 0$, where

$$\hat{\mathbb{U}}_j(\beta) = n_j^{-1} \sum_{i=1}^{n} I(G_i = j)\, \mathrm{U}_j(\beta, Y_i, \mathbf{X}_i) + \lambda_{jn}\beta, \quad (3.1)$$

$\mathbf{U}_j(\beta;\cdot,\cdot)$ is the estimating function relating the observed data to the parameters of interest, and $0 \leq \lambda_{jn}=o\left(n_j^{-\frac{1}{2}}\right)$ is a tuning parameter chosen to ensure stable fitting when the dimension of β is not too small compared to the sample size. For example, one may let $\mathbf{U}_j(\beta;Y_i,\mathbf{X}_i)=\mathbf{X}_i\left\{Y_i-g\left(\beta_j^{\intercal}\psi(\mathbf{X_i})\right\}\right.$ under generalized linear models, where $\psi(\cdot)$ is a prespecified finite dimensional vector of potentially non-linear transformation functions which would allow one to capture non-linear effects. We choose penalty $\lambda_{jn}=o\left(n_j^{-\frac{1}{2}}\right)$ so that $n^{\frac{1}{2}}\left(\hat{\beta}-\bar{\beta}\right)$ still converges to a zero-mean normal random vector to enable easy inference. Then, the model based estimate of $D(\mathbf{X})$ can be obtained as $D\left(\hat{\beta};\mathbf{X}\right)=\mu_1\left(\hat{\beta}_0;\mathbf{X}\right)-\mu_0\left(\hat{\beta};\mathbf{X}\right)$, where $\hat{\beta}=\left(\hat{\beta}_0^{\intercal},\hat{\beta}_1^{\intercal}\right)^{\intercal}$ is a vector of estimated model parameters.

Next, we propose to estimate $\mathscr{I}_{calib}$ and $\mathbb{V}_{\mathscr{I}_{calib}}$ non-parametrically with $\hat{\mathscr{I}}_{calib}(\mathbf{X})=I\left[\hat{\Delta}\left\{D\left(\hat{\beta},\mathbf{X}\right)\right\}\geq\xi\right]$ and $\hat{\mathbb{V}}_{\hat{\mathscr{I}}_{calib}}$ respectively, where

$$\begin{aligned}\hat{\mathbb{V}}_{\mathscr{I}}&=n_1^{-1}\sum_{i=1}^{n}\mathscr{I}(\mathbf{X}_i)(Y_i-\xi)I(G_i=1)+n_0^{-1}\sum_{i=1}^{n}\{1-\mathscr{I}(\mathbf{X}_i)\}Y_iI(G_i=0),\\ \hat{\Delta}(s)&=\frac{\sum_{i=1}^{n}K_h\left(D\left(\hat{\beta};\mathbf{X}_i\right)-s\right)Y_iI(G_i=1)}{\sum_{i=1}^{n}K_h\left(D\left(\hat{\beta};\mathbf{X}_i\right)-s\right)I(G_i=1)}-\frac{\sum_{i=1}^{n}K_h\left(D\left(\hat{\beta};\mathbf{X}_i\right)-s\right)Y_iI(G_i=0)}{\sum_{i=1}^{n}K_h\left(D\left(\hat{\beta};\mathbf{X}_i\right)-s\right)I(G_i=0)}\end{aligned}\tag{3.2}$$

representing a non-parametric smoothed estimator of $\bar{\Delta}(s)$ and $K(\cdot)$ is a smooth symmetric density function with $h\to 0$ as $n\to 1$.

We show in Web Appendix A that, under mild regularity conditions, $P\left\{\hat{\mathscr{I}}_{calib}(\mathbf{X})\neq\mathscr{I}_{calib}(\mathbf{X})\right\}\to 0$ and $\hat{\mathbb{V}}_{\hat{\mathscr{I}}_{calib}}\to\mathbb{V}_{\mathscr{I}_{calib}}$ in probability. Furthermore, we show that $n^{\frac{1}{2}}\left(\hat{\mathbb{V}}_{\hat{\mathscr{I}}_{calib}}-\mathbb{V}_{\mathscr{I}_{calib}}=n^{\frac{1}{2}}\left(\hat{\mathbb{V}}_{\widetilde{\mathscr{I}}_{calib}}-\mathbb{V}_{\mathscr{I}_{calib}}\right)+o_p(1)\right.$, which converges in distribution to a zero mean normal random variable with variance $\sigma^2=E\left(\Psi_{\nu_i}^2\right)$, where Ψ_{ν_i} is defined in (A.1) of Web Appendix A and $\widetilde{\mathscr{I}_{calib}}(\mathbf{X})=I\left[\bar{\Delta}\left\{D\left(\hat{\beta},\mathbf{X}\right)\geq\xi\right]\right.$ is the estimated treatment assignment rule knowing $\bar{\Delta}(\cdot)$.

This indicates that $\hat{\Delta}(s)$ in the non-parametric calibration step does not contribute any additional variability for the estimation of $\mathbb{V}_{\mathscr{I}_{calib}}$ at the first order. Alternative choices of $\hat{\Delta}(s)$ such as a local likelihood estimator can also be valid provided that $\hat{\Delta}(s)-\bar{\Delta}(s)=o_p\left(n^{-1/4}\right)$.

3.2 Estimation of the Threshold Under Monotonicity

Under the assumptions that $\bar{\Delta}(s)$ is monotone, $\bar{\Delta}(s)=\xi$ has a unique solution at c°, and the product $P\left\{D\left(\bar{\beta},\mathbf{X}\right)>\bar{\Delta}^{-1}(\xi)\right\}P\left\{D\left(\bar{\beta},\mathbf{X}\right)<\bar{\Delta}^{-1}(\xi)\right\}>0$. The optimal ITR based on $D\left(\bar{\beta},\mathbf{X}\right)$ takes the form $\mathscr{I}_{calib}(\mathbf{X})=I\left(D\left(\bar{\beta},\mathbf{X}\right)>c^\circ\right)$, where $c^\circ = \text{argmax}_c\,\mathbb{V}_{I_c}$ is an interior point of the support of $D\left(\bar{\beta},\mathbf{X}\right)$. Hence, one may also approximate $\mathscr{I}_{calib}(\mathbf{X})$ via direct maximization as $\hat{I}_{\hat{\phi}}(\mathbf{X}$, for $\hat{\phi}=argmax_c\,\hat{\mathbb{V}}_{\hat{I}_c}$, where we define $\hat{I}_c(\mathbf{X})=I\left\{D\left(\hat{\beta},\mathbf{X}\right)\geq c\right\}$.

We show, in Web Appendix B, that $\hat{\phi}$ is a consistent estimator of c°. However, due to the non-smoothness in the empirical value function $\hat{\mathbb{V}}_{\hat{I}_c}$, we have $\hat{\phi}-c^\circ=O_p\left(n^{-1/3}\right)$, suggesting that direct maximization results in an estimator with a much slower convergence rate and hence large variability in $\hat{\phi}$. Furthermore, following from Theorem §1.1 of Kim and Pollard (1990), $n^{1/3}\left(\hat{\phi}-c_0\right)$ converges in distribution to $\text{argmin}_t\,\mathscr{Z}(t)$, where $\mathscr{Z}(t)$ is a Gaussian process. However, $\text{argmin}_t\,\mathscr{Z}(t)$ does not generally have an explicit form.

Since co is also the solution to $\bar{\Delta}(c)-\xi=0$, we propose to estimate c° as $\hat{c}$, the solution to $\bar{\Delta}(c)=\xi$. As shown in Web Appendix B, $\hat{c}-c^\circ=O_p\left\{(nh)^{-1/2}+h^2\right\}$ and $\sqrt{nh}\,(\hat{c}-c^\circ)$ converges in distribution to a normal when $h = O(n^{-\nu})$ with $\nu \in = [1/5, 1/2)$. In addition, $\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}\rightarrow\mathbb{V}_{\mathscr{I}_{calib}}$ and $n^{\frac{1}{2}}\left(\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}-\mathbb{V}_{\mathscr{I}_{calib}}\right)=n^{\frac{1}{2}}\left(\hat{\mathbb{V}}_{\hat{I}_{c^\circ}}-\mathbb{V}_{\mathscr{I}_{calib}}\right)+O_p(1)=n^{\frac{1}{2}}\left(\hat{\mathbb{V}}_{\mathscr{I}_{calib}}-\mathbb{V}_{\mathscr{I}_{calib}}\right)+O_p(1)$. This suggests that, under the monotonicity assumption, the variability of $\hat{c}$ is ignorable at the first order when making inference about $\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$ and the estimator $\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$ is asymptotically equivalent to $\hat{\mathbb{V}}_{\mathscr{I}_{calib}}$ obtained by first estimating $\bar{\Delta}(s)$ via smoothing.

Finally, using similar arguments as those given inWeb Appendix B, one may also show that $n^{\frac{1}{2}}\left(\hat{\mathbb{V}}_{\hat{I}_{\hat{\phi}}}-\mathbb{V}_{\mathscr{I}_{calib}}\right)=n^{\frac{1}{2}}\left(\hat{\mathbb{V}}_{\hat{I}_{c^\circ}}-\mathbb{V}_{\mathscr{I}_{calib}}\right)+O_p(1)$. However, in finite sample, the large variability in $\hat{\phi}$ leads to $\hat{\mathbb{V}}_{\hat{I}_{\hat{\phi}}}$ having higher variability than that of $\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$. We further illustrate these points in the simulation section.

3.3 Bias Correction and Interval Estimation

In practice, for a small or moderate sample size *n*, the value of the marker-guided ITR could be substantially over-estimated due to overfitting (Zhao et al., 2013). To correct for the bias, we use the standard cross-validation (CV) technique. Specifically, we first randomly splits

the data into $\mathscr{K}$ disjoint subsets of about equal size and label them as $\mathbb{E}_k$, where $k = 1, \ldots \mathscr{K}$. For each k, we use all the observations not in $\mathbb{E}_k$ to obtain $\hat{\beta}_j^{(-k)}$ for β_j via (3.1), and compute the score $D\left(\hat{\beta}^{(-k)};\mathbf{X}_i\right)$ as well as $\hat{\Delta}^{(-k)}(s)$. Then, we use observations in $\mathbb{E}_k$ to obtain

$$\hat{\mathbb{V}}_{\mathscr{I}_{calib},k}=\frac{\sum_{i\in\mathbb{E}_k} I\left\{\hat{\Delta}^{(-k)}\left(D\left(\hat{\beta}^{-k};\mathbf{X}_i\right)\right)\geq\xi, G_i=1\right\}(Y_i-\xi)}{\sum_{i\in\mathbb{E}_k} I(G_i=1)}+\frac{\sum_{i\in\mathbb{E}_k} I\left\{\hat{\Delta}^{-k}\left(D\left(\hat{\beta}^{(-k)};\mathbf{X}_i\right)\right)<\xi, G_i=0\right\}Y_i}{\sum_{i\in\mathbb{E}_k} I(G_i=0)}$$

Finally, we obtain the CV estimator for $\mathbb{V}_{\mathscr{I}_{calib}}$ as $\mathbb{V}^{(cv)}_{\mathscr{I}_{calib}}=\mathscr{K}^{-1}\sum_{k=1}^{\mathscr{K}}\hat{\mathbb{V}}_{\mathscr{I}_{calib},k}$

The variance of $\hat{\mathbb{V}}_{\mathscr{I}_{calib}}$ involves unknown density-like functions, which makes it difficult to estimate directly, especially when the number of covariates in the model is not small. To circumvent this issue, we use a perturbation-resampling technique to obtain a good approximation to the distribution of our proposed estimators. Specifically, let $\{W_i, i = 1; \ldots, n\}$ be n independent and identically distributed random variables generated from a known distribution with mean 1 and variance 1. The perturbed version of $\hat{\mathbb{V}}_{\mathscr{I}_{calib}}$ can be obtained as

$$\hat{\mathbb{V}}^*_{\mathscr{I}_{calib}}=\frac{\sum_{i=1}^{n} I\left\{\hat{\Delta}^{(*)}\left(D\left(\hat{\beta}^*;\mathbf{X}_i\right)\right)\geq\xi, G_i=1\right\}(Y_i-\xi)W_i}{\sum_{i=1}^{n} I(G_i=1)W_i}+\frac{\sum_{i=1}^{n} I\left\{\hat{\Delta}^{*}\left(D\left(\hat{\beta}^*;\mathbf{X}_i\right)\right)<\xi, G_i=0\right\}Y_iW_i}{\sum_{i=1}^{n} I(G_i=0)W_i},$$

where

$$\begin{aligned}\hat{\Delta}^*(s)=&\sum_{i=1}^{n}K_h\left(D\left(\hat{\beta}^*;\mathbf{X}_i\right)\right.\\&-sY_iI(G_i\\&=1)W_i\\&/\left\{\sum_{i=1}^{n}K_h\left(D\left(\hat{\beta}^*;\mathbf{X}_i\right)\right.\right.\\&-sI(G_i=1)W_i-\sum_{i=1}^{n}K_h\left(D\left(\hat{\beta}^*;\mathbf{X}_i\right)-s\right)Y_iI(G_i=0)W_i/\left\{\sum_{i=1}^{n}K_h\left(D\left(\hat{\beta}^*;\mathbf{X}_i\right)\right.\right.\\&\left.\left.-s)I(G_i=0)W_i\right\}\end{aligned}$$

and $\hat{\beta}^*_j$ is the solution to $n_j^{-1}\sum_{i=1}^{n} I(G_i=j)\mathbf{U}_j(\beta;Y_i,\mathbf{X}_i)W_i+\lambda_{jn}\beta=0$.

Using arguments similar to Park and Wei (2003) and Cai et al. (2005), we can show that, the distribution of $n^{\frac{1}{2}}\left(\hat{\mathbb{V}}_{\mathscr{I}_{calib}}-\mathbb{V}_{\mathscr{I}_{calib}}\right)$ can be approximated by the conditional distribution of $n^{1/2}\left(\hat{\mathbb{V}}^*_{\mathscr{I}_{calib}}-\hat{\mathbb{V}}_{\mathscr{I}_{calib}}\right)$ given the data. Therefore, a 100(1 – α)% confidence interval (CI) for $\mathbb{V}_{\mathscr{I}_{calib}}$ is constructed as $\hat{\mathbb{V}}_{\mathscr{I}_{calib}}\pm z_{100(1-\alpha/2)}\hat{\sigma}_{\mathbb{V}}$ or $\hat{\mathbb{V}}^{(cv)}_{\mathscr{I}_{calib}}\pm z_{100(1-\alpha/2)}\hat{\sigma}_{\mathbb{V}}$, where $\hat{\sigma}_{\mathbb{V}}$ is obtained as the standard error (SE) of $\hat{\mathbb{V}}^*_{\mathscr{I}_{calib}}$ and z_q is the qth percentile of the standard normal distribution. When $h = O(n^{-\nu})$ with $\nu \in (1/5, 1/2)$, the resampling method can also be

used to obtain CIs for c° based on $\hat{c}$. However, we acknowledge that it is unclear whether the proposed resampling procedures or the standard bootstrap can be used to approximate the distribution of $\hat{\zeta}$ due to its non-regular distribution. Our numerical results seem to suggest that the perturbation works reasonably well in finite sample.

3.4 Incremental Value (IncV) of Markers in Improving ITR

When a set of new markers $\mathbf{X}_{new}$ is available to further improve ITR, it is important to assess their IncV, especially if these markers are costly or invasive. *Using the population average outcome*, we may quantify the IncV of the new markers by comparing the ITRs constructed with $\mathbf{X}_{old}$ and $\mathbf{X}_{update}=(\mathbf{X}_{old}^\intercal, \mathbf{X}_{new}^\intercal)^\intercal$, denoted respectively by $\mathscr{I}_{calib}^{old}$ and $\mathscr{I}_{calib}^{update}$, with respect to their value functions, i.e. the cost-adjusted population average outcomes. Specifically, the IncV of $\mathbf{X}_{new}$ can be quantified as $\mathrm{V}=\mathbb{V}_{\mathscr{I}_{calib}^{update}} - \mathbb{V}_{\mathscr{I}_{calib}^{old}}$. Based on the inference procedures described above, we obtain a plug-in estimate of IncV as $\hat{IncV}=\hat{\mathbb{V}}_{\mathscr{I}_{calib}^{update}} - \hat{\mathbb{V}}_{\mathscr{I}_{calib}^{old}}$.

The CI of IncV can be constructed via the perturbation-resampling approach as well. For each set of perturbation random variables $\{W_i, i = 1, \ldots, n\}$, we follow the perturbation procedures described in Section 3.3 to obtain perturbed counterpart of $(\hat{\mathbb{V}}_{\mathscr{I}_{calib}^{old}}, \hat{\mathbb{V}}_{\mathscr{I}_{calib}^{update}})$ as $(\hat{\mathbb{V}}^*_{\mathscr{I}_{calib}^{old}}, \hat{\mathbb{V}}^*_{\mathscr{I}_{calib}^{update}})$. Then the perturbed counterpart of $\hat{IncV}$ can be obtained as $\hat{IncV}^*=\hat{\mathbb{V}}^*_{\mathscr{I}_{calib}^{update}} - \hat{\mathbb{V}}^*_{\mathscr{I}_{calib}^{old}}$. The empirical distribution of $n^{\frac{1}{2}}\left(\hat{IncV}^* - \hat{IncV}\right)$ conditional on the data can be used to approximate the distribution of $n^{\frac{1}{2}}\left(\hat{IncV} - IncV\right)$ and construct the CI accordingly.

4. Simulation Studies

To evaluate the performance of the proposed method under practical settings, we conducted extensive simulation studies to examine the finite-sample properties of the proposed point and interval estimation estimators. In addition, we compared the performance of the proposed procedures to that of Qian and Murphy (2011) and Zhang et al. (2012a) with respect to the achievable value function. Throughout the simulation studies, we let $n_0 = n_1 = 250$ and set $\xi = 0$ for simplicity. All the CIs were obtained using 500 perturbed samples with $W_i \sim \exp(1)$. All the results are summarized based on 1000 simulated datasets.

4.1 Evaluation of the Proposed Method

We first evaluated the proposed point and interval estimation procedures for the value functions. We generated covariates from

$$\begin{bmatrix} log\,(X_1) \\ X_2 \\ log\,(X_3) \\ X_4 \end{bmatrix} \sim N\left(\begin{bmatrix} 1.5 \\ 5.0 \\ -3.0 \\ 2.0 \end{bmatrix}, \begin{bmatrix} 4.0 & 0.8 & 0 & 0 \\ 0.8 & 16.0 & 0 & 0 \\ 0 & 0 & 8.0 & 3.2 \\ 0 & 0 & 3.2 & 8.0 \end{bmatrix} \right); \quad X_5 \sim Bernoulli\,(0.5)$$

Given $\check{\mathbf{X}}=(1, log\,(X_1), X_2, log\,(X_3), X_4, X_5)^{\intercal}$, two types of outcomes were generated: (1) a continuous outcome from a linear model

$$Y_i=\beta_1^{\intercal}\check{\mathbf{X}}_i I\,(G_i=1)+\beta_0^{\intercal}\check{\mathbf{X}}_i I\,(G_i=0)+\epsilon_i, \quad (4.1)$$

where $\epsilon_i \sim N(0, 1)$ with $\beta_1 = (1, 0.5, 1, 1.5, 1, 1)^{\top}$ and $\beta_0 = (6, 2.5, -3, -1, 1.5, -2)^{\top}$; (2) a binary outcome from a logistic regression model

$$logit\left\{P\left(Y_i=1|\check{\mathbf{X}}_i, G_i\right)\right\}=\beta_1^{\intercal}\check{\mathbf{X}}_i I\,(G_i=1)+\beta_0^{\intercal}\check{\mathbf{X}}_i I\,(G_i=0), \quad (4.2)$$

where $\beta_1 = (0.1, 0., 0.1, 0.15, 0.2, 0.1)^{\top}$ and $\beta_0 = (0.6, 0.25, -0.2, -0.1, 0.15, -0.2)^{\top}$.

To derive ITRs, we fitted various working models with either linear (for continuous *Y*) or logistic regression (for binary *Y*) with the following different sets of covariates:

$$\begin{array}{llll} (M_1): & \mathbf{X}_{M_1}=\check{\mathbf{X}} & (M_2): & \mathbf{X}_{M_2}=\left(\check{\mathbf{X}}^{\intercal}, X_6, X_7, \ldots, X_{20}\right)^{\intercal} \\ (M_3): & \mathbf{X}_{M_3}=(1, X_2, X_4, X_5)^{\intercal} & (M_4): & \mathbf{X}_{M_4}=(1, X_1, X_2, X_3)^{\intercal}. \end{array}$$

For the overfitted model (M_2), we generated $X_6 \sim$ Bernoulli(0.7) and $(X_7, \ldots, X_{20})^{\top}$ $N(\mu, 8 \cdot I_{14\times 14} + 8)$, where $\mu = (3, 1, 2, -3, 3, 1, 2, 1.5, 2.5, 1, 0.5, 2, 1, 1)^{\top}$; $I_{14\times 14}$ denotes a 14 × 14 identity matrix. To obtain model based estimate of $D(\mathbf{X})$, we fit simple linear regression and logistic regression models with penalty parameter λ_{jn} set to 0 since the models sizes are reasonably small. It can be seen that $\Delta(\cdot)$ is an increasing function under those four working models and hence $\hat{\mathbb{V}}_{\hat{\mathscr{I}}_{calib}}$ is asymptotically equivalent to $\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$ and $\hat{\mathbb{V}}_{\hat{I}_{\hat{\phi}}}$.

The results for estimating $\mathbb{V}_{\mathscr{I}_{calib}}$ via $\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$, $\hat{\mathbb{V}}_{\hat{I}_{\hat{\phi}}}$, and $\hat{\mathbb{V}}_{\hat{\mathscr{I}}_{calib}}$ together with a 2-fold CV procedure are shown in Table 1. For all estimators of $\mathbb{V}_{\mathscr{I}_{calib}}$, the biases are negligible and the estimated SEs are close to the empirical SEs. The 95% CIs have empirical coverage levels close to the nominal level. As expected, the CV procedure generally provides estimators with lower bias compared to the apparent estimates. Since $\mathscr{I}_{calib}$ and $\hat{I}_{\hat{c}}$ are equivalent under monotonicity, the results for $\hat{\mathbb{V}}_{\hat{\mathscr{I}}_{calib}}$ and $\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$ are almost identical to each other.

Although $\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$ and $\hat{\mathbb{V}}_{\hat{I}_{\hat{\phi}}}$ are also asymptotically equivalent, we note that $\hat{\mathbb{V}}_{\hat{I}_{\hat{\phi}}}$ tends to have slightly larger bias and variation in finite samples. This could be in part due to the larger

variability in estimating the corresponding optimal threshold values as shown in Table 2. The bias and variance of $\hat{\phi}$ are substantially larger than those of $\hat{c}$. The efficiency of $\hat{\phi}$ relative to $\hat{c}$ is only about 66-75% for most of the settings we considered. This confirms the disadvantage of estimating $c°$ by directly maximizing the non-smooth empirical objective function $\hat{\mathbb{V}}_{\hat{I}_c}$.

We also evaluated the finite sample performance of the inference procedure for the IncV of new markers. In Table 3, we presented results on the estimated IncV of X_1 and X_3 in improving the value function by comparing the ITRs derived from the full model (M_1) to those from model (M_3) with no information on X_1 and X_3. Our proposed procedures based on $\hat{\mathscr{J}}_{calib}$ and $\hat{\mathbb{V}}_{\hat{\mathscr{J}}_{calib}}$ give minimally biased estimates of the IncV and the resampling procedures provide good estimates for the SEs and CIs. Since the model sizes are small, both the apparent estimates and the CV estimates lead to reasonable interval estimates with proper coverage levels. For comparison, we provided the results based on $\hat{I}_{\hat{\phi}}$ as well which also leads to consistent estimates of the IncV since $\Delta(\cdot)$ is monotone for both models. Similar to the results shown earlier, the direct maximization of $\hat{\mathbb{V}}_{\hat{I}_c}$ resulted in larger variability in $\hat{I}_{\hat{\phi}}$, which subsequently lead to higher variability in estimating IncV in finite sample.

4.2 Comparisons to Existing Methods

Additional simulation studies were conducted to compare our calibration methods with those proposed in Qian and Murphy (2011) (QM) and Zhang et al. (2012a) (Zhang). The QM method employs model based ITRs but guard against model mis-specification by including non-linear basis functions and then obtains stable parameter estimates by imposing L_1 penalization. In our simulations, we included all linear effects and two-way interactions for their method. To apply the Zhang method, we let the propensity score be 0.5 and search for the optimal ITR using the linear regression followed by a CART procedure. The complexity parameter was first set at 0.001 to build a large tree which is then pruned via a 10-fold CV.

To compare the performance of these methods, we generated $\mathbf{X}_{20\times 1} \sim N(0, 2.4I_{20\times 20}+1.6)$ and a continuous Y that depends non-linearly to the first 5 covariates through

$$Y_i=\beta_1^{\intercal}\tilde{\mathbf{X}}_i I\,(G_i=1)+\beta_0^{\intercal}\tilde{\mathbf{X}}_i I\,(G_i=0)+\epsilon_i \qquad with \qquad \epsilon_i \sim N\,(0,1) \quad (4.3)$$

where $\tilde{\mathbf{X}}=\left(1, e^{X_1/3}, X_2, e^{X_3/4}, X_4, I\,(X_5>0), X_2^2, \sin(X_4)\right)^{\top}$, β_0 = (12, 5, −6, −2, 3, −4, 7, −2)$^{\top}$ and β_1 = (2, 1, 2, 3, 2, 2, 3, 3)$^{\top}$. To construct the ITRs, we investigated four sets of working linear models with covariates being

$$\left(M_1'\right):\mathbf{X}_{M_1'}=\tilde{\mathbf{X}} \qquad \left(M_2'\right):\mathbf{X}_{M_2'}=\left(\tilde{\mathbf{X}}^{\intercal}, X_6, X_7, \ldots, X_{20}\right)^{\intercal}$$
$$\left(M_3'\right):\mathbf{X}_{M_3'}=(1, X_1, X_2, X_3, X_4, X_5)^{\intercal} \qquad \left(M_4'\right):\mathbf{X}_{M_4'}=(1, X_1, X_2, X_3)^{\intercal}.$$

For (M_3') and (M_4'), Δ(·) is non-monotone. Since our proposal also allows non-linear transformations, we derive $\hat{\mathscr{I}}_{calib}$ based on $D\left(\hat{\beta}, \mathbf{X}\right)$ using (i) linear models ($\hat{\mathscr{I}}^{L}_{calib}$) and (ii) linear + two-way interactions with ridge penalty ($\hat{\mathscr{I}}^{L+I}_{calib}$). For comparison, we also consider the model based rule by assigning patients according to $D\left(\hat{\beta}, \mathbf{X}\right) \geq 0$ for both linear effects (Model^{L}) and linear + interaction effects (Model$L+I$). The ridge penalty parameters for Model$L+I$ and $\hat{\mathscr{I}}^{L+I}_{calib}$ were chosen via the CV. For all scenarios, we estimated ITRs with training sets of sizes $n_0 = n_1 = 250$ and evaluated their corresponding value functions on independent validation datasets with $n_0 = n_1 = 10^5$. We use large validation sets to ensure that the variability observed in the empirical value function reects mainly the variability due to estimating the ITRs. Table 4 summarizes the mean and empirical SE (ESE) of the empirical value functions for (1) Model^{L}; (2) $\hat{\mathscr{I}}^{L}_{calib}$; (3) Model^{L+I} ; (4) $\hat{\mathscr{I}}^{L+I}_{calib}$; (5) QM; and (6) Zhang.

Under the true model (M_1') with 5 covariates, all methods perform very similarly. Our calibrated method performs almost identically to the model based method, suggesting that little price is paid for the additional calibration. The QM method also did not pay much price for including the non-informative two-way interactions in this case. With the over-fitted model (M_2'), except for Model^{L+I}, all other methods achieve similar level of value function with slight difference in the variability. The QM method performs quite well considering that about 200 covariates are included in the fitting with only 5 are informative. This is not too surprising since L_1 penalization is expected to work well when the signals are strong and sparse, as in the present case. Both the QM and our method with $\hat{\mathscr{I}}^{L}_{calib}$ achieved value functions almost identical to those obtained under (M_1'). Our method with $\hat{\mathscr{I}}^{L+I}_{calib}$ and the Zhang method result in slightly higher variability compared to the QM and $\hat{\mathscr{I}}^{L}_{calib}$. Model^{L+I} has a slightly lower value function with larger variability, suggesting instability in ridge penalized modeling fitting. However, the calibration appears to have the ability to reduce the instability with a much more stable $\hat{\mathscr{I}}^{L+I}_{calib}$ compared to Model^{L+I}. In general, the price paid for overfitting under correct model specification seems to be relatively low. This is in part because all methods are building on top of correctly specified models and hence the estimated regression parameters are maximizing the value function asymptotically. As a result the variability due to estimating all parameters in the ITR contributes at a second order, similar to those argued in Zhao et al. (2012).

With the mis-specified working models (M_3') and (M_4'), our proposed calibration method performs better than all other competing methods with respect to the achievable value function and/or the variability. For example, for (M_3'), the average value function was 41.95 and 41.99 for $\hat{\mathscr{I}}^{L}_{calib}$ and $\hat{\mathscr{I}}^{L+I}_{calib}$ respectively, 40.61 for the QM method, and 41.75 for the Zhang method. Since the true underlying effects are quite non-linear, the model based method with linear effects perform poorly in this setting with value function of only 39.65. It seems that there is a slight increase in the achievable value function from our calibration

method by including the non-linear basis in the working models. Under mis-specified models, the calibration method tends to produce more stable ITRs than those from the QM and Zhang methods. For example, the ESE was 0.17 and 0.15 for $\hat{\mathscr{I}}^{L}_{calib}$ and $\hat{\mathscr{I}}^{L+I}_{calib}$ respectively, 0.43 for the QM method, and 0.72 for the Zhang method.

5. Example: Application to an HIV/AIDS Clinical Trial

In this section, we use a randomized clinical trial from the AIDS Clinical Trials Group Protocal 320 (ACTG 320) to illustrate our proposed methods. A total of 1156 zidovudine-experienced patients with advanced human immunodeficiency virus (HIV) type-I infection were enrolled in ACTG 320 and assigned to either a 2-drug combination of zidovudine and lamivudine (treatment 0) or a 3-drug combination of zidovudine, lamivudine and indinavir (treatment 1). The objective of this study was to assess the added value of a protease inhibitor (indivar) to the dual nucleoside reverse-transcriptase inhibitors (xidovudine and lamivudine) (Hammer et al., 1997). The overall success of the 3-drug combination on various study endpoints was so significant that the study was terminated early by the Data Safety Monitoring Board. However, since the benefit of 3-drug combination therapy over the 2-drug alternative potentially differs across patients, it would be interesting to identify subgroups of patients who can be managed almost as well using the less potent 2-drug therapy.

For our analyses, the potential baseline predictors for constructing ITRs consist of age (years), $CD4_0$, $logCD4_0$, and $\log_{10}RNA_0$, where $CD4_k$ and RNA_k denote, respectively, the CD4 cell count (cells/mm^3) and the RNA measure (copies/ml) at week k. We restricted our study to the 856 subjects who had complete information on these variables and on the outcome of interest, 427 of which were in the 3-drug combination group. Because it is relatively expensive to measure RNA, particularly in resource-limited settings, it would be of interest to examine whether $\mathbf{X}_{new} = \log_{10}RNA_0$ is useful for improving the ITRs. Thus, we compared the optimal ITRs based on the various working models with the following two sets of predictors: (M_{old}) : $\mathbf{X}_{old}=(1, \text{Age}, CD4_0, logCD4_0)^{\top}$

$$(M_{update}) : \mathbf{X}_{update}=(1,\ Age,\ CD4_0,\ logCD4_0,\ log_{10}\, RNA_0)^{\intercal}$$

To quantify the effectiveness of the therapy, we considered the immune response Y defined as the change in logCD4 from baseline to week 24. Table 5 summarizes the estimated regression coefficients for fitted linear models with (M_{update}) and (M_{old}). To construct optimal ITRs under these two models, we let $\xi = 0.277$, which is about the within subject variability of logCD4 counts (Hughes et al., 1994), indicating that 3-drug therapy is only preferred if the gain in immune response exceeds ξ. All estimators of the value functions are based on 500 repeated two-fold CVs. The SEs are based on 500 perturbations.

Under (M_{update}), the maximum cost-adjusted value function is $\hat{\mathbb{V}}_{\hat{\mathscr{I}}^{update}_{calib}} = 0.610$ with a 95% CI (0.526, 0.692). Its corresponding optimal threshold is equal to 0.236 with a 95% CI=(−0.040, 0.512). Under (M_{old}), i.e. without the RNA information, we have $\hat{\mathbb{V}}_{\hat{\mathscr{I}}^{old}_{calib}}$

=0.606 with a 95% CI (0.522, 0.690) and $\hat{c}$ = 0.231 with a 95% CI=(−0.004, 0.466). Finally, the estimated difference $\hat{IncV}$ is equal to 0.004 with the 95% CI (−0.035, 0.043). This implies that despite the highly significant difference between the effect of RNA on the outcome between the two treatment groups (as shown in Table 5), including RNA information in the ITR does not improve the value function. Therefore, RNA is not necessary to better assign future patients to a specific treatment. By including two-way interactions of all variables under (M_{update}) and (M_{old}), our calibrated method yields ITRs with estimated values 0.614 for (M_{update}) and 0.609 for (M_{old}), similar to their linear effect counterparts. We also applied the QM and Zhang methods to ACTG 320 data set and observed slightly lower value functions compared to those from our methods. The value function associated with the models (M_{update}) and (M_{old}) are, respectively, 0.594 and 0.585 based on the Zhang method, and 0.578 and 0.569 based on the QM method.

It would also be of interest to compare to the simple rules that assign all patients to the 3-drug combination group or to the 2-drug group. The value functions are estimated as 0.562 if all treated with 3-drug and 0.177 if all treated with 2-drug. Comparing to assigning all patients to 3-drug, $\hat{\mathscr{I}}_{calib}^{update}$ leads to an increased value of 0.48 with a 95% CI [0.003, 0.093], suggesting an improvement by adopting the ITR.

6. Remarks

In this paper, we have proposed a robust procedure that uses multiple baseline covariates to develop and evaluate ITRs. While the procedure builds upon an existing two-step framework, this paper provides additional insights into how to develop an optimal ITR based on working models and how to evaluate such ITRs. Fitting the data with semi-parametric models in step I, combined with the non-parametric estimation in step II, provides robust estimates of ITRs and valid estimates of their associated cost-adjusted population average outcomes. Our proposed ITRs are optimal across all rules based on the given set of covariates when the fitted working models are correct or nearly correct as discussed in section 2. The ITR is optimal among all rules based on the estimated scores when the fitted models are mis-specified. To account for the variability in estimating various parameters, we proposed perturbation-resampling procedures that can be used to assess the variability in the estimators.

While in traditional statistical methods, model misspecification may hinder predictions and lead to inaccurate assignment rules, the methods developed in this paper rely on a layer of calibration in step II to guard against model misspecification. We provide justifications for when the proposed procedures lead to ITRs that are globally optimal under correct or near correct model specification and optimal within a class of ITR rules under model misspecification. Under model misspecification, the proposed ITR could be suboptimal when compared to the global optimal ITR. Hence, it would be crucial to provide working models that can approximate the true model reasonably well. Incorporating non-linear effects through basis function specification could be helpful and one may use the proposed inference procedure for comparing value functions as a tool for selecting important bases functions.

Through simulation studies and theoretical studies, we also demonstrate that direct optimization of the empirical value function $\hat{\mathbb{V}}_{\hat{f}_c}$ could lead to rather unstable ITRs, with high variability and slow convergence rate in estimating the optimal threshold values, even when the underlying conditional treatment difference function $\Delta(c)$ is monotone in c.

When there are new biomarkers introduced to assist in treatment selection, it is important to evaluate their value in improving population average outcomes. It is important to note that a variable highly differentially associated with $Y^{(1)}$ and $Y^{(0)}$ may not necessarily be important for improving ITRs. This is somewhat similar to the phenomenon observed in the risk prediction literature: a variable highly significant in regression modeling may not result in large improvement in prediction. However, the IncV with respect to improving ITRs is more subtle than the typical prediction setting. Taking an extreme case scenario with a single marker X and cost $\xi = 0$, suppose $D(X)$ is strictly increasing and $D(X) >> 0$ for all X. Then obviously X will be selected as important for predicting treatment response by any variable selection procedure. On the other hand, X would be deemed as not important for treatment selection since all subjects should be assigned to treatment 1 and having the information on X does not change the treatment assignment or the corresponding value function. Thus, if one is interested in measuring the importance of markers in guiding treatment selection, it would be valuable to directly assess the IncV with respect to the value function as proposed in this paper. When the dimension of new markers is not small, it would be crucial to employ the cross-validation to correct for the overfitting bias as suggested by Zhao et al. (2013). Procedures for efficiently selecting the informative markers warrant further research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The work is partially supported by grants R01- GM079330, R01 HL089778, R01 AI052817, and T32 NS048005 awarded by the National Institutes of Health.

References

Baker S, Kramer B, Sargent D, Bonetti M. Biomarkers, subgroup evaluation, and clinical trial design. Discovery Medicine. 2012; 13:187–192. [PubMed: 22463794]

Bonetti M, Gelber R. A graphical method to assess treatment–covariate interactions using the Cox model on subsets of the data. Statistics in medicine. 2000; 19:2595–2609. [PubMed: 10986536]

Bonetti M, Gelber R. Patterns of treatment effects in subsets of patients in clinical trials. Biostatistics. 2004; 5:465–481. [PubMed: 15208206]

Cai T, Tian L, Wei L. Semiparametric Box–Cox power transformation models for censored survival observations. Biometrika. 2005; 92:619–632.

Cai T, Tian L, Wong P, Wei L. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011; 12:270–282. [PubMed: 20876663]

Foster J, Taylor J, Ruberg S. Subgroup identification from randomized clinical trial data. Statistics in Medicine. 2011; 30:2867–2880. [PubMed: 21815180]

Hammer S, Squires K, Hughes M, Grimes J, Demeter L, Currier J, Eron J Jr, Feinberg J, Balfour H Jr, Deyton L, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human

immunodeficiency virus infection and cd4 cell counts of 200 per cubic millimeter or less. New England Journal of Medicine. 1997; 337:725–733. [PubMed: 9287227]

Holland PW. Statistics and causal inference. Journal of the American statistical Association. 1986; 81:945–960.

Huang Y, Gilbert PB, Janes H. Assessing treatment-selection markers using a potential outcomes framework. Biometrics. 2012; 68:687–696. [PubMed: 22299708]

Hughes MD, Stein DS, Gundacker HM, Valentine FT, Phair JP, Volberding PA. Within-subject variation in cd4 lymphocyte count in asymptomatic human immunodeficiency virus infection: implications for patient monitoring. Journal of Infectious Diseases. 1994; 169:28–36. [PubMed: 7903975]

Imai K, Strauss A. Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. Political Analysis. 2011; 19:1–19.

Janes H, Pepe M, Bossuyt P, Barlow W. Measuring the performance of markers for guiding treatment decisions. Annals of Internal Medicine. 2011; 154:253–259. [PubMed: 21320940]

Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. Journal of American Medical Association. 2007; 298:1209–1212.

Kim J, Pollard D. Cube root asymptotics. The Annals of Statistics. 1990; 18:191–219.

Park Y, Wei L. Estimating subject-specific survival functions under the accelerated failure time model. Biometrika. 2003; 90:717–723.

Qian M, Murphy S. Performance guarantees for individualized treatment rules. Annals of Statistics. 2011; 39:1180–1210. [PubMed: 21666835]

Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling. 1986; 7:1393–1512.

Rothwell P. Can overall results of clinical trials be applied to all patients? The Lancet. 1995; 345:1616–1619.

Rothwell P, Mehta Z, Howard S, Gutnikov S, Warlow C. From subgroups to individuals: general principles and the example of carotid endarterectomy. The Lancet. 2005; 365:256–265.

Rubin D. Randomization analysis of experimental data: The fisher randomization test comment. Journal of the American Statistical Association. 1980; 75:591–593.

Rubin D. Statistics and causal inference: Comment: Which ifs have causal answers. Journal of the American Statistical Association. 1986; 81:961–962.

Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology. 1974; 66:688.

Rubin DB, et al. Bayesian inference for causal effects: The role of randomization. The Annals of Statistics. 1978; 6:34–58.

Song X, Pepe M. Evaluating markers for selecting a patient's treatment. Biometrics. 2004; 60:874–883. [PubMed: 15606407]

Song X, Zhou X-H. Evaluating markers for treatment selection based on survival time. UW Biostatistics Working Paper Series pages Working Paper. 2009; 349:1–27.

Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. Stat. 2012a; 1:103–114. [PubMed: 23645940]

Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012b; 68:1010–1018. [PubMed: 22550953]

Zhao L, Tian L, Cai T, Claggett B, Wei L. Effectively selecting a target population for a future comparative study. Journal of the American Statistical Association. 2013; 108:527–539. [PubMed: 24058223]

Zhao Y, Zeng D, Rush A, Kosorok M. Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association. 2012; 107:1106–1118. [PubMed: 23630406]

Table 1

Empirical bias (Bias), empirical standard error (ESE), average of the estimated standard error (ASE) and the empirical coverage level of the 95% CIs (95%-CP) for $\mathbb{V}_{\mathscr{I}_{calib}}$ via $\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$ and $\hat{\mathbb{V}}_{\mathscr{I}_{calib}}$ under various working models.

(a) Continuous outcome

				Apparent			2-fold CV		
Model	**Method**	**True**	**ASE**	**Bias**	**ESE**	**95%-CP**	**Bias**	**ESE**	**95%-CP**
	$\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$		0.629	0.109	0.609	0.952	0.073	0.602	0.955
M_1	$\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$	9.510	0.621	0.047	0.605	0.946	0.026	0.602	0.947
	$\hat{\mathbb{V}}_{\mathrm{I}_{\mathrm{calib}}}$		0.621	0.047	0.605	0.946	0.026	0.602	0.947
	$\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$		0.628	0.108	0.610	0.949	0.076	0.605	0.954
M_2	$\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$	9.510	0.621	0.048	0.603	0.946	0.022	0.600	0.946
	$\hat{\mathbb{V}}_{\mathrm{I}_{\mathrm{calib}}}$		0.621	0.048	0.603	0.946	0.022	0.600	0.946
	$\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$		0.632	0.124	0.629	0.948	0.089	0.623	0.944
M_3	$\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$	8.842	0.621	0.044	0.620	0.945	−0.020	0.616	0.945
	$\hat{\mathbb{V}}_{\mathrm{I}_{\mathrm{calib}}}$		0.621	0.044	0.620	0.945	−0.020	0.616	0.945
	$\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$		0.639	0.212	0.633	0.962	0.146	0.630	0.956
M_4	$\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$	8.834	0.637	0.137	0.627	0.955	0.107	0.623	0.944
	$\hat{\mathbb{V}}_{\mathrm{I}_{\mathrm{calib}}}$		0.630	0.074	0.621	0.951	0.057	0.613	0.944

(b) Binary outcome

				Apparent			2-fold CV		
Model	**Method**	**True**	**ASE**	**Bias**	**ESE**	**95%-CP**	**Bias**	**ESE**	**95%-CP**
	$\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$		0.042	0.028	0.037	0.927	0.016	0.037	0.935
M_1	$\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$	0.766	0.039	0.018	0.036	0.948	0.008	0.036	0.963
	$\hat{\mathbb{V}}_{\mathrm{I}_{\mathrm{calib}}}$		0.039	0.019	0.035	0.946	0.007	0.036	0.960
	$\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$		0.043	0.042	0.037	0.905	0.010	0.037	0.972
M_2	$\hat{\mathbb{V}}_{\hat{I}_{\hat{c}}}$	0.766	0.041	0.033	0.035	0.920	−0.005	0.036	0.974
	$\hat{\mathbb{V}}_{\mathrm{I}_{\mathrm{calib}}}$		0.040	0.034	0.035	0.921	−0.004	0.036	0.976

(b) Binary outcome

				Apparent			2-fold CV		
Model	**Method**	**True**	**ASE**	**Bias**	**ESE**	**95%-CP**	**Bias**	**ESE**	**95%-CP**
M_3	$\hat{V}_{\hat{I}_{\hat{C}}}$	0.754	0.041	0.045	0.036	0.902	0.029	0.037	0.933
	$\hat{V}_{\hat{I}_{\hat{C}}}$		0.039	0.029	0.035	0.946	0.019	0.035	0.949
	$\hat{V}_{I_{calib}}$		0.038	0.030	0.035	0.940	0.020	0.035	0.943
M_4	$\hat{V}_{\hat{I}_{\hat{C}}}$	0.749	0.048	0.039	0.042	0.900	0.031	0.041	0.915
	$\hat{V}_{\hat{I}_{\hat{C}}}$		0.042	0.032	0.041	0.915	0.024	0.040	0.924
	$\hat{V}_{I_{calib}}$		0.041	0.019	0.040	0.924	0.015	0.039	0.940

Table 2

Empirical bias (Bias), empirical standard error (ESE), average of the estimated standard errors (ASE) and empirical coverage levels of 95% CIs (95%-CP) for estimating the optimal threshold c based on $\hat{c} = \text{argmax}_c \hat{\mathbb{V}}_{\hat{f}_c}$ and $\hat{c} = \{c : \hat{\Delta}(c) = 0\}$.

		Continuous					Binary				
Model	**Method**	**True**	**Bias**	**ESE**	**ASE**	**95%-CP**	**True**	**Bias**	**ESE**	**ASE**	**95%-CP**
M_1	$\hat{c}$	0.00	−0.178	1.667	1.753	0.960	0.00	−0.012	0.115	0.121	0.961
	$\bar{c}$		0.036	1.358	1.407	0.967		−0.001	0.100	0.106	0.971
M_2	$\hat{c}$	0.00	−0.201	1.647	1.757	0.964	0.00	−0.021	0.118	0.128	0.958
	$\bar{c}$	−0.0	−0.007	1.355	1.407	0.972		0.004	0.112	0.118	0.972
M_3	$\hat{c}$	−0.019	−0.212	1.764	1.851	0.967	−0.023	−0.022	0.109	0.114	0.954
	$\bar{c}$		0.101	1.480	1.550	0.965		−0.008	0.094	0.100	0.964
M_4	$\hat{c}$	−0.121	0.190	1.693	1.853	0.979	0.017	−0.030	0.103	0.105	0.951
	$\bar{c}$		0.071	1.491	1.547	0.946		−0.018	0.099	0.102	0.964

Table 3

Empirical bias (Bias), empirical standard errors (ESE), average of the estimated standard errors (ASE) and empirical coverage levels of the 95% CIs for the IncV comparing the optimal ITR under the full model versus the reduced model, where the optimal ITRs are estimated based on $\hat{I}_{\hat{\phi}}$, and $\hat{\mathscr{I}}_{calib}$.

				Apparent			2-fold CV		
Response	**Method**	**True**	**ASE**	**Bias**	**ESE**	**95%-CP**	**Bias**	**ESE**	**95%-CP**
Continuous	$\hat{I}_{\hat{\phi}}$	0.663	0.215	0.025	0.217	0.946	−0.019	0.218	0.948
	$\hat{I}_{calib}$		0.187	0.013	0.183	0.948	−0.007	0.184	0.952
Binary	$\hat{I}_{\hat{\phi}}$	0.011	0.026	0.005	0.024	0.973	0.002	0.024	0.966
	$\hat{I}_{calib}$		0.021	0.004	0.020	0.966	0.001	0.019	0.957

Table 4

Empirical mean (Mean) and the corresponding empirical standard error (ESE) for $\mathbb{V}_{\mathscr{I}_{calib}}$ via different methods under various working models.

		Model^{L}	**$\hat{I}^{L}_{calib}$**	**Model^{L+I}**	**$\hat{I}^{L+I}_{calib}$**	**QM**	**Zhang**
M_1'	Mean	42.39	42.39	42.39	42.39	42.38	42.06
	ESE	0.13	0.13	0.13	0.13	0.13	0.15
M_2'	Mean	42.39	42.39	40.90	42.01	42.35	42.04
	ESE	0.13	0.13	0.51	0.16	0.14	0.16
M_3'	Mean	39.65	41.95	40.47	41.99	40.61	41.75
	ESE	0.95	0.17	0.39	0.15	0.43	0.72
M_4'	Mean	39.61	41.89	40.42	41.92	40.58	41.72
	ESE	1.01	0.17	0.52	0.15	0.57	0.70

Table 5

Estimated regression coefficients for the immune response (change in CD4 counts from baseline to week 24) based on the ACTG 320 Data.

Model	Treatment		Intercept	Age	$CD4_0$	$\log(CD4_0)$	$\log_{10}(RNA_0)$
M_{update}	2-drug	Estimate	1.065	0.002	0.004	−0.338	0.011
		Std. Error	0.350	0.004	0.001	0.057	0.049
		p-value	0.002	0.568	< 0.001	< 0.001	0.829
	3-drug	Estimate	1.317	0.009	−0.001	−0.328	0.128
		Std. Error	0.348	0.004	0.001	0.062	0.051
		p-value	< 0.001	0.023	0.199	< 0.001	0.013
M_{old}	2-drug	Estimate	1.124	0.002	0.004	−0.338	
		Std. Error	0.218	0.004	0.001	0.057	
		p-value	< 0.001	0.584	< 0.001	< 0.001	
	3-drug	Estimate	1.984	0.008	−0.002	−0.318	
		Std. Error	0.225	0.004	0.001	0.062	
		p-value	< 0.001	0.035	0.066	< 0.001	