11-15-2013

Characterizing Expected Benefits of Biomarkers in Treatment Selection

Ying Huang
Fred Hutchinson Cancer Center, yhuang@fhcrc.org

Eric Laber
North Carolina State University, eblaber@ncsu.edu

Holly Janes
Fred Hutchinson Cancer Research Center, hjanes@fhcrc.org

Suggested Citation

Huang, Ying; Laber, Eric; and Janes, Holly, "Characterizing Expected Benefits of Biomarkers in Treatment Selection" (November 2013). *UW Biostatistics Working Paper Series.* Working Paper 396.
http://biostats.bepress.com/uwbiostat/paper396

1 Introduction

In many clinical settings for disease prevention and control, there is significant heterogeneity in patient response to an intervention. Biomarkers associated with this heterogeneity, such as demographic or genetic characteristics, can be used to help patients and clinicians select interventions such that a therapy is only delivered to a patient who is likely to benefit from it. Examples of such treatment selection markers include: the K-RAS gene mutations for selecting antiepidermal growth factor receptor therapy for colorectal cancer (Karapetis *and others*, 2008; Allegra *and others*, 2009); the FGFR2 and MRP30 genes for selecting hormone therapy in breast cancer prevention of postmenopausal women (Prentice *and others*, 2009; Huang *and others*, 2011); and the Oncotype-Dx, a 21-gene score for selecting adjuvant chemotherapy for the treatment of estrogen-receptor positive breast cancer (Paik *and others*, 2006; Harris *and others*, 2007; Albain *and others*, 2010).

Statistical measures of the value of a marker-based treatment selection rule are essential in the development of treatment selection markers. The test of a marker by treatment interaction is a common strategy for identifying treatment selection markers, however, recently there has been a growing emphasis on developing measures of treatment selection ability which are directly linked to clinical outcomes (Janes *and others*, 2011). Much of this work focuses on the effect of treatment on the targeted disease outcome of interest. For example, the population benefit of using a marker-based treatment-selection strategy, defined as the reduction in population disease rate through treatment-selection (Song and Pepe, 2004; Zhang *and others*, 2012ab, 2013; Zhao *and others, 2012*); the accuracy for classifying a subject into treatment-effective or ineffective categories based on a potential outcomes framework

(Huang *and others*, 2012); the difference in average treatment effect between marker-positive group and the overall population (Foster *and others*, 2011); the distribution of the disease risk difference conditional on a marker or marker model between comparative interventions (Cai *and others*, 2011; Huang *and others*, 2012; Janes *and others*, 2013); and the average treatment difference in a subgroup of subjects whose estimated treatment difference were greater than some threshold (Zhao *and others*, 2013) all quantify, in some fashion, the impact of the marker on the rate of disease. In addition to its effect on the targeted disease of interest, an intervention may impact the population through other costs such as side effect burden or monetary cost. Thus, another important consideration in assessing a treatment selection rule is the proportion of subjects selected for treatment (Janes *and others*, 2013). If two treatment selection strategies are equally effective in controlling population disease rate, then the strategy that recommends fewer people for treatment is more appealing due to less treatment cost. One approach to incorporate both the effect on targeted disease and other treatment effects in the assessment of treatment selection markers, is to use a decision-theoretic framework which puts the treatment effect on disease and additional treatment effects on the same scale by means of a treatment/disease cost ratio. For example, one can define a net benefit measure as the reduction in overall disease and treatment cost comparing a marker-based treatment strategy with a default strategy of treating no one (Vickers *and others*, 2007; Rapsomaniki *and others*, 2012; Janes *and others*, 2013).

While useful for comparing treatment selection strategies, the net benefit of a particular marker-based strategy does not provide a direct quantification of the benefit gained by the marker when treatment/disease cost ratio varies, since the default strategy of no treatment may or may not be the optimal strategy absent the marker information. In this paper, we develop the *expected benefit*, a new measure built upon the net benefit, to quantify the benefit in treatment selection gained by adopting a marker-based rule. This measure quantifies the reduction in overall population cost due to both disease and treatment comparing a marker-based treatment selection rule with the optimal treatment strategy absent the marker information. The proposed method is built upon an earlier decision-theoretic framework (Vickers *and others*, 2007; Rapsomaniki *and others*, 2012; Janes *and others*, 2013), but further accommodates the fact that optimal treatment in the absence of marker information varies with the cost of treatment. In addition, we propose a novel method to standardize the expected benefit of a treatment selection strategy relative to the benefit that can potentially be achieved via a perfect treatment selection rule. While the latter is in general not identifiable, we show that upper and lower bounds instead can be established based on a potential outcomes framework. Using a generalized linear model of disease risk as a function of marker and treatment, we consider the problem of maximizing the expected benefit as a function of treatment/disease cost ratio. We develop model-based estimators for the corresponding expected benefit and its standardized value in a randomized trial setting and develop asymptotic theory for the estimators. The expected benefit is not a smooth function of the generative model causing the standard bootstrap to fail; we develop a novel adaptive bootstrap confidence interval that provides consistent inference. We also investigate alternative estimators of the optimal expected benefit based on a working model, which are robust to model misspecification.

In Section 2, we introduce the concept of expected benefit, derive bounds on the expected benefit of a perfect selection rule, and define the standardized expected benefit. We develop estimation methods and theoretical results in Section 3. Simulation studies are presented in Section 4 where we investigate finite sample performance of the estimators. Application of the methodology to an example in the Diabetes Control and Complications Trial is presented in Section 5 where we evaluate the expected benefit of the baseline hemoglobin A1C for selecting diabetes treatment. We then summarize the paper and make concluding remarks.

2 Methods

We consider the setting of a randomized trial with two arms, $T = 0, 1$ indicating the placebo and treatment arms respectively. Let D be a binary outcome that the treatment targets, which we call 'disease,' with $D = 0, 1$ indicating control and case status respectively. For example, in a disease prevention setting, D may indicate disease status; whereas in the setting of treating patients with an existing disease, D may indicate disease recurrence or disease-related death. Let $\rho_0 = P(D = 1|T = 0)$ and $\rho_1 = P(D = 1|T = 1)$ indicate disease prevalence in placebo and treatment arms. Let Y denote the, possibly multivariate, biomarker of interest. Let $A(Y)$ be a treatment-selection rule based on the marker, which takes value 1,0 corresponding to the recommendation for or against the treatment. Let i be subject indicator. With a randomized trial with N participants, the data we observe are i.i.d $(Y_i, T_i, D_i), i = 1, \ldots, N$.

As in Vickers *and others* (2007), we assume the cost of the treatment either due to side effects or due to monetary cost can be quantified as c times the cost per disease outcome, where c is a non-negative utility parameter indicating the ratio of treatment cost relative to disease cost. For example, in Vickers *and others* (2007) c was chosen to be 5% for treating breast cancer with adjuvant therapy according to a patient survey; in Rapsomaniki *and others* (2012), c was chosen to be 20% for preventing the emergence of cardiovascular disease within 10 years using statin.

Without loss of generality, let the cost of per disease outcome be 1, then the total cost due to disease and treatment averaged across a population for a strategy $A(Y)$ and for the optimal treatment strategy absent the marker information can be computed respectively as follows in units of disease cost.

For cost ratio c, the population-averaged total cost due to disease and treatment by applying a treatment selection rule $A(Y)$ is $E_A(D) + P\{A(Y) = 1\} \times c$, where the expectation

$E_A(D)$ is taken with respect to the rule $A(Y)$. Based on a randomized trial where T is randomized to subjects, the cost can be expressed as

$$\begin{aligned} & P\{D=1, A(Y)=0|T=0\} + P\{D=1, A(Y)=1|T=1\} + P\{A(Y)=1\} \times c \\ = \; & P(D=1|T=0) - [E\{DA(Y)|T=0\} + E\{DA(Y)|T=1\}] + E\{A(Y)\} \times c \\ = \; & \rho_0 - E\left[A(Y)\left\{\Delta(Y) - c\right\}\right], \end{aligned} \tag{1}$$

where $\Delta(Y) = P\left(D=1|T=0,Y\right) - P\left(D=1|T=1,Y\right)$ is the risk difference conditional on Y between placebo and treatment arms.

Without any biomarker information, the optimal treatment selection rule that minimizes the total disease plus treatment cost is to treat everyone if $\rho_0 - \rho_1 > c$ and treating no one otherwise. Thus, the total cost based on this treatment selection rule is

$$\rho_0 - \left[\rho_0 - \rho_1 - c\right]_+, \tag{2}$$

where $[u]_+ = max(0,u)$ is the positive-part function.

We define the expected benefit $EB_A(c)$ for a treatment-strategy $A(Y)$ and cost ratio c as the difference between (2) and (1), i.e., the reduction in the total cost using $A(Y)$ relative to the optimal rule in the absence of biomarkers:

$$EB_A(c) = E\left[A(Y)\left\{\Delta(Y) - c\right\}\right] - \left[\rho_0 - \rho_1 - c\right]_+. \tag{3}$$

Note that the first component of (3) is exactly the net benefit measure considered by Vickers *and others* (2007), Rapsomaniki *and others* (2012), and Janes *and others* (2013). That is, the *expected benefit* equals the net benefit when the optimal strategy absent marker information is to treat no one. The second component of (3) can be viewed as the net benefit of an optimal treatment selection rule absent any marker information. Thus, the expected benefit of a marker-based treatment selection rule can be interpreted as the incremental value in net benefit compared to the optimal treatment strategy without the biomarker. In other words, if one computes the improvements in the net benefit for the marker guided treatment selection strategy versus two potential strategies: treating everyone and treating no one, then the minimum of the two incremental values corresponds to the expected benefit of the marker-based strategy.

In practice, it is difficult to agree upon one single utility parameter c. Rather an expected benefit curve of $EB_A(c)$ versus c can be used. Examples of expected benefit curves are shown in Figures 1 (a)(b), which we describe in more detail in next section. Hereafter, to simplify notation we write $EB(c)$ with the understanding that a underlying strategy $A(Y)$ is implicitly involved.

2.1 Perfect Treatment Selection and Standardized Expected Benefit

In this section, we define the expected benefit for a perfect treatment selection rule which can be used to standardize the expected benefit of a marker-based rule. This type of standardization makes the measure of benefit invariant to the choice of disease cost. Thus, it puts the measure on a relative scale between 0 and 1 which is common across different settings where marker, treatment, disease, and study populations may differ. Standardization has been commonly performed with measures of a biomarker's capacity for risk prediction, (e.g., the standardized total gain (Bura and Gastwirth, 2001; Huang and Pepe, 2009; Gu and Pepe, 2009) and the relative utility curve (Baker *and others*, 2009)), but not yet for treatment selection.

We define the new concept of a perfect treatment selection rule using a potential outcomes framework. Let $D(0), D(1)$ denote the pair of potential outcomes if a subject receives placebo or treatment respectively. The four possible values of $D(0), D(1)$ are shown below with $q_1, \ldots, q_4$ denoting the unobserved population proportion of subjects falling into each category. Subjects with $D(0) = 1, D(1) = 0$ can be described as treatment-benefitted, subjects with $D(0) = D(1) = 1$ or $D(0) = D(1) = 0$ are treatment-unaffected, and subjects with $D(0) = 0, D(1) = 1$ are treatment-harmed, where benefits and harms are with respect to the targeted disease of interest.

$D(0)$	$D(1)$		*proportion*
1	0	benefitted	q_1
1	1	unaffected	q_2
0	0	unaffected	q_3
0	1	harmed	q_4

For a particular cost ratio $c \geq 0$, a perfect treatment selection rule will identify all treatment-benefitted subjects for treatment and recommend against treatment for others. This will lead to a population-averaged total disease and treatment cost of $q_2 + q_1 c = \rho_0 - q_1 + q_1 c$, which corresponds to expected benefit $PEB(c) = q_1(1-c) - [\rho_0 - \rho_1 - c]_+$.

While in general q_1 is not identifiable in the observed data, upper and lower bounds can be identified using a disease risk model. Let $q_k(Y), k = 1, 2, 3, 4$ indicate the probability that a subject with marker Y falls into the i^{th} potential outcome category, and let $\rho_0(Y) = P(D = 1|Y, T = 0), \rho_1(Y) = P(D = 1|Y, T = 1)$. We have

$$
\begin{aligned}
&q_1(Y) + q_2(Y) = \rho_0(Y) \Rightarrow q_1(Y) \leq \rho_0(Y), \\
&q_1(Y) + q_3(Y) = 1 - \rho_1 \Rightarrow q_1(Y) \leq 1 - \rho_1(Y), \\
&q_1(Y) - q_4(Y) = \rho_0(Y) - \rho_1(Y) \Rightarrow q_1(Y) \geq \rho_0(Y) - \rho_1(Y),
\end{aligned}
$$

which implies $max\{0, \rho_0(Y) - \rho_1(Y)\} \leq q_1(Y) \leq \min\{\rho_0(Y), 1 - \rho_1(Y)\}$. Taking an expectation over Y, we have $E\left[\{\Delta(Y)\}_+\right] \leq q_1 \leq \rho_0 - E\left[\{\rho_0(Y) + \rho_1(Y) - 1\}_+\right]$. Note that alternative nonparametric bounds for q_1 can be derived without relying on any biomarker or model information: $\max(0, \rho_0 - \rho_1) = [E\{\Delta(Y)\}]_+ \leq q_1 \leq \min(\rho_0, 1 - \rho_1) = \rho_0 - [E\{\rho_0(Y) + \rho_1(Y) - 1\}]_+$. Incorporating risk model information nevertheless leads to narrower bounds of q_1 since $[E\{\Delta(Y)\}]_+ \leq E\left[\{\Delta(Y)\}_+\right]$ and $[E\{\rho_0(Y) + \rho_1(Y) - 1\}]_+ \leq E\left[\{\rho_0(Y) + \rho_1(Y) - 1\}_+\right]$, and will be the focus of this paper. These types of restrictions on the probability of potential outcome category have also been recognized by others, e.g., Gadbury *and others* (2004); Huang *and others* (2012); Zhang *and others* (2013).

Based on the model of the risk of D conditional on Y and T, we can construct a lower bound for the expected benefit of a perfect treatment selection rule as

$$PEB^l(c) = E\left[\{\Delta(Y)\}_+\right] \times (1-c) - \left[\rho_0 - \rho_1 - c\right]_+, \tag{4}$$

and an upper bound

$$PEB^u(c) = (\rho_0 - E\left[\{\rho_0(Y) + \rho_1(Y) - 1\}_+\right]) \times (1-c) - \left[\rho_0 - \rho_1 - c\right]_+. \tag{5}$$

In summary, the uncertainty in identifying the expected benefit of a perfect treatment selection marker is caused by the non-identifiability of the percent of "benefitted" individuals, in other words, the inability to separate "benefitted" individuals from "unaffected" individuals among diseased subjects in the placebo arm. In the special case where the treatment has a monotone effect on the targeted disease and will not cause any harm (so $q_4 = 0$), we have $q_1 = \rho_0 - \rho_1$. Thus, under monotonicity the expected benefit of a perfect treatment selection rule can be uniquely identified as $(\rho_0 - \rho_1) \times (1-c) - \left[\rho_0 - \rho_1 - c\right]_+$, which is equal to its lower bound in (4) since $E\left[\{\Delta(Y)\}_+\right] = E\{\Delta(Y)\} = \rho_0 - \rho_1$ under monotonicity.

Finally, dividing the expected benefit of a marker-based treatment strategy by the bounds of expected benefit from perfect treatment selection, we obtain bounds for the standardized expected benefit: $SEB^l(c) = EB(c)/PEB^u(c)$ and $SEB^u(c) = EB(c)/PEB^l(c)$.

Expected benefit from a perfect treatment selection sets a reference for gauging the benefit of a particular treatment selection rule or the difference in benefit between treatment selection rules. In Figures 1(a)(b), we show the expected benefit of two treatment selection markers and lower and upper bounds for perfect treatment selection derived from corresponding marker-based risk model. Marker 1 (Figure 1(a)) has small expected benefit with a large potential for improvement. For example, at a cost ratio 0.05, its expected benefit of 0.005

is far from the perfect selection rule: a perfect selection rule can have an expected benefit 8.8-26.9 times that of Marker 1; corresponding standardized expected benefit for Marker 1 at cost ratio 0.05 ranges between 3.7% and 11.4% (Figure 1(c)). In contrast, there is less potential to improve over a better marker (Marker 2) (Figure 1(b)). At a cost ratio 0.05, a perfect selection rule can have expected benefit 1.7-2.8 times that of Marker 2, which has expected benefit 0.05 and standardized value ranging from 35.7% to 58.0% (Figure 1(d)).

3 Derivation of Treatment Selection Rules for Optimizing Expected Benefit

In this section, we consider methods for maximizing the expected benefit of a marker-based treatment selection rule and for estimating the benefit and its standardized value for varying c. We first need to construct a treatment selection rule $A(Y)$. In this paper we consider the class of selection rules $A(Y)$ which compare $h(Y)$, a function of Y with a threshold value δ to be determined, where $h(Y)$ is constructed by fitting a generalized linear model (GLM) of D on Y and T. Next we propose two strategies for optimizing the selection rule $A(Y)$ and calculating (standardized) expected benefit. Both methods use the same strategy to find $h(Y)$, but differ in the determination of δ. The first relies on a well-calibrated model for risk difference $\Delta(Y)$, the second does not and is thus more robust to model-misspecification. The first strategy however is more efficient under a correctly specified model.

3.1 Model-Based Approach

Based on equation (1), it can be seen that a marker-based rule $A(Y)$ that optimizes expected benefit at cost ratio c is equal to 1 whenever $\Delta(Y) > c$ and 0 otherwise. For details see, for example, Vickers *and others* (2007) and Janes *and others* (2013).

We consider modeling the risk of D conditional on Y and T with $g\{P(D=1|Y,T)\} = \beta_0 + \beta_1 T + \beta_2^T Y + \beta_3^T YT$, where g is a known link function, for example, the logit or inverse normal CDF. Let $\hat{\beta}$ denote the maximum likelihood estimator (MLE) of β, and let $\hat{\Delta}(Y)$ denote the corresponding estimator of $\Delta(Y)$. When the model for risk difference is well-calibrated, i.e., when $\hat{\Delta}(Y)$ is a good estimator of $\Delta(Y)$, a model based estimator of expected benefit can be constructed based on (1): $\widehat{EB}(c) = \sum_{i=1}^{N} \left(\hat{\Delta}_i - c\right)_+ /N - \left(\sum_{i=1}^{N} \widehat{\Delta}_i/N - c\right)_+$, where $\hat{\Delta}_i = \hat{\Delta}(Y_i)$ is the estimate of Δ for subject i. Note that a good calibration of the risk model itself is sufficient for the risk difference Δ to be well-calibrated, but not necessary. Hosmer-Lemeshow type techniques can be used for evaluating both types of calibrations (Hosmer and Lemesbow, 1980; Huang and Pepe, 2010; Janes *and others*, 2013).

We estimate the lower bound on the expected benefit of a perfect treatment selection rule as $\widehat{PEB}^{l}(c) = \sum_{i=1}^{N} \left(\hat{\Delta}_i\right)_+ \times (1-c)/N - \left(\sum_{i=1}^{N} \hat{\Delta}_i/N - c\right)_+$ and the upper bound as $\widehat{PEB}^{u}(c) = \sum_{i=1}^{N} \left\{\widehat{Risk}_{0i} - \left(\widehat{Risk}_{0i} + \widehat{Risk}_{1i} - 1\right)_+\right\} \times (1-c)/N - \left(\sum_{i=1}^{N} \hat{\Delta}_i/N - c\right)_+$, where $\widehat{Risk}_0$ and $\widehat{Risk}_1$ are model-based estimates of $P(D=1|Y, T=0)$ and $P(D=1|Y, T=1)$ respectively. Corresponding lower and upper bounds on $SEB(c)$ can be estimated as $\widehat{EB}(c)/\widehat{PEB}^{u}(c)$ and $\widehat{EB}(c)/\widehat{PEB}^{l}(c)$. Next we present asymptotic theory for the model-based estimators.

3.2 Asymptotic Theory for the Model-Based Estimator

Under standard regularity conditions listed in Supplementary Appendix, when $\rho_0 - \rho_1 \neq c$, $\widehat{EB}(c)$, $\widehat{PEB}^{l}(c)$, and $\widehat{PEB}^{u}(c)$ are asymptotically normal as stated in Theorems 1 and 2.

Theorem 1 Under the specified regularity conditions, $\widehat{EB}(c)$ is asymptotically normally distributed as $N \to \infty$ for $c \neq \rho_0 - \rho_1$. In particular, we have

$$\sqrt{N}\left\{\widehat{EB}(c) - EB(c)\right\} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{\Psi_{1i} + \Psi_{2i} - I(\rho_0 - \rho_1 > c)\left(\Psi_{3i} + \Psi_{4i}\right)\right\} + o_p(1),$$

where

$$\Psi_{1i} = \frac{\partial EB(c)}{\partial \beta} I^{-1}(\beta) l(\beta)_i, \qquad \Psi_{2i} = (\Delta_i - c)_+ - E\left\{(\Delta - c)_+\right\},$$

$$\Psi_{3i} = \frac{\partial E\{\Delta(\beta)\}}{\partial \beta} I^{-1}(\beta) l(\beta)_i, \qquad \Psi_{4i} = \Delta_i - (\rho_0 - \rho_1),$$

with $I(\beta)$ and $l(\beta)$ the information matrix and efficient influence function for $P(D|Y,T)$.

Theorem 2 Under the specified regularity conditions, $\widehat{PEB}^{l}(c)$ and $\widehat{PEB}^{u}(c)$ are asymptotically normally distributed as $N \rightarrow \infty$ for $c \neq \rho_0 - \rho_1$. In particular:

$$(i) \qquad \sqrt{N}\left\{\widehat{PEB}^{l}(c) - PEB^{l}(c)\right\} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{(1-c)(\Psi_{5i}+\Psi_{6i}) - I(\rho_0-\rho_1>c)(\Psi_{3i}+\Psi_{4i})\} + o_p(1),$$

$$(ii) \qquad \sqrt{N}\left\{\widehat{PEB}^{u}(c) - PEB^{u}(c)\right\} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{(1-c)(\Psi_{7i}+\Psi_{8i}) - I(\rho_0-\rho_1>c)(\Psi_{3i}+\Psi_{4i})\} + o_p(1),$$

where

$$\Psi_{5i} = \frac{\partial E(\Delta_+)}{\partial\beta}I^{-1}(\beta)l(\beta)_i, \qquad \Psi_{6i} = (\Delta_i)_+ - E(\Delta_+),$$

$$\Psi_{7i} = \frac{\partial\left[E\{Risk_0(\beta)\} - E\{Risk_0(\beta)+Risk_1(\beta)-1\}_+\right]}{\partial\beta}I^{-1}(\beta)l(\beta)_i,$$

$$\Psi_{8i} = Risk_{0i}(\beta) - \{Risk_{0i}(\beta)+Risk_{1i}(\beta)-1\}_+ - \left[\rho_0 - E\{Risk_0(\beta)+Risk_1(\beta)-1\}_+\right].$$

When $c \neq \rho_0 - \rho_1$, asymptotic normality of $\widehat{SEB}(c)^l$ and $\widehat{SEB}(c)^u$ then follows from Theorems 1 and 2 and the Delta method.

When $c = \rho_0 - \rho_1$, it can be shown that $\sqrt{N}\left\{\left(\sum_{i=1}^{N}\hat{\Delta}_i/N - c\right)_+ - (\rho_0-\rho_1-c)_+\right\}$ converges to a mixture of 0 and a truncated normal distribution. As a result, asymptotic normality of $\widehat{EB}(c)$, $\widehat{PEB}(c)$, or $\widehat{SEB}(c)$ does not hold. Even when asymptotic normality of these estimators does hold, we recommend the bootstrap for constructing confidence intervals since computation of the asymptotic variance of these estimators requires numerical differentiation. In practice, standard bootstrap percentile CI can lead to undercoverage when $c \approx \rho_0 - \rho_1$, we adopt an adaptive bootstrap confidence interval following the ideas of Berger and Boos (1994), Laber and Murphy (2011), and Robins (2004). Specifically, the proposed interval is equivalent to the standard bootstrap percentile CI when c is far from $\rho_0 - \rho_1$ and is equivalent to a projection interval otherwise, which is the union of bootstrap intervals as described below. Because the behavior of the confidence interval is automatically dictated by the data, we term it 'adaptive.'

Let $b = 1, \ldots, B$ index bootstrap samples drawn from the original data with replacement. We add a superscript b, to indicate that a statistic has been computed using a bootstrap sample. We construct an adaptive projection confidence interval as follows. For any $r \in \mathbb{R}$ define $\widehat{EB}^{b}_{r}(c) = \sum_{i=1}^{N} \left(\hat{\Delta}_i^b - c\right)_+ /N - \left(\sum_{i=1}^{N} \hat{\Delta}_i^b/N - c\right) I(r > 0)$, $\widehat{PEB}^{lb}_{r}(c) = \sum_{i=1}^{N} \left(\hat{\Delta}_i^b\right)_+ \times (1 - c) - \left(\sum_{i=1}^{N} \hat{\Delta}_i^b/N - c\right)_+ I(r > 0)$, and $\widehat{PEB}^{ub}_{r}(c) = \sum_{i=1}^{N} \left\{ \widehat{Risk}^{b}_{0i} - \left(\widehat{Risk}^{b}_{0i} + \widehat{Risk}^{b}_{1i} - 1\right)_+ \right\} /N \times (1 - c) - \left(\sum_{i=1}^{N} \hat{\Delta}_i^b/N - c\right) I(r > 0)$. Let $\zeta_{EB(c),\eta}(r)$, $\zeta_{PEB^l(c),\eta}(r)$, and $\zeta_{PEB^u(c),\eta}(r)$ denote $(1 - \eta) \times 100\%$ percentile bootstrap confidence intervals formed by taking empirical percentiles of $\widehat{EB}^{b}_{r}(c)$, $\widehat{PEB}^{lb}_{r}(c)$, and $\widehat{PEB}^{ub}_{r}(c)$ over bootstrap samples respectively. Let $\Gamma_\alpha(c)$ denote an asymptotically valid $(1 - \alpha) \times 100\%$ confidence interval for $\rho_0 - \rho_1 - c$. The $(1 - \eta - \alpha) \times 100\%$ projection intervals for $EB(c)$, $PEB^l(c)$ and $PEB^u(c)$ are given respectively by $\bigcup_{r \in \Gamma_\alpha(c)} \zeta_{EB(c),\eta}(r)$, $\bigcup_{r \in \Gamma_\alpha(c)} \zeta_{PEB^l(c),\eta}(r)$, and $\bigcup_{r \in \Gamma_\alpha(c)} \zeta_{PEB^u(c),\eta}(r)$. Let P^b denote probability taken with respect to the bootstrap sampling algorithm, conditional on the observed data. The following results (Theorem 3 and Corollary 1) are proved in the Supplementary Appendix B.

Theorem 3 [Projection bootstrap intervals] Assume $\Delta(Y)$ has a continuous and bounded density function. Let $\alpha, \eta \in (0, 1)$, and let c be fixed. Then,

1. $P^b \left(EB(c) \in \bigcup_{r \in \Gamma_\alpha(c)} \zeta_{EB(c),\eta}(r) \right) \geq 1 - \alpha - \eta + o_p(1)$;
2. $P^b \left(PEB^l(c) \in \bigcup_{r \in \Gamma_\alpha(c)} \zeta_{PEB^l(c),\eta}(r) \right) \geq 1 - \alpha - \eta + o_p(1)$;
3. $P^b \left(PEB^u(c) \in \bigcup_{r \in \Gamma_\alpha(c)} \zeta_{PEB^u(c),\eta}(r) \right) \geq 1 - \alpha - \eta + o_p(1)$.

If $E\Delta(Y) \neq c$ then the right hand side of the foregoing inequalities can be replaced with equality to $1 - \eta + o_P(1)$.

Corollary 1 Let τ_N be a sequence of positive random variables satisfying $\tau_N \to 0$ and $\sqrt{N}\tau_N \to \infty$ almost surely as $N \to \infty$. Define $\mathfrak{A}(c) = \Gamma_\alpha(c)$ if $|\hat{\rho}_0 - \hat{\rho}_1 - c| \le \tau_N$ and $\{\sum_{i=1}^N \hat{\Delta}_i/N - c\}$ otherwise. Assume the conditions of Theorem 3. Then,

1. $P^b\left(EB(c) \in \bigcup_{r\in\mathfrak{A}(c)} \zeta_{EB(c),\eta}(r)\right) \ge 1 - \alpha - \eta + o_p(1)$;
2. $P^b\left(PEB^l(c) \in \bigcup_{r\in\mathfrak{A}(c)} \zeta_{PEB^l(c),\eta}(r)\right) \ge 1 - \alpha - \eta + o_p(1)$;
3. $P^b\left(PEB^u(c) \in \bigcup_{r\in\mathfrak{A}(c)} \zeta_{PEB^u(c),\eta}(r)\right) \ge 1 - \alpha - \eta + o_p(1)$.

If $E\Delta(Y) \neq c$ then the right hand side of the foregoing inequalities can be replaced with equalities. Note that for $|\hat{\rho}_0 - \hat{\rho}_1 - c| > \tau_N$, $\bigcup_{r\in\mathfrak{A}(c)} \zeta_{EB(c),\eta}(r)$, $\bigcup_{r\in\mathfrak{A}(c)} \zeta_{PEB^l(c),\eta}(r)$, and $\bigcup_{r\in\mathfrak{A}(c)} \zeta_{PEB^u(c),\eta}(r)$ in the corollary refer to standard $(1-\eta)\times 100\%$ bootstrap confidence interval for $EB(c)$, $PEB^l(c)$ and $PEB^u(c)$.

Remark 1. Berger and Boos (1994) recommend choosing α to quite small in which case $1 - \eta \approx 1 - \eta - \alpha$. Consequently, the originally proposed projection confidence interval is nearly exact in large samples provided $E\Delta(Y) \neq c$, but potentially conservative otherwise. However, Corollary 1 suggests a procedure which provides exact coverage when $E\Delta(Y) \neq c$ and is thus both adaptive and less conservative than the projection interval. For these reasons, it is recommended in practice.

Remark 2. The conditions of the preceding theorem can be relaxed at the expense of a possibly more conservative confidence interval. In the Supplementary Appendix C we provide a locally consistent projection confidence interval that does not require $\Delta(Y)$ have smooth bounded density. However, this interval requires taking a union over a larger set and is thus potentially more conservative in some settings. We defer the detailed investigation of this CI to future work.

3.3 Robust Estimation Methods

The validity of the model-based approach for estimating the expected benefit of an optimal treatment selection rule depends critically on model calibration. In practice, one may adopt a working model $\Delta^\star$ based on a GLM and focus on constructing a decision rule $A(Y) = I\{\Delta^\star(Y) \geq \delta\}$ that has large benefit regardless of whether or not $\Delta^\star$ is well-calibrated.

Write the expected benefit based on $\Delta^\star$ and threshold δ as $E\{I(\Delta^\star(Y) > \delta)\{\Delta(Y) - c\}\} - [\rho_0 - \rho_1 - c]_+$, which in a randomized trial can be represented as

$$\begin{aligned} &\{P(D=1|\Delta^\star(Y) > \delta, T=0) - P(D=1|\Delta^\star(Y) > \delta, T=1) - c\} \, P(\Delta^\star(Y) > \delta) \\ - \;\; & [\rho_0 - \rho_1 - c]_+ \, , \end{aligned} \tag{6}$$

and estimated nonparametrically by

$$\begin{aligned} &\left\{ \frac{\sum_{i=1}^N D_i \left(\hat{\Delta}_i^\star > \delta\right)(1-T_i)}{\sum_{i=1}^N \left(\hat{\Delta}_i^\star > \delta\right)(1-T_i)} - \frac{\sum_{i=1}^N D_i \left(\hat{\Delta}_i^\star > \delta\right) T_i}{\sum_{i=1}^N \left(\hat{\Delta}_i^\star > \delta\right) T_i} - c \right\} \frac{1}{N} \sum_{i=1}^N \left(\hat{\Delta}_i^\star > \delta\right) \\ - \;\; & \left[\frac{\sum_{i=1}^N Y_i(1-T_i)}{\sum_{i=1}^N (1-T_i)} - \frac{\sum_{i=1}^N Y_i(T_i)}{\sum_{i=1}^N T_i} - c \right]_+ . \end{aligned} \tag{7}$$

A natural nonparametric analogue of the model-based estimator based on threshold value c can be constructed by entering $\delta = c$ into (7). This estimator is unbiased for the expected benefit of $A(Y) = \Delta^\star(Y) > c$ whether or not $\Delta^\star$ is well calibrated. However, when $\Delta^\star$ is not well-calibrated, the rule based on comparing $\Delta^\star$ with c is not optimal among all rules of the form $\Delta^\star > \delta$. An optimal δ among this class that maximizes (6) can instead be identified by maximizing (7) over δ. Furthermore, following Zhang *and others* (2012), an augmented version of (7) can be constructed as

$$\begin{aligned} &(7) + \frac{1}{N} \sum_{i=1}^N \frac{T_i \left(\hat{\Delta}_i^\star > \delta\right) + (1-T_i)\left(\hat{\Delta}_i^\star \leq \delta\right) - \pi(Y,\delta)}{\pi(Y,\delta)} \times \\ &\left\{ \widehat{Risk}_1 \times I(\hat{\Delta}_i^\star > \delta) + \widehat{Risk}_0 \times I(\hat{\Delta}_i^\star \leq \delta) \right\} , \end{aligned} \tag{8}$$

where $\pi(Y,\delta) = P(T=1)(\hat{\Delta}^\star > \delta) + P(T=0)(\hat{\Delta}^\star \leq \delta)$. The optimal δ can be constructed as the maximizer of (8) for potential efficiency gain. These robust estimation methods are aimed for scenarios where risk model is prone to misspecification. Since validity of the model-based bounds in (4) and (5) for perfect selection benefit relies on well-calibrated risk model, here we do not consider those bounds anymore.

Let $\hat{\delta}$ be the estimate of δ through maximization of either (7) or (8), the corresponding value of expected benefit in the training data tends to overestimate the true expected benefit of the rule $A(Y) = \hat{\Delta}^\star > \hat{\delta}$. In practice one can use cross-validation to correct for this bias, as we demonstrate in our simulation studies.

4 Simulation Studies

In this section, we conduct simulation studies to investigate our estimators of (standardized) expected benefit. We consider a two-arm 1:1 randomized trial with $T = 0$ and $T = 1$ indicating placebo and treatment arm respectively. Assume we have a biomarker Y which follows standard normal distribution, we consider a linear logistic model for the risk of a binary disease D conditional on Y and T: $\text{logit}P(D = 1|Y,T) = \beta_0 + \beta_1 T + \beta_2 Y + \beta_3 YT$. The risk model parameters are chosen such that disease prevalences are $\rho_0 = 0.25$ and $\rho_1 = 0.125$ in placebo and treatment arm respectively. We consider cost ratios $c = 0, 0.105, 0.125, 0.145, 0.175$, which correspond to expected benefit value of 0.043, 0.059, 0.063, 0.048, 0.029. The pairs of lower and upper bounds for expected benefit from perfect treatment selection equal to $\{0.043, 0.098\}$, $\{0.130, 0.180\}$, $\{0.147, 0.196\}$, $\{0.144, 0.191\}$, and $\{0.139, 0.184\}$ respectively. Sample sizes of 200, 500, and 2000 are used in the simulation studies.

Performance of the model-based estimators for $EB(c)$, $PEB(c)$ and $SEB(c)$ are shown in Tables 1 and 2. With a sample size of 200, model-based estimators have minimal bias for each measure. Coverage of 95% percentile bootstrap CI is close to nominal level when

c is away from $\rho_0 - \rho_1$, whereas an under-coverage is observed when $c = \rho_0 - \rho_1$, which is not alleviated with the increase of sample size. The adaptive bootstrap CI fixes the under-coverage problem where we use the projection interval (with $\alpha = 0.01$) when estimated $\rho_0 - \rho_1$ is close to c (defined as $|\hat{\rho}_0 - \hat{\rho}_1 - c| \leq \hat{SE}(\hat{\rho}_0 - \hat{\rho}_1) \times \Phi^{-1}(0.9)$ in the simulation study).

If we use the same threshold c for risk difference but use a nonparametric method instead for estimating $EB(c)$, a decrease of efficiency is observed compared to the model-based estimator (details omitted). In our simulation setting, variances of the model-based estimators are around 40% $\sim$ 70% of the variances of the nonparametric estimators.

Based on the same logistic model fitting, we further consider finding $\hat{\delta}$ by maximizing the empirical estimate of $EB(c)$ (7) and its augmented version (8). Table 3 presents performance of the derived treatment decision rule using $\hat{\delta}$ versus using $\delta = c$ in the population. Estimating δ leads to a rule with smaller expected benefit with larger variability for small sample size of 200, but has small impact on treatment selection performance when sample size is as large as 2000. In general, using the augmented estimator for δ estimation leads to small increase in treatment selection performance and decreased variability. Note that under correct model specification, the estimated expected benefit based on robust methods is expected to be suboptimal compared to the model-based estimator. It is when the model is misspecified that the robust estimation methods may yield higher expected benefit.

In Supplementary Table 1, we present results for estimating expected benefit of treatment selection rule $\hat{\Delta}^{\star} > \hat{\delta}$, using naive estimators based on the training data or alternative estimators based on random cross-validation. For the latter, we randomly split the data into 2/3 of training set and 1/3 of test set, fitting a logistic model to the training data and compute $\hat{\delta}$, then compute the expected benefit based on the test data. An average of expected benefit is computed over 500 splits. We see that expected benefit estimated from training data can have severe over-estimation even with a sample size as large as 2000, which is corrected by cross-validation. Also presented in Supplementary Table 1 are naive and CV estimates for expected benefits using the model-based rule. Here the over-fitting in naive estimate is much less severe compared to nonparametric estimator and is minimal when sample size goes above 500.

5 Data Example

In this section, we illustrate the estimation of expected benefit using an example from the Diabetes Control and Complications Trial (DCCT) (Control and Group, 1993). DCCT was a large-scale randomized controlled trial designed to compare intensive and conventional diabetes therapy with respect to their effects on the development and progression of the early vascular and neurologic complications of diabetes. Overall 1441 patients with insulin-dependent diabetes mellitus were enrolled from 1983 to 1999, including 726 primary prevention cohort patients who were free of any microvascular complications and 715 secondary prevention cohort patients who had mild preexisting retinopathy or other complications. Their appearances of progression of retinopathy and other complications assessed regularly and the trial was terminated on 1999 with significant evidence of treatment efficacy resulting in an average followup of 6.5 years.

One outcome of which the treatment in DCCT shows significant effect in reducing the risk is micro-albuminuria, a sign of kidney damage, defined as albumin excretion rate greater than 40mg/24hr. Our analysis here consists of 579 subjects in the secondary prevention cohort who did not have micro-albuminuria and neuropathy at baseline. We consider baseline homoglobin A1C (HBA1C) as a biomarker for selecting treatment: a linear logistic regression model of micro-albuminuria developed during the study versus treatment and baseline HBA1C and their interaction shows a significant interaction between treatment and HBA1C.

We estimate the expected benefit of HBA1C and its standard value. The curve of model-based estimator of $EB(c)$ versus c is presented in Supplementary Figure 1(a). Also displayed are estimated lower and upper bounds of expected benefit for perfect treatment selection. Corresponding bounds for standardized expected benefit of HBA1C are displayed in Supplementary Figure 1(b). For a series of chosen cost ratio, the model-based estimates and their 95% CI are shown in Table 4. For example, at cost ratio $c = 0$, i.e., no cost of more intensive diabetes therapy, HBA1C has a EB of 0.005 while the EB of a perfect treatment selection rule can range from 0.005 to 0.206, such that standard EB of HBA1C is above 2.3%. For a bigger cost ratio such as $c = 0.05$, i.e., the cost of intensive therapy is 5% the cost of micro-albuminuria, HB1AC has EB of 0.019, which explains 8% - 38.8% benefits of a perfect treatment selection rule. Supplementary Table 2 presents cross-validated EB for the model-based estimator and for the robust estimators where the selection threshold is non-parametrically derived. In general we see reduction in EB resulted from CV. Optimization of the threshold leads to slightly better CV estimate compared to model-based estimator.

6 Concluding Remarks

In this paper we developed an expected benefit measure for characterizing the capacity of biomarkers in treatment selection. Built upon a decision-theoretic framework, this measure integrates the benefit of a marker-based treatment selection rule on reducing population disease rate with the additional treatment cost through the specification of a treatment/disease cost ratio. We also developed a new concept of a perfect treatment selection rule in the sense that it correctly makes treatment recommendation for patients according to whether or not they will benefit from the treatment. We developed bounds for expected benefit of a perfect treatment selection rule based on the model of disease risk conditional on marker and treatment. These bounds can be used to standardize the expected benefit of a treatment selection rule, potentially facilitating comparison of markers across different study settings.

The idea of generating bounds for treatment selection can be readily applied to other treatment selection measures such as the population disease rate resulted from treatment selection (Song and Pepe, 2004). An interesting fact about these model-based bounds is that their width depends on how well the risk model used to construct the bounds can identify the percent of "benefitted" in the population. A model that better predicts heterogeneity in treatment responses in terms of larger variability in Δ tends to move up the lower bound for PEB through the increase of $E\{\Delta(Y)\}_+$. In other words, tighter bounds reflect a better knowledge in selecting treatment-benefitted subjects. In particular, in the scenario where we compare two nested models, sharper bounds for perfect treatment selection can be estimated from the more complicated model. In general when we have several risk models in a population to assess expected benefit of perfect treatment selection. Tighter bounds can be constructed using bounds derived from individual risk model. Specifically, at a given cost ratio, the minimum benefit of perfect selection can be constructed as the maximum among corresponding values in individual lower bounds, and the maximum benefit of perfect selection can be constructed as the minimum among corresponding values in individual upper bounds.

We considered the problem of maximizing the expected benefit based on a GLM and proposed two strategies. The model-based estimator was more efficient under well-calibrated models whereas the nonparametric and augmented estimators were robust to misspecification of working model and provided a way for sensitivity analysis. One advantage of using the common GLM method for deriving the treatment selection rule is the computation simplicity as the GLM model can be easily implemented with standard statistical software. We note that there are alternative ways to construct the treatment selection rule $A(Y)$ in the field, such as maximizing the measure of interest directly as adopted in Zhang *and others* (2012ab, 2013), and Zhao *and others* (2012); expected benefit of the treatment selection rule derived can be similarly estimated, through procedures such as cross-validation. To make inference using the proposed model-based estimator, we proposed an adaptive bootstrap procedure to handle the presence of non-regularity when cost ratio is close to the average treatment effect. This idea of using data to adaptively construct bootstrap confidence interval has a great potential to be used in other types of biomarker evaluation and comparison problems where non-regularity can occur at some point in the parameter space.

In this paper, we consider cost ratio c to be a constant and a series of c can be chosen for sensitivity analysis. In practice, the cost ratio might be a function of biomarker. For example, the cost of mammography use for breast cancer prevention might depend on women's age (Gail, 2009). It is straightforward to extend the concept of expected benefit allowing $c = C(Y)$ to be a function of biomarker value in scenarios where information is available for modeling $C(Y)$ as proposed in Janes *and others* (2013).

Finally, while the concepts of perfect and/or standardized expected benefits are restricted to binary disease outcomes, the concept of expect benefit itself and the methods developed can be readily generalized to handle continuous outcomes.

Acknowledgment

This work was supported by NIH grant R01 GM106177-01, R01 CA152089, P30CA015704, U01 CA086368, and PO1 CA053996. *Conflict of Interest*: None declared.

References

Albain, K.S., Barlow, W.E., Shak, S., Hortobagyi, G.N., Livingston, R.B., Yeh, I., Ravdin, P., Bugarini, R., Baehner, F.L., Davidson, N.E. *and others*. (2010). Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on

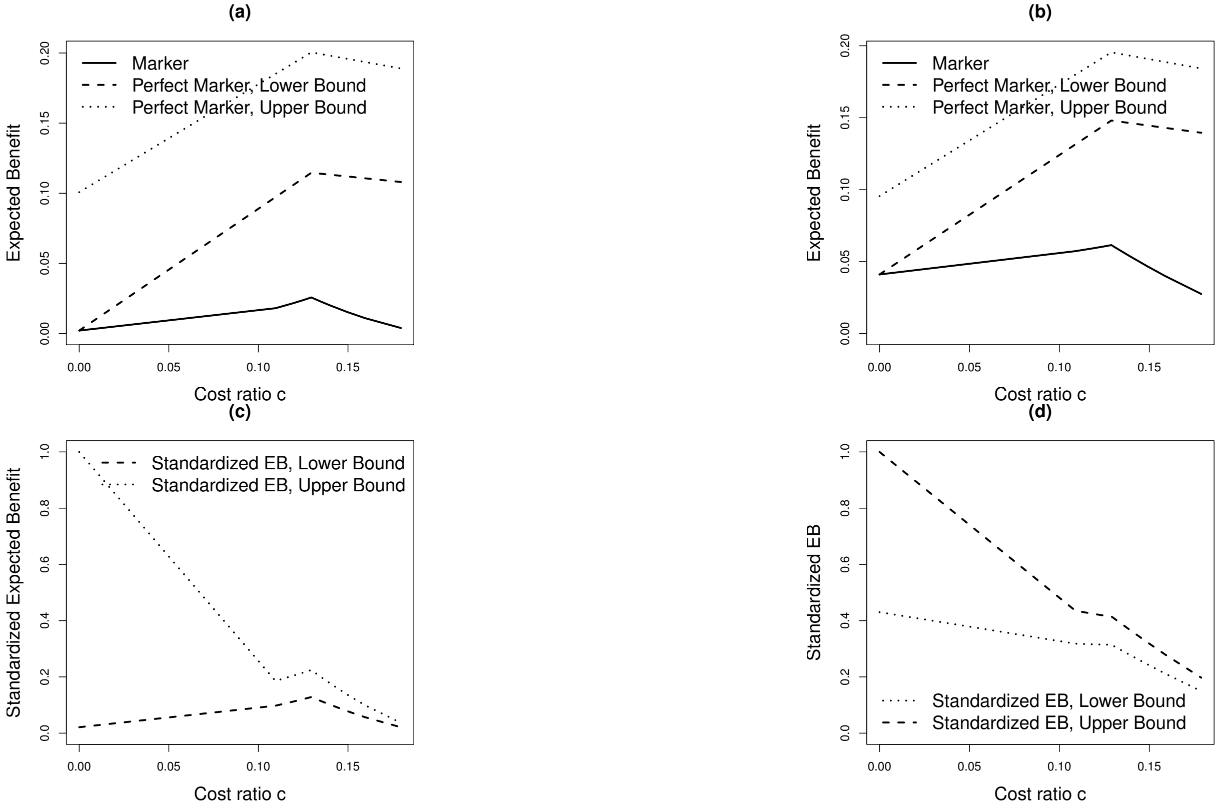

Figure 1: Expected benefit curves for Marker 1 (a) and Marker 2 (b) and bounds for perfect treatment selection rule, and corresponding standardized expected benefit curves for Marker 1 (c) and Marker 2 (d). Disease prevalence among untreated and treated subjects is 0.25 and 0.12 respectively. Each marker Y follows a N(2,1) distribution and its relationship to disease status is described by a logistic risk model, $\text{logit}P(D = 1|Y, T) = \beta_0 + \beta_1 T + \beta_2 Y + \beta_3 YT$, where T is an indicator of treatment assignment. For Marker 1, $\beta_0 = 0.69, \beta_1 = 0.2, \beta_2 = -1, \beta_3 = -1$. For Marker 2, $\beta_0 = -0.158, \beta_1 = 3.495, \beta_2 = -0.5, \beta_3 = -4$. Note that in this setting the optimal treatment strategy in the absence of marker information is to treat everyone if $c < 0.13$ and treat no one otherwise.

chemotherapy: a retrospective analysis of a randomised trial. *The lancet oncology* **11**(1), 55–65.

Allegra, C.J., Jessup, J.M., Somerfield, M.R., Hamilton, S.R., Hammond, E.H., Hayes, D.F., McAllister, P.K., Morton, R.F. and Schilsky, R.L. (2009). American society of clinical oncology provisional clinical opinion: testing for kras gene

Table 1: Performance of the Model-Based Estimator for EB with $\rho_0 - \rho_1 = 0.125$.

Cost ratio c	0.000	0.105	0.125	0.145	0.175
$EB(c)$	0.043	0.059	0.063	0.048	0.029
N			Bias×1000		
200	2.02	-7.90	-15.55	-6.33	2.17
500	1.22	-3.26	-10.72	-3.00	1.50
2000	0.24	-0.33	-5.62	-0.15	0.66
			$SE \times \sqrt{N}$		
200	0.29	0.27	0.27	0.29	0.29
500	0.29	0.28	0.29	0.33	0.35
2000	0.29	0.31	0.31	0.40	0.38
	Coverage of 95% percentile bootstrap CI				
200	95.10	91.80	83.80	95.20	96.90
500	95.10	94.90	84.60	96.30	95.80
2000	94.50	96.40	85.10	96.80	95.50
	Coverage of 95% adaptive bootstrap CI				
200	95.14	96.06	93.22	96.08	97.02
500	95.08	96.72	93.74	96.58	95.96
2000	94.48	96.40	96.90	96.68	95.54

mutations in patients with metastatic colorectal carcinoma to predict response to anti–epidermal growth factor receptor monoclonal antibody therapy. *Journal of Clinical Oncology* **27**(12), 2091–2096.

Baker, S.G., Cook, N.R., Vickers, A. and Kramer, B.S. (2009). Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**(4), 729–748. PMCID:PMC2804257.

Berger, Roger L and Boos, Dennis D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89**(427), 1012–1016.

Bura, E. and Gastwirth, J. L. (2001). The binary regression quantile plot: assessing the importance of predictors in binary regression visually. *Biometrical Journal* **43**(1), 5–21.

Cai, T., Tian, L., Wong, P.H. and Wei, LJ. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* **12**(2), 270–282.

Control, The Diabetes and Group, Complications Trial Research. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *New England Journal of Medicine* **329**(14), 977–986.

Foster, J.C., Taylor, J.M.G. and Ruberg, S.J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30**(24), 2867–2880.

Gadbury, Gary L, Iyer, Hari K and Albert, Jeffrey M. (2004). Individual treatment effects in randomized trials with binary outcomes. *Journal of statistical planning and inference* **121**(2), 163–174.

Gail, M.H. (2009). Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. *Journal of the National Cancer Institute* **101**(13), 959–963. PMCID:PMC2704229.

Gu, W. and Pepe, M.S. (2009). Measures to summarize and comparie the predictive capacity of markers. *International Journal of Biostatistics* **5**(1), 1557–4679.

Harris, L., Fritsche, H., Mennel, R., Norton, L., Ravdin, P., Taube, S., Somerfield, M.R., Hayes, D.F. and Bast Jr, R.C. (2007). American society of clinical oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *Journal of Clinical Oncology* **25**(33), 5287–5312.

Hosmer, David W and Lemesbow, Stanley. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods* **9**(10), 1043–1069.

Huang, Ying, Ballinger, Dennis G, Dai, James Y, Peters, Ulrike, Hinds, David A, Cox, David R, Beilharz, Erica, Chlebowski, Rowan T, Rossouw, Jacques E, McTiernan, Anne *and others*. (2011). Genetic variants in the mrps 30 region and postmenopausal breast cancer risk. *Genome Medicine* **3**(6), 42–42.

Huang, Ying, Gilbert, Peter B and Janes, Holly. (2012). Assessing treatment-selection markers using a potential outcomes framework. *Biometrics* **68**(3), 687–696.

Huang, Y. and Pepe, MS. (2009). A parametric roc model-based approach for evaluating the predictiveness of continuous markers in case–control studies. *Biometrics* **65**(4), 1133–1144. PMCID:PMC2794984.

Huang, Y. and Pepe, M.S. (2010). Assessing risk prediction models in case–control studies using semiparametric and nonparametric methods. *Statistics in medicine* **29**(13), 1391–1410. PMCID:PMC3045657.

Janes, H., Brown, M. D., Pepe, M.S. and Huang, Y. (2013*a*). Statistical methods for evaluating and comparing biomarkers for patient treatment selection. *International Journal of Biostatistics (submitted)*.

Janes, H., Pepe, M.S., Bossuyt, P.M. and Barlow, W.E. (2011). Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine* **154**(4), 253.

Janes, H., Pepe, M.S. and Huang, Y. (2013*b*). A general framework for evaluating markers used to select patient treatment. *Medical Decision Making*.

Karapetis, C.S., Khambata-Ford, S., Jonker, D.J., O'Callaghan, C.J., Tu, D., Tebbutt, N.C., Simes, R.J., Chalchal, H., Shapiro, J.D., Robitaille, S. *and* *others*. (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine* **359**(17), 1757–1765.

Laber, Eric B and Murphy, Susan A. (2011). Adaptive confidence intervals for the test error in classification. *Journal of the American Statistical Association* **106**(495), 904–913.

Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., Cronin, M., Baehner, F.L., Watson, D., Bryant, J. *and others*. (2006). Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor–positive breast cancer. *Journal of Clinical Oncology* **24**(23), 3726–3734.

Prentice, R.L., Huang, Y., Hinds, D.A., Peters, U., Pettinger, M., Cox, D.R., Beilharz, E., Chlebowski, R.T., Rossouw, J.E., Caan, B. *and others*. (2009). Variation in the fgfr2 gene and the effects of postmenopausal hormone therapy on invasive breast cancer. *Cancer Epidemiology Biomarkers & Prevention* **18**(11), 3079–3085.

Rapsomaniki, E., White, I.R., Wood, A.M. and Thompson, S.G. (2012). A framework for quantifying net benefits of alternative prognostic models. *Statistics in Medicine* **31**(2), 114–130.

Robins, James M. (2004). Optimal structural nested models for optimal sequential decisions. In: *Proceedings of the second Seattle Symposium in Biostatistics*. Springer. pp. 189–326.

Song, X. and Pepe, M.S. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics* **60**(4), 874–883.

Vickers, A.J., Kattan, M.W. and Sargent, D.J. (2007). Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials* **8**(1), 14.

Zhang, Baqun, Tsiatis, Anastasios A, Laber, Eric B and Davidian, Marie. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68**(4), 1010–1018.

Zhang, Zhiwei, Wang, Chenguang, Nie, Lei and Soon, Guoxing. (2013). Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Zhao, Lihui, Tian, Lu, Cai, Tianxi, Claggett, Brian and Wei, Lee-Jen. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association* (just-accepted).

Zhao, Yingqi, Zeng, Donglin, Rush, A John and Kosorok, Michael R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**(499), 1106–1118.

Table 2: Performance of the Model-Based Estimator for Bounds of $PEB(c)$ and $SEB(c)$.

Cost ratio c	0.000	0.105	0.125	0.145	0.175	0.000	0.105	0.125	0.145	0.175
	$PEB^u(c)$					$PEB^l(c)$				
	0.098	0.180	0.196	0.191	0.184	0.043	0.130	0.147	0.144	0.139
N	Bias×1000									
200	-0.10	-11.73	-20.31	-12.08	-4.93	2.02	-9.82	-18.45	-10.27	-3.18
500	-0.53	-5.53	-13.37	-6.04	-2.00	1.22	-3.97	-11.84	-4.55	-0.56
2000	-0.28	-0.98	-6.38	-1.02	-0.32	0.24	-0.51	-5.92	-0.57	0.11
	$SE \times \sqrt{N}$									
200	0.33	0.29	0.31	0.34	0.39	0.29	0.25	0.28	0.32	0.38
500	0.33	0.30	0.32	0.37	0.44	0.29	0.26	0.29	0.36	0.43
2000	0.34	0.32	0.32	0.43	0.45	0.29	0.28	0.30	0.42	0.45
	Coverage of 95% percentile bootstrap CI									
200	94.90	88.20	77.80	90.70	95.90	95.10	89.60	78.70	92.70	96.70
500	94.50	91.50	77.70	93.80	95.50	95.10	94.20	80.70	96.10	95.50
2000	94.50	95.80	80.10	96.50	95.00	94.50	96.40	82.20	96.70	95.20
	Coverage of 95% adaptive bootstrap CI									
200	94.90	94.28	91.08	93.98	95.84	95.14	95.04	91.22	94.90	96.62
500	94.52	95.02	91.44	94.72	95.50	95.08	96.26	92.02	96.36	95.62
2000	94.50	96.08	96.38	96.54	95.04	94.48	96.44	96.32	96.56	95.24
	$SEB^l(c)$					$SEB^u(c)$				
	0.436	0.327	0.323	0.253	0.156	1.000	0.452	0.429	0.336	0.207
N	Bias×1000									
200	9.12	-30.47	-58.66	-27.63	4.30	0.00	-42.79	-74.27	-39.34	-0.86
500	10.19	-10.73	-38.06	-12.60	3.42	0.00	-17.01	-47.53	-19.03	0.24
2000	2.66	-0.74	-19.51	-0.96	2.05	0.00	-2.30	-23.24	-2.43	1.36
	$SE \times \sqrt{N}$									
200	2.02	1.23	1.24	1.31	1.36	0.00	1.41	1.46	1.57	1.65
500	2.12	1.25	1.27	1.45	1.60	0.00	1.38	1.44	1.71	1.94
2000	2.21	1.35	1.29	1.66	1.70	0.00	1.45	1.39	1.89	2.06
	Coverage of 95% percentile bootstrap CI									
200	95.10	95.80	91.40	96.80	97.30	100.00	94.10	88.80	96.50	97.40
500	95.10	96.70	90.80	96.90	95.60	100.00	95.50	87.80	96.60	95.80
2000	94.20	96.20	89.40	96.80	95.60	100.00	96.40	87.20	97.00	95.40
	Coverage of 95% adaptive bootstrap CI									
200	95.12	97.82	95.88	97.24	97.20	100.00	97.02	94.86	97.02	97.40
500	95.10	97.94	95.72	97.28	95.74	100.00	97.04	94.80	96.88	95.86
2000	94.16	96.18	97.82	96.68	95.58	100.00	96.44	97.28	96.84	95.42

Table 3: MEAN (SD) of expected benefit in the population using model-based selection rule(PAR) or selection rules where threshold value is nonparametrically identified (NPAR, AUG). NPAR is associated with empirically estimator and AUG is associated with augmented estimator.

	Cost ratio c	0.000	0.105	0.125	0.145	0.175
N=200	PAR	0.0406 (0.0035)	0.0542 (0.0067)	0.0577 (0.0072)	0.0420 (0.0080)	0.0213 (0.0075)
	NPAR	0.0337 (0.011)	0.0474 (0.0136)	0.0514 (0.0136)	0.0368 (0.0127)	0.0179 (0.0106)
	AUG	0.0353 (0.0097)	0.0484 (0.0128)	0.0523 (0.0126)	0.0376 (0.0117)	0.0187 (0.0092)
N=500	PAR	0.042 (0.0012)	0.0572 (0.002)	0.0611 (0.0022)	0.0459 (0.0025)	0.0257 (0.0035)
	NPAR	0.0385 (0.006)	0.0536 (0.0066)	0.0577 (0.0067)	0.0426(0.0071)	0.0227 (0.007)
	AUG	0.0395 (0.0044)	0.0539 (0.0061)	0.0579 (0.0065)	0.0427 (0.0071)	0.0227 (0.0069)
N=2000	PAR	0.0427 (3e-04)	0.0585 (5e-04)	0.0626 (6e-04)	0.0476 (7e-04)	0.028 (8e-04)
	NPAR	0.0415 (0.002)	0.0571 (0.0022)	0.0611 (0.0024)	0.0463 (0.0024)	0.0266 (0.0032)
	AUG	0.0417 (0.0016)	0.0572 (0.0022)	0.0612 (0.0023)	0.0463 (0.0024)	0.0264 (0.0035)

Table 4: Estimate and 95%CI of expected benefit in DCCT example.

Cost ratio c	0	0.05	0.10	0.12
$EB(c)$	0.005 (0, 0.166)	0.019 (0, 0.123)	0.035 (0.001, 0.102)	0.028 (0, 0.119)
$PEB^l(c)$	0.005 (0 ,0.166)	0.05 (0.031 ,0.157)	0.086 (0.029 ,0.149)	0.084 (0.029 ,0.149)
$PEB^u(c)$	0.206 (0.157 ,0.352)	0.242 (0.192 ,0.335)	0.267 (0.216 ,0.329)	0.261 (0.211 ,0.343)
$SEB^l(c)$	0.023 (0 ,0.498)	0.08 (0 ,0.382)	0.131 (0.003 ,0.334)	0.107 (0.001 ,0.366)
$SEB^u(c)$	1 (1 ,1)	0.388 (0 ,0.802)	0.408 (0.026 ,0.809)	0.333 (0.005 ,0.823)

Supplementary Material

Supplementary Figures

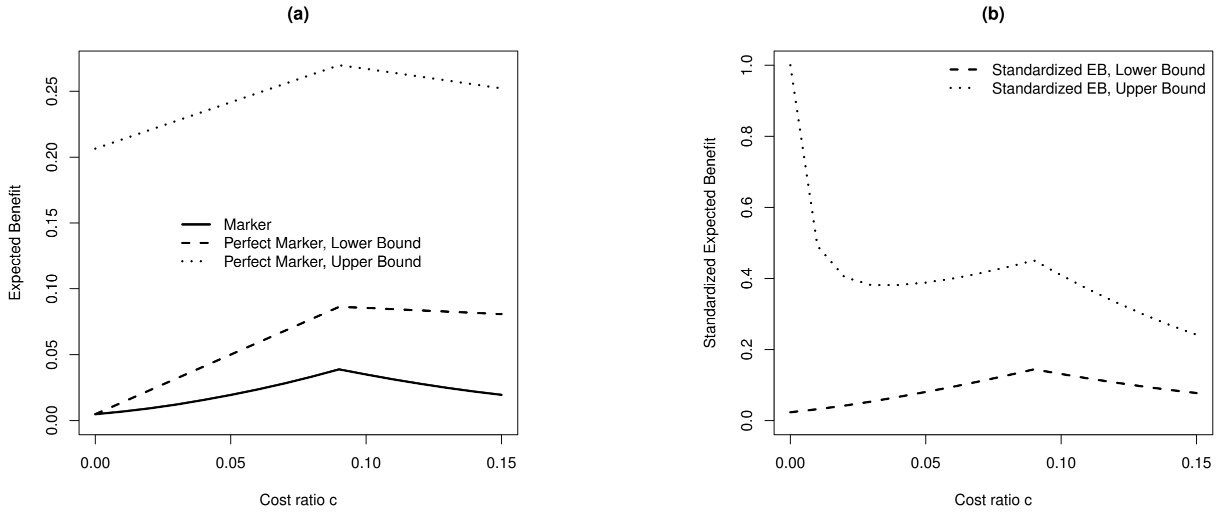

Figure 1: Expected benefit curves of HBA1C and the bounds for perfect biomarker for guiding the prevention of micro-albuminuria in the DCCT example.

Supplementary Tables

Table 1: Performance of naive and cross-validated estimates of expected benefit. PAR indicates model-based selection rule, NPAR and AUG are selection rules where threshold value is nonparametrically identified. NPAR is associated with empirically estimator and AUG is associated with augmented estimator.

	Cost ratio c		0.000			0.105			0.125	
		PAR	NPAR	AUG	PAR	NPAR	AUG	PAR	NPAR	AUG
$N = 200$	$EB(c;\hat{\beta})^{\star}$	0.0406	0.0337	0.0353	0.0542	0.0474	0.0484	0.0577	0.0514	0.0523
	Naive*	0.0454	0.0586	0.0568	0.0518	0.0653	0.0659	0.0485	0.0625	0.0636
	CV†	0.0391	0.0403	0.0394	0.0380	0.0393	0.0384	0.0327	0.0342	0.0336
$N = 500$	$EB(c;\hat{\beta})$	0.0420	0.0385	0.0395	0.0572	0.0536	0.0539	0.0611	0.0577	0.0579
	Naive	0.0439	0.0513	0.0506	0.0559	0.0633	0.0637	0.0531	0.0608	0.0616
	CV	0.0410	0.0418	0.0414	0.0501	0.0509	0.0502	0.0463	0.0471	0.0464
$N = 2000$	$EB(c;\hat{\beta})$	0.0427	0.0415	0.0417	0.0585	0.0571	0.0572	0.0626	0.0611	0.0612
	Naive	0.0433	0.0468	0.0465	0.0588	0.0622	0.0623	0.0575	0.0607	0.0609
	CV	0.0428	0.0432	0.0432	0.0575	0.0579	0.0575	0.0555	0.0558	0.0555
	Cost ratio c			0.145			0.175			
		PAR	NPAR	AUG	PAR	NPAR	AUG			
$N = 200$	$EB(c;\hat{\beta})$	0.0420	0.0368	0.0376	0.0213	0.0179	0.0187			
	Naive	0.0429	0.0580	0.0595	0.0318	0.0488	0.0510			
	CV	0.0256	0.0274	0.0275	0.0129	0.0155	0.0171			
$N = 500$	$EB(c;\hat{\beta})$	0.0459	0.0426	0.0427	0.0257	0.0227	0.0227			
	Naive	0.0462	0.0549	0.0559	0.0313	0.0417	0.0431			
	CV	0.0389	0.0397	0.0390	0.0234	0.0245	0.0243			
$N = 2000$	$EB(c;\hat{\beta})$	0.0476	0.0463	0.0463	0.0280	0.0266	0.0264			
	Naive	0.0480	0.0519	0.0523	0.0293	0.0339	0.0344			
	CV	0.0462	0.0465	0.0462	0.0274	0.0278	0.0274			

$EB(c;\hat{\beta})^{\star}$ is expected benefit of a treatment selection rule based on estimated risk model given cost ratio c;
Naive* is the naive estimate of $EB(c;\hat{\beta})^{\star}$ based on training data;
CV† is the estimate of $EB(c;\hat{\beta})^{\star}$ based on random cross-validation.

Table 2: Cross-Validated estimates of $EB(c)$ in DCCT example.

Cost ratio c	0	0.05	0.10	0.12
PAR(Naive)	0.005	0.019	0.035	0.028
PAR(CV)	0.0051	0.016	0.020	0.012
NPAR(CV)	0.0060	0.018	0.022	0.012
AUG(CV)	0.0072	0.021	0.023	0.013

Supplementary Appendix

Here we provide rough sketches of the proofs of theorems stated in the paper.

Appendix A. Proof of Theorems 1 and 2

We assume the following conditions hold:

i) $\sqrt{N}(\hat{\beta} - \beta) = n^{-1/2} \sum_{i=1}^{n} I(\beta) l(\beta)_i + o_p(1)$

ii) $\rho_0 - \rho_1 \neq c$

iii) $E(Risk_0(\beta) - Risk_1(\beta))$ is differentiable with respect to β at the true β value

iv) $EB(c;\beta)$, $E(Risk_0(\beta) - Risk_1(\beta))_+$, $E(Risk_0(\beta))$, $E(Risk_0(\beta) + Risk_1(\beta) - 1)$ are differentiable with respect to β at the true β value

We prove the result for $EB(c)$ and the proofs for $PEB^l(c)$ and $PEB^u(c)$ in Theorem 2 follow similar arguments.

$$\begin{aligned}
& \sqrt{N}\left\{\widehat{EB}(c) - EB(c)\right\} \\
= \; & \sqrt{N}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\Delta}_i - c\right)_+ - E\left(\Delta - c\right)_+\right\} - \sqrt{N}\left\{\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\Delta}_i - c\right)_+ - I(\rho_0 - \rho_1 > c)\right\} \\
= \; & \sqrt{N}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\Delta}_i - c\right)_+ - \frac{1}{N}\sum_{i=1}^{N}(\Delta_i - c)_+\right\} + \sqrt{N}\left\{\frac{1}{N}\sum_{i=1}^{N}(\Delta_i - c)_+ - E(\Delta - c)_+\right\} \\
- \; & \sqrt{N}\left\{\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\Delta}_i - c\right)_+ - (\rho_0 - \rho_1 - c)_+\right\} \\
= \; & \sqrt{N}\left[E\left\{\Delta(\hat{\beta}) > c\right\}_+ - E(\Delta - c)_+\right] + \sqrt{N}\left\{\frac{1}{N}\sum_{i=1}^{N}(\Delta_i - c)_+ - E(\Delta - c)_+\right\} \\
- \; & A + o_p(1)
\end{aligned}$$

$$\begin{aligned}
\text{where} \quad A \; = \; & \sqrt{N}\left\{\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\Delta}_i - c\right)_+ - (\rho_0 - \rho_1 - c)_+\right\} \\
= \; & \sqrt{N}\left\{\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\Delta}_i - c\right)_+ - \left(\frac{1}{N}\sum_{i=1}^{N}\Delta_i - c\right) \times I(\rho_0 - \rho_1 - c > 0)\right\} \\
+ \; & \sqrt{N}\left\{\left(\frac{1}{N}\sum_{i=1}^{N}\hat{\Delta}_i - c\right) \times (I\rho_0 - \rho_1 - c > 0) - (\rho_0 - \rho_1 - c)_+\right\}
\end{aligned}$$

which when $\rho_0 - \rho_1 \neq c$ by equi-continuity equals to

$$\begin{aligned}
& \sqrt{N} \times (\rho_0 - \rho_1 - c) \times \left\{ I\left(\frac{1}{N} \sum_{i=1}^{N} \hat{\Delta}_i - c > 0 \right) - I(\rho_0 - \rho_1 - c > 0) \right\} \\
+ \quad & \sqrt{N} \left\{ \frac{1}{N} \sum_{i=1}^{N} \hat{\Delta}_i - (\rho_0 - \rho_1) \right\} \times I(\rho_0 - \rho_1 - c > 0) + o_p(1) \\
= \quad & \sqrt{N} \left\{ \frac{1}{N} \sum_{i=1}^{N} \hat{\Delta}_i - (\rho_0 - \rho_1) \right\} \times I(\rho_0 - \rho_1 - c > 0) + o_p(1),
\end{aligned}$$

which equals to $\sqrt{N}\left\{\sum_{i=1}^{N} \hat{\Delta}_i / N - (\rho_0 - rho_1)\right\}$ for $\rho_0 - \rho_1 > c$ and equals to 0 for $\rho_0 - \rho_1 < c$.

Appendix B. Proof of Theorem 3 and Corollary 1

We prove the result for $EB(c)$ as the proofs for $PEB^l(c)$ and $PEB^u(c)$ are similar.

Suppose τ_N is a positive sequence of random variables converging to zero almost surely with n and satisfying $\sqrt{N}\tau_N \to \infty$ almost surely as $n \to \infty$. Define the event $\mathcal{E} \triangleq \{|\hat{\rho}_0 - \hat{\rho}_1 - c| \leq \tau_N\}$ then $1_{\mathcal{E}} \to 1_{\rho_0 - \rho_1 = c}$ in probability. Thus, the validity of the confidence interval follows if: (i) the projection interval provides the correct coverage when $\rho_0 - \rho_1 = c$; and (ii) the standard bootstrap confidence interval provides the correct coverage when $\rho_0 - \rho_1 \neq c$. We next sketch the argument that the projection interval is valid in both (i) and (ii).

Define $EB_r(c) \triangleq E\left[\Delta(Y) - c\right]_+ - E(\Delta(Y) - c)1_{r \geq 0}$. We show that $\sqrt{N}(\widehat{EB}_r(c) - EB_r(c))$ and $\sqrt{N}(\widehat{EB}_r^b(c) - \widehat{EB}_r(c))$ converge to the same limiting distribution in probability. Thus, the validity of the proposed confidence intervals follows from standard arguments for the validity of projection intervals (see, for example, Berger and Boos (1994)). To simplify our proofs we assume that Y is bounded with probability one. Let $l(\beta^*)$ denote the influence function of $\sqrt{N}(\widehat{\beta} - \beta^*)$. Without loss of generality we assume $c = 0$.

For $\theta \in \mathbb{R}^{2\dim(Y)+2}$ define

$$\Delta(Y; \theta) \triangleq g^{-1}\left(\theta_0 + \theta_2^T Y\right) - g^{-1}\left(\theta_0 + \theta_1 + (\theta_2 + \theta_3)^T Y\right),$$

where g is the logit function. Note that $\Delta(Y) = \Delta(Y;\beta)$ and $\widehat{\Delta}(Y) = \Delta(Y;\widehat{\beta})$. Define $\dot{\Delta}(Y;\theta) \triangleq (d/d\theta)\Delta(Y;\theta)$, then for any compact set $\mathcal{K} \subseteq \mathbb{R}^{2\dim(Y)+2}$ the class of functions $\{||\dot{\Delta}(y;\theta)|| : \mathbb{R}^{\dim(Y)} \to \mathbb{R}\,,\ \theta \in \mathcal{K}\}$ is Donsker (see, for example, Kosorok (2008)). Write $\widehat{E}_N$ to denote expectation with respect to the empirical distribution. Then

$$\sqrt{N}(\widehat{EB}_r(0)-EB_r(0)) = \sqrt{N}\left(\widehat{E}_N\left[\widehat{\Delta}(Y)\right]_+ - \widehat{E}_N\widehat{\Delta}(Y)1_{r\geq 0}\right) - \sqrt{N}\left(E\left[\Delta(Y)\right]_+ - E\Delta(Y)1_{r\geq 0}\right),$$

which we can expand to equal

$$\sqrt{N}\widehat{E}_N\left(\left[\widehat{\Delta}(Y)\right]_+ - [\Delta(Y)]_+\right) - \widehat{E}_N\sqrt{N}\left(\widehat{\Delta}(Y) - \Delta(Y)\right)1_{r\geq 0} \\ + \sqrt{N}(\widehat{E}_N - E)\left([\Delta(Y)]_+ - \Delta(Y)1_{r\geq 0}\right),$$

which equals

$$\widehat{E}_N\left(\left[\mathbb{Z}_N(Y) + \sqrt{N}\Delta(Y)\right]_+ - \left[\sqrt{N}\Delta(Y)\right]_+\right) \\ + \sqrt{N}(\widehat{E}_N - E)\left([\Delta(Y)]_+ - \Delta(Y)1_{r\geq 0} - 1_{r\geq 0}E(\dot{\Delta}(Y;\beta^*)^T)l(\beta^*)\right) + o_P(1),$$

where $\mathbb{Z}_N \triangleq \sqrt{N}(\widehat{\Delta}(Y) - \Delta(Y)) = \dot{\Delta}(Y;\tilde{\beta})^T\sqrt{N}(\widehat{\beta} - \beta^*)$ for some $\tilde{\beta}$ intermediate to $\widehat{\beta}$ and β^*. We now argue that the leading term in the above display is equal to $\sqrt{N}(\widehat{E}_N - E)\dot{\Delta}(Y;\beta^*)^Tl(\beta^*)1_{\Delta(Y)\geq 0} + o_P(1)$. The leading term in the above display is equal to

$$\widehat{E}_N\mathbb{Z}_N(Y)1_{\Delta(Y)\geq 0}1_{\sqrt{N}|\Delta(Y)|\geq|\mathbb{Z}_N(Y)|} \\ + \widehat{E}_N\left(\left[\mathbb{Z}_N(Y) + \sqrt{N}\Delta(Y)\right]_+ - \left[\sqrt{N}\Delta(Y)\right]_+\right)1_{\sqrt{N}|\Delta(Y)|\leq|\mathbb{Z}_N(Y)|}. \quad (1)$$

Note that $P\left(|\sqrt{N}\Delta(Y)| \leq |\mathbb{Z}_N(Y)|\right)$ is bounded above by

$$P\left(|\Delta(Y)| \leq \sup_{y\in\mathrm{supp}(Y)} ||\dot{\Delta}(y;\tilde{\beta})||\,||\widehat{\beta} - \beta^*||\right) \leq 2C\sup_{y\in\mathrm{supp}(Y)} ||\dot{\Delta}(y;\widetilde{\beta})||\,||\widehat{\beta} - \beta^*|| = o_P(1),$$

where C is an upper bound on the density of $\Delta(Y)$. Using $|[a+b]_+ - [b]_+| \leq [a]_+$ the second term in (1) is bounded above in magnitude by $\widehat{E}_N\left[\mathbb{Z}_N(Y)\right]_+ 1_{\sqrt{N}|\Delta(Y)|\leq|\mathbb{Z}_N|} = o_P(1)$.

The first term in (1) is equal to $\widehat{E}_N \mathbb{Z}_N(Y) 1_{\Delta(Y)\geq 0} + o_P(1)$, which in turn is equal to $\mathbb{E}\dot{\Delta}(Y;\beta^*)^T 1_{\Delta(Y)\geq 0}\sqrt{N}(\widehat{E}_N - E)l(\beta^*) + o_P(1)$.

Assembling the arguments made above, it follows that

$$\sqrt{N}(\widehat{EB}_r(0) - EB_r(0)) = \nu^T\sqrt{N}(\widehat{E}_N - E)\begin{pmatrix} [\Delta(Y)]_+ \\ \Delta(Y)1_{r\geq 0} \\ l(\beta^*) \end{pmatrix} + o_P(1),$$

where $\nu = (1,\, 1_{r\geq 0}, E\dot{\Delta}(Y;\beta^*)^T(1_{\Delta(Y)\geq 0} - 1_{r\geq 0}))^T$.

Following the same arguments, it can be shown that $\sqrt{N}(\widehat{EB}^b_r(0) - \widehat{EB}_r(0))$ equals

$$\sqrt{N}(\widehat{EB}^b_r(0) - EB^b_r(0)) = \nu^T\sqrt{N}(\widehat{E}_N - E)\begin{pmatrix} [\Delta(Y)]_+ \\ \Delta(Y)1_{r\geq 0} \\ l(\beta^*) \end{pmatrix} + o_{P^b}(1),$$

where ν is defined as above and we write $r_N = o_{P^b}(1)$ to mean $P^b(|r_N| \geq \epsilon) = o_P(1)$ for any $\epsilon > 0$. Note that $\sqrt{N}(\widehat{\Delta}^b(Y) - \widehat{\Delta}(Y)) = \dot{\Delta}(Y;\widetilde{\beta})^T\sqrt{N}(\widehat{E}^b_N - \widehat{E}_N)I(\beta^*) + o_{P^b}(1)$ where $\widetilde{\beta}$ is intermediate to $\widehat{\beta}^b$ and $\widehat{\beta}$, and

$$\begin{aligned} &P\left(|\widehat{\Delta}(Y)| \leq \sup_{y\in\mathrm{supp}(Y)} ||\dot{\Delta}(y;\widetilde{\beta})||\, ||\widehat{\beta}^b - \widehat{\beta}||\right) \\ &\qquad \leq P\left(|\Delta(Y)| \leq \sup_{y\in\mathrm{supp}(Y)} ||\dot{\Delta}(y;\widetilde{\beta})||\, ||\widehat{\beta}^b - \widehat{\beta}|| + \sup_{y\in\mathrm{supp}(Y)} |\widehat{\Delta}(y) - \Delta(y)|\right) = o_{P^b}(1), \end{aligned}$$

where again $\widetilde{\beta}$ is intermediate to $\widehat{\beta}^b$ and $\widehat{\beta}$.

It remains to show that the bootstrap confidence interval for $\widehat{EB}(c)$ is consistent when $\rho_0 - \rho_1 \neq 0$. In the above notation this requires showing $\sqrt{N}(\widehat{EB}^b_{\hat{\rho}^b_0 - \hat{\rho}^b_1}(0) - \widehat{EB}_{\hat{\rho}_0-\hat{\rho}_1}(0))$ and $\sqrt{N}(\widehat{EB}_{\hat{\rho}_0-\hat{\rho}_1}(0) - EB_{\rho_0-\rho_1}(0))$ converge to the same limiting distributions in probability. However, since $\hat{\rho}_0 - \hat{\rho}_1$ is a regular, (strongly) consistent estimator of $\rho_0 - \rho_1$ and $\rho_0 - \rho_1 \neq 0$, it follows that

$$\sqrt{N}(\widehat{EB}^b_{\hat{\rho}^b_0 - \hat{\rho}^b_1}(0) - \widehat{EB}_{\hat{\rho}_0-\hat{\rho}_1}(0)) = \sqrt{N}(\widehat{EB}^b_{\rho_0-\rho_1}(0) - \widehat{EB}_{\rho_0-\rho_1}(0)) + o_{P^b}(1),$$

and

$$\sqrt{N}(\widehat{EB}_{\hat{\rho}_0-\hat{\rho}_1}(0) - EB_{\rho_0-\rho_1}(0)) = \sqrt{N}(\widehat{EB}_{\rho_0-\rho_1}(0) - \widehat{EB}_{\rho_0-\rho_1}(0)) + o_P(1).$$

Thus, the projection interval proof for $r = \rho_0 - \rho_1$ applies.

Appendix C. Details supporting Remark 2

Remark 2: [Locally consistent confidence interval for $EB(c)$.]

For any $\eta \in \mathbb{R}^{\dim(\beta)}$ and $r \in \mathbb{R}$ define

$$\begin{aligned}\widehat{\theta}(\eta, r) &\triangleq \widehat{E}_N \Delta(Y; \widehat{\beta}_N) 1_{\Delta(Y;\eta)\geq 0}(1-c) - \widehat{E}_N(\Delta(Y;\eta) - c)1_{r-c\geq 0},\\ \theta(\eta, r) &\triangleq E\Delta(Y;\beta)1_{\Delta(Y;\eta)\geq 0}(1-c) - E(\Delta(Y;\eta)-c)1_{r-c\geq 0}.\end{aligned}$$

Note that $\theta(\beta, E\Delta(Y;\beta)) = EB(c)$. For every fixed η, r pair it can be shown that $\sqrt{N}(\widehat{\theta}(\eta, r) - \theta(\eta, r))$ is regular, asymptotically normal, and for any $\delta \in (0,1)$ a $(1-\delta)\times 100\%$ confidence interval for $\theta(\beta, E\Delta(Y;\beta))$ can be obtained via the bootstrap. Denote such an interval by $\xi_\delta(\eta, r)$. Thus, were β^* and $E\Delta(Y;\beta)$ known, one could bootstrap $\sqrt{N}\left(\widehat{\theta}(\beta, E\Delta(Y;\beta)) - \theta(\beta, E\Delta(Y;\beta))\right)$ (but holding β and $E\Delta(Y;\beta)$ to be fixed) to obtain a valid confidence interval for $\theta(\beta, E\Delta(Y;\beta))$. Of course, neither β nor $E\Delta(Y;\beta)$ are known; however, for any $\alpha \in (0,1)$ standard methods can be used to construct a $(1-\alpha)\times 100\%$ joint confidence region for $(\beta, E\Delta(Y;\beta))$, say Γ_α. Then, it follows that

$$\bigcup_{(\eta,r)\in\Gamma_\alpha} \xi_\delta(\eta, r),$$

is a valid $(1-\delta-\alpha)\times 100\%$ confidence interval for $\theta(\beta, E\Delta(Y;\beta)) = EB(c)$. This procedure involves the union of smooth, regular, confidence intervals and is therefore also regular (i.e., locally consistent). The above interval takes a union over a larger set and is therefore potentially more conservative than the interval described in Section 4.2. On the other hand, a smooth density for $\Delta(Y)$ is no longer required (details omitted).

References

Berger, Roger L and Boos, Dennis D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89**(427), 1012–1016.

Kosorok, Michael R. (2008). *Empirical Processes and Semiparametric Inference*. Springerverlag New York.