Biostatistics (2015), **16**, 2, *pp.* 383–399
doi:10.1093/biostatistics/kxu039
Advance Access publication on September 3, 2014

Characterizing expected benefits of biomarkers in treatment selection

YING HUANG*

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle WA, 98109, USA

yhuang@fhcrc.org

ERIC B. LABER

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, NC, 27695-8203, USA

HOLLY JANES

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle WA, 98109, USA

Summary

Biomarkers associated with heterogeneity in subject responses to treatment hold potential for treatment selection. In practice, the decision regarding whether to adopt a treatment-selection marker depends on the effect of using the marker on the rate of targeted disease and on the cost associated with treatment. We propose an expected benefit measure that incorporates both effects to quantify a marker's treatment-selection capacity. This measure builds upon an existing decision-theoretic framework, but is expanded to account for the fact that optimal treatment absent marker information varies with the cost of treatment. In addition, we establish upper and lower bounds on the expected benefit for a perfect treatment-selection rule which provides the basis for a standardized expected benefit measure. We develop model-based estimators for these measures in a randomized trial setting and evaluate their asymptotic properties. An adaptive bootstrap confidence interval is proposed for inference in the presence of non-regularity. Alternative estimators robust to risk model misspecification are also investigated. We illustrate our methods using the Diabetes Control and Complications Trial where we evaluate the expected benefit of baseline hemoglobin A1C in selecting diabetes treatment.

Keywords: Adaptive bootstrap; Biomarker; Expected benefit; Potential outcomes; Treatment selection.

1. Introduction

In many clinical settings for disease prevention and treatment, there is significant heterogeneity in subject response to the same treatment. Biomarkers associated with this heterogeneity, such as demographic or genetic characteristics, can be used to help subjects select treatment to ensure that a therapy is only delivered to subjects who are likely to benefit from it.

Statistical measures for quantifying the performance of candidate treatment-selection markers are essential for developing these markers and evaluating their clinical impact. Testing for a marker by

*To whom correspondence should be addressed.

treatment interaction is a common strategy for identifying treatment-selection markers. However, there is a growing emphasis on developing measures of treatment-selection ability that are directly linked to clinical outcomes. Much of this work focuses on the effect of using the marker on the targeted disease. For example, the reduction in population disease rate as a result of treatment selection (Song and Pepe, 2004; Zhang, Tsiatis, Laber, *and others*, 2012); the accuracy for classifying a subject into treatment-effective or ineffective categories (Huang *and others*, 2012); the distribution of the disease risk difference conditional on a marker in the population or in the marker-positive group (Cai *and others*, 2011; Foster *and others*, 2011; Huang *and others*, 2012; Zhao *and others*, 2013), all quantify, in some fashion, the impact of using the marker to select treatment on the rate of disease. In practice, a treatment may affect the population through not only its effect on the targeted disease but also 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*, 2014). One way to incorporate both disease risk and proportion treated is to adopt a decision-theoretic framework that puts the marker's effect on disease and the proportion treated on the same scale by means of a treatment-disease cost ratio. An example is the net benefit measure characterizing the reduction in the sum of disease and treatment cost comparing a marker-based treatment strategy with the strategy of treating no one (Vickers *and others*, 2007, Rapsomaniki *and others*, 2012).

In this paper, we develop a new measure named *expected benefit* that extends the net benefit measure to quantify the reduction in the sum of disease and treatment cost by using the marker. This measure is based on the comparison between a marker-based treatment-selection rule and the optimal treatment strategy absent the marker information, where the latter is allowed to vary with the treatment–disease cost ratio. 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 using a perfect treatment-selection rule. While the latter is not identifiable in general, we show that upper and lower bounds can be established through a potential outcomes framework. We develop a model-based strategy for deriving the treatment-selection rule and for estimating the corresponding (standardized) expected benefit using data from a randomized trial, and develop asymptotic theory for the estimators. The expected benefit is not a smooth function of the generative model which can cause the standard bootstrap to fail; therefore, we develop a novel adaptive bootstrap confidence interval that provides consistent inference. We also investigate alternative strategies for deriving the treatment-selection rule and estimators of expected benefit that are robust to model misspecification.

In Section 2, we introduce the concept of expected benefit, derive bounds on the expected benefit of a perfect treatment-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 the Diabetes Control and Complications Trial is presented in Section 5. We then conclude the paper with a summary and discussion.

2. METHODS

We consider the setting of a randomized trial with two arms, $T = 0, 1$ indicating the untreated and treated groups, respectively. Let D be a binary outcome that the treatment is intended to prevent, which we call "targeted disease," with $D = 0, 1$ indicating control and case status, respectively, and $\rho_0 = P(D = 1|T = 0)$ and $\rho_1 = P(D = 1|T = 1)$ indicating disease prevalence in untreated and treated groups. Let Y denote the biomarker of interest, which may be univariate or multivariate. Let $A(Y)$ be a treatment-selection rule based on the marker, which takes values 1 and 0 corresponding to the recommendation for or against the treatment, respectively. Let i be subject indicator. With N participants in the trial, we observe i.i.d. data $(Y_i, T_i, D_i),\ i = 1, \ldots, N$.

As in Vickers *and others* (2007), we assume the cost of the treatment due to side effects, subject burden, and/or 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), based on a patient survey, c was chosen to be 5% for treating breast cancer with adjuvant chemotherapy, which corresponds to assuming that the cost of death is 20 times the cost of chemotherapy. Without loss of generality, let the cost per disease outcome be 1 such that disease and treatment *cost* will be represented in units of the burden per disease outcome. The total cost of a treatment-selection strategy $A(Y)$ and of the optimal treatment strategy absent the marker information can thus be computed as follows.

At cost ratio c, the total cost of a treatment-selection rule $A(Y)$, i.e. the sum of disease and treatment cost, is equal to $\sum_{a=0}^{1} P\{A(Y)=a\} \times P\{D=1|A(Y)=a, T=a\} + P\{A(Y)=1\} \times c$. As shown in supplementary material available at *Biostatistics* online, Appendix A, this equals

$$\rho_0 - E[A(Y) \times \{\Delta(Y) - c\}], \tag{2.1}$$

where $\Delta(Y) = P(D=1|T=0, Y) - P(D=1|T=1, Y)$ is the absolute risk difference conditional on Y between untreated and treated. Without the biomarker, treatment will be applied either to all subjects or to no one. If treating all, the total cost is $\rho_1 + c$, and if treating none, the total cost is ρ_0. Therefore, absent the marker, the optimal treatment-selection rule that minimizes the total cost is to treat everyone if $\rho_0 > \rho_1 + c$, and to treat no one otherwise (Vickers *and others*, 2007). The total cost of the optimal marker-independent rule is therefore

$$\rho_0 - [\rho_0 - \rho_1 - c]_+, \tag{2.2}$$

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

We define the expected benefit for a rule $A(Y)$ and cost ratio c as the difference between (2.2) and (2.1), i.e. $\mathrm{EB}_A(c)$ is the reduction in the total cost using $A(Y)$ relative to the optimal rule absent the marker:

$$\mathrm{EB}_A(c) = E[A(Y) \times \{\Delta(Y) - c\}] - [\rho_0 - \rho_1 - c]_+. \tag{2.3}$$

Note that the first component of (2.3) is exactly the net benefit measure of $A(Y)$ (Vickers *and others*, 2007); the second component of (2.3) is 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 with the optimal treatment strategy without the biomarker. If $c = 0$ and $\rho_0 > \rho_1$, this reduces to the decrease in the disease rate, a measure advocated by Janes *and others* (2014). Hereafter, to simplify notation we write $\mathrm{EB}(c)$ with the understanding that the underlying strategy $A(Y)$ is implicit.

The incremental value in net benefit by using the marker is informative for several reasons. Net benefit itself is useful for comparing treatment-selection strategies because the difference in net benefit between two models is equal to the difference in their expected benefits, as the second component of (2.3) does not depend on the marker. However, the net benefit of a marker-based strategy does not always properly quantify the absolute benefit gained by the marker for different choices of the treatment–disease cost ratio because the default strategy of no treatment may not be the optimal strategy absent the marker information. The expected benefit measure, in contrast, takes into account the optimal treatment choice absent the marker. Moreover, when we evaluate whether a new model improves over an existing marker/model, the expected benefit of the existing marker can serve as a useful reference for gauging whether the difference in expected benefit between models is meaningful.

In practice, it is difficult to agree upon one single parameter c. An expected benefit curve, which plots $\mathrm{EB}_A(c)$ versus c, can be used to gauge marker value at multiple values of c. Examples of EB curves are shown in Figure 1 in Section 5.

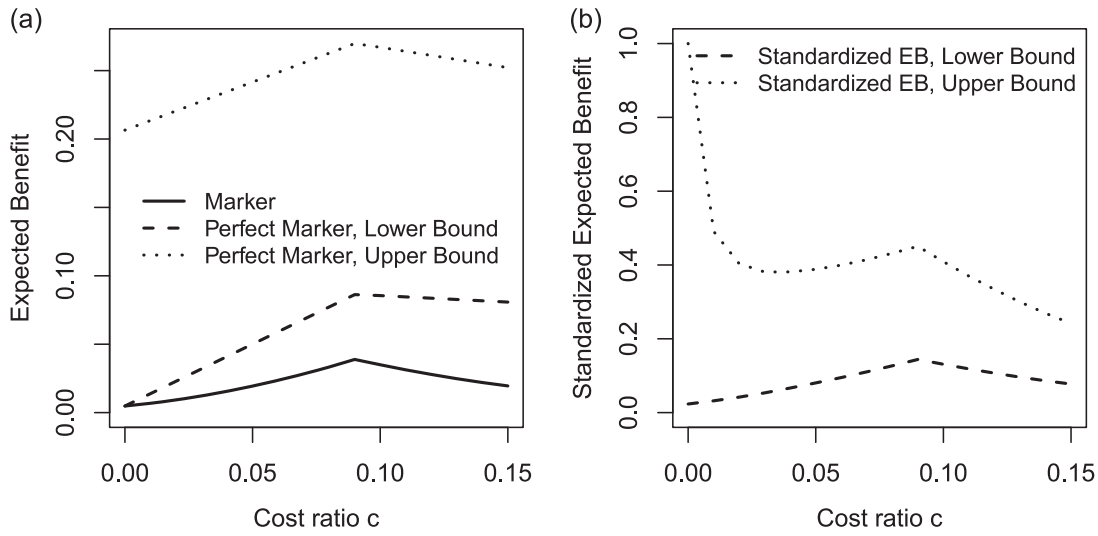

Fig. 1. Expected benefit curves of HBA1C and the bounds for perfect biomarker for guiding the prevention of microalbuminuria in the DCCT example.

2.1 *Perfect treatment-selection capacity and standardized expected benefit*

In this section, we derive the expected benefit of a perfect treatment-selection rule which can be used to standardize the expected benefit of a marker-based rule. This type of standardization puts the expected benefit measure on a scale between 0 and 1 and makes it invariant to the choice of disease cost. Standardization has been used when measuring a biomarker's capacity for risk prediction, e.g. in Bura and Gastwirth (2001) and Baker *and others* (2009), but not yet for treatment selection.

We define a perfect treatment-selection rule using a potential outcomes framework. Let $D(1)$, $D(0)$ denote the pair of potential outcomes under treatment or no treatment, respectively. The four possible values of $D(0)$, $D(1)$ are shown below with q_1, q_2, q_3, and q_4 denoting the unobserved population proportion of subjects falling into each category, where benefits and harms are with respect to the targeted disease of interest.

$D(0)$	$D(1)$		Proportion
1	0	Benefited	q_1
1	1	Unaffected	q_2
0	0	Unaffected	q_3
0	1	Harmed	q_4

Given cost ratio $c \geqslant 0$, a perfect treatment-selection rule will recommend treatment for all treatment-benefited subjects and recommend against treatment for all others. This will lead to a population-averaged total disease and treatment cost of $q_2 + q_1 c = \rho_0 - q_1 + q_1 c$. Thus, the expected benefit of a perfect treatment-selection rule, i.e. the "perfect expected benefit" (PEB), is $\mathrm{PEB}(c) = q_1(1-c) - [\rho_0 - \rho_1 - c]_+$.

While in general q_1 is not identifiable given the observed data, upper and lower bounds can be identified using a disease risk model, i.e. the model of the risk of D conditional on Y and T. Let $q_k(Y)$, $k = 1, 2, 3$, and 4 indicate the probability that a subject with marker Y falls into the ith potential outcome category.

Let $\rho_0(Y) = P(D = 1|Y, T = 0)$, $\rho_1(Y) = P(D = 1|Y, T = 1)$. Then

$$q_1(Y) + q_2(Y) = \rho_0(Y) \Rightarrow q_1(Y) \leqslant \rho_0(Y),$$
$$q_1(Y) + q_3(Y) = 1 - \rho_1(Y) \Rightarrow q_1(Y) \leqslant 1 - \rho_1(Y),$$
$$q_1(Y) - q_4(Y) = \rho_0(Y) - \rho_1(Y) \Rightarrow q_1(Y) \geqslant \rho_0(Y) - \rho_1(Y),$$

which implies $\max\{0, \rho_0(Y) - \rho_1(Y)\} \leqslant q_1(Y) \leqslant \min\{\rho_0(Y), 1 - \rho_1(Y)\}$. Taking an expectation over Y, we have $E[\{\Delta(Y)\}_+] \leqslant q_1 \leqslant \rho_0 - E[\{\rho_0(Y) + \rho_1(Y) - 1\}_+]$. Note that alternative nonparametric bounds for q_1 can be derived without relying on any biomarker model: $\max(0, \rho_0 - \rho_1) = [E\{\Delta(Y)\}]_+ \leqslant q_1 \leqslant \min(\rho_0, 1 - \rho_1) = \rho_0 - [E\{\rho_0(Y) + \rho_1(Y) - 1\}]_+$. The bounds constructed based on the disease risk model are narrower since $[E\{\Delta(Y)\}]_+ \leqslant E[\{\Delta(Y)\}_+]$ and $[E\{\rho_0(Y) + \rho_1(Y) - 1\}]_+ \leqslant E[\{\rho_0(Y) + \rho_1(Y) - 1\}_+]$, 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), and Zhang *and others* (2013).

Using disease risk model, we obtain lower and upper bounds for the perfect expected benefit

$$\mathrm{PEB}^{\mathrm{l}}(c) = E[\{\Delta(Y)\}_+] \times (1 - c) - [\rho_0 - \rho_1 - c]_+, \tag{2.4}$$
$$\mathrm{PEB}^{\mathrm{u}}(c) = (\rho_0 - E[\{\rho_0(Y) + \rho_1(Y) - 1\}_+]) \times (1 - c) - [\rho_0 - \rho_1 - c]_+. \tag{2.5}$$

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

In summary, the expected benefit of a perfect treatment-selection rule can be derived but is non-identifiable due to the non-identifiability of the percent of "benefited" individuals (or equivalently, the percent of "harmed" or "unbenefited" since q_1 can be determined when any of q_2, q_3, or q_4 is fixed). An alternative to constructing bounds on q_1 using a risk model is to conduct a sensitivity analysis treating the percent harmed (q_4) as a sensitivity parameter, and to compute PEB for each fixed q_4 value. A narrower bound for PEB might be achieved when a narrow range of q_4 is plausible based on biological assumptions. 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 $q_1 = \rho_0 - \rho_1$, and PEB can be uniquely identified as $(\rho_0 - \rho_1) \times (1 - c) - [\rho_0 - \rho_1 - c]_+$, which is equal to its lower bound in (2.4) since $E[\{\Delta(Y)\}_+] = E\{\Delta(Y)\} = \rho_0 - \rho_1$ under monotonicity.

The expected benefit from perfect treatment selection sets a reference for gauging the benefit of a particular treatment-selection rule or the difference in benefit between two rules. A demonstration of the comparison between two markers is presented in supplementary material available at *Biostatistics* online, Figure 1.

3. Derivation of treatment-selection rules and estimation of expected benefit

In this section, we consider methods for deriving marker-based treatment-selection rules in order to maximize expected benefit and to estimate their (standardized) expected benefit. We consider the class of selection rules $A(Y)$ which depends on the sign of a smooth function of Y, namely $h(Y)$, i.e. $A(Y) = I\{h(Y) > 0\}$. We propose two strategies for optimizing the selection rule $A(Y)$ and estimating (standardized) expected benefit. The first requires correct modeling of the risk difference $\Delta(Y)$, whereas the second does not and is thus more robust to model misspecification. However, the first strategy is more efficient under a correctly specified model.

3.1 *Model-based approach*

Based on (2.3), 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. That is, $h(Y) = \Delta(Y) - c$. For details, see, e.g. Vickers *and others* (2007). We model the disease risk with a generalized linear model (GLM): $g\{P(D=1|Y,T)\} = \beta_0 + \beta_1 T + \beta_2^T Y + \beta_3^T YT$, where g is a known link function, e.g. the logit or inverse normal CDF. Let $\hat{\beta}$ denote the maximum likelihood estimator (MLE) of β, and let $\hat{\Delta}(Y)$ denote the MLE of $\Delta(Y)$. A model-based treatment-selection rule can be constructed as $A(Y) = I\{\hat{\Delta}(Y) > c\}$. Assuming that the model for $\Delta(Y)$ is correctly specified, a model-based estimator of expected benefit can be constructed based on (2.3): $\widehat{\mathrm{EB}}(c) = \sum_{i=1}^N (\hat{\Delta}_i - c)_+/N - \left(\sum_{i=1}^N \hat{\Delta}_i/N - c\right)_+$, where $\hat{\Delta}_i = \hat{\Delta}(Y_i)$ is the estimate of $\Delta(Y)$ for subject i. Note that a good fit of the risk model itself is sufficient for the good fit of the model for $\Delta(Y)$, but not necessary. Hosmer–Lemeshow type techniques can be used to evaluate both types of calibration (Huang and Pepe, 2010; Janes *and others*, 2014).

We estimate the lower bound on PEB with $\widehat{\mathrm{PEB}}^{\mathrm{l}}(c) = \sum_{i=1}^N (\hat{\Delta}_i)_+ \times (1-c)/N - (\sum_{i=1}^N \hat{\Delta}_i/N - c)_+$ and the upper bound with $\widehat{\mathrm{PEB}}^{\mathrm{u}}(c) = \sum_{i=1}^N \{\widehat{\mathrm{Risk}}_{0i} - (\widehat{\mathrm{Risk}}_{0i} + \widehat{\mathrm{Risk}}_{1i} - 1)_+\} \times (1-c)/N - \left(\sum_{i=1}^N \hat{\Delta}_i/N - c\right)_+$, where $\widehat{\mathrm{Risk}}_0$ and $\widehat{\mathrm{Risk}}_1$ 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 $\mathrm{SEB}(c)$ can be estimated as $\widehat{\mathrm{EB}}(c)/\widehat{\mathrm{PEB}}^{\mathrm{u}}(c)$ and $\widehat{\mathrm{EB}}(c)/\widehat{\mathrm{PEB}}^{\mathrm{l}}(c)$.

3.2 *Robust approaches*

Optimality of a model-based treatment-selection rule $A(Y) = I\{\hat{\Delta}(Y) > c\}$ and validity of the model-based estimator for corresponding expected benefit rely critically on correct specification of the model for $\Delta(Y)$. Here we describe alternative approaches for characterizing the optimal treatment-selection rule and for estimating expected benefit that are more robust to model misspecification, in the sense that good performance of the derived rule and unbiasedness of the estimates of EB do not require correct specification of $\Delta(Y)$.

Suppose that we are interested in picking the best rule among all rules of the form $A(Y) = I\{h(Y) > 0\}$ with $h(Y)$ belonging to a pre-specified class, e.g. $h(Y)$ can be a linear function of Y. The optimal rule can be estimated by maximizing an estimate of the expected benefit. Following the notations in Section 2.1, let $D(t)$ indicate the potential disease outcome if a subject were to receive treatment t, $t = 0, 1$. Let $D(A) = D(0) \times I\{A(Y) = 0\} + D(1) \times I\{A(Y) = 1\}$ be the potential outcome that would be observed if a randomly chosen subject from the population were to be assigned treatment according to rule $A(Y)$. The expected benefit under $A(Y)$ can be represented as $\mathrm{EB}(c) = \rho_0 - E\{D(A)\} - E\{A(Y)\} \times c - [\rho_0 - \rho_1 - c]_+$. Let $C_A = T \times A(Y) + (1-T) \times \{1 - A(Y)\}$ be the indicator of observing $D(A)$, i.e. $C_A = 1$ if $A(T) = T$. Then to estimate $E\{D(A)\}$, one can use the inverse-probability weighted estimator (IPWE)

$$\frac{1}{N}\sum_{i=1}^N \frac{C_{Ai} D_i}{P(C_A = 1|Y_i)}, \tag{3.1}$$

or a doubly robust augmented IPWE

$$\frac{1}{N}\sum_{i=1}^N \frac{C_{Ai} D_i}{P(C_A = 1|Y_i)} - \frac{C_{Ai} - P(C_A = 1|Y_i)}{P(C_A = 1|Y_i)} m(Y_i; \hat{\beta}), \tag{3.2}$$

where $P(C_A = 1|Y) = P\{A(Y) = T|Y\} = P\{T = A(Y) = 1|Y\} + P\{T = A(Y) = 0|Y\} = P(T = 1|Y) \times A(Y) + P(T = 0|Y) \times \{1 - A(Y)\}$; $m(Y_i; \hat{\beta})$, an estimate of $E\{D(A)|Y_i\}$ equals $\widehat{\mathrm{Risk}}_1(Y_i)A(Y_i) -$

$\widehat{\text{Risk}}_0(Y_i)\{1 - A(Y_i)\}$ based on some working risk model. The doubly robust estimator augments the IPWE of $E\{D(A)\}$ with a term that involves the risk of D given Y and T. It has the double-robustness property in that it is consistent for $E\{D(A)\}$ if either $P(T = 1|Y)$ or the risk model is correctly specified. In a randomized trial, $P(T = 1|Y)$ is known, so consistency of the estimator is always achievable; the second term in (3.2) "augments" the empirical estimate so as to increase asymptotic efficiency as shown in Zhang, Tsiatis, Laber, *and others* (2012).

Based on the IPWE and the augmented estimator of $E\{D(A)\}$ in (3.1) and (3.2), corresponding empirical and augmented estimators for EB(c) are

$$\frac{\sum_{i=1}^N D_i \times [1 - I\{h(Y_i) > 0\}] \times (1 - T_i)}{\sum_{i=1}^N (1 - T_i)} - \frac{\sum_{i=1}^N D_i \times I\{h(Y_i) > 0\} \times T_i}{\sum_{i=1}^N T_i}$$
$$- c \times \frac{1}{N} \sum_{i=1}^N I\{h(Y_i) > 0\} - \left[\frac{\sum_{i=1}^N D_i \times (1 - T_i)}{\sum_{i=1}^N (1 - T_i)} - \frac{\sum_{i=1}^N D_i \times T_i}{\sum_{i=1}^N T_i} - c \right]_+, \tag{3.3}$$

and

$$(3.3) + \frac{1}{N} \sum_{i=1}^N \frac{T_i \times I\{h(Y_i) > 0\} + (1 - T_i) \times I\{h(Y_i) > 0\} - \pi(Y; h)}{\pi(Y; h)}$$
$$\times [\widehat{\text{Risk}}_1 \times I\{h(Y_i) > 0\} + \widehat{\text{Risk}}_0 \times I\{h(Y_i) \leqslant 0\}], \tag{3.4}$$

respectively, with $\pi(Y; h) = P(T = 1)I\{h(Y) > 0\} + P(T = 0)I\{h(Y) \leqslant 0\}$.

As in Zhang, Tsiatis, *and others* (2012) and Zhao *and others* (2012), the problem of maximizing the expected benefit estimators in (3.3) and (3.4) can be transformed into a weighted classification problem. Consider deriving a rule based on a linear marker combination $h(Y) = \alpha_0 + \alpha_1^T Y$. As shown in supplementary material available at *Biostatistics* online, Appendix C, the values α_0, α_1 that maximize these expect benefit estimates can be shown to be the minimizers of

$$\sum_{i=1}^N |W_i| I\{\text{sgn}(W_i) \neq \text{sgn}(\alpha_0 + \alpha_1 Y_i)\}, \tag{3.5}$$

with $W_i = -D_i T_i/(N_1/N) + D_i(1 - T_i)/(N_0/N) - c$ for maximizing (3.3) and $W_i = -D_i T_i/N_1 + D_i(1 - T_i)/N_0 - \{P(T = 1) - T_i\}/P(T = 0) \times \widehat{\text{Risk}}_0 + \widehat{\text{Risk}}_1/N - c/N$ for maximizing (3.4), where N_1 and N_0 are sample sizes in treated and untreated groups. We consider two algorithms in the simulation studies to derive α_0, α_1 for minimizing (3.5), one fitting a weighted linear logistic regression model regressing binary outcome $\text{sgn}(W_i)$ versus Y_i with individual weight $|W_i|$, the other directly solving for α_0, α_1 by minimizing (3.5) through a grid search.

Finally, we consider an alternative robust approach for deriving $h(Y)$ that is computationally simpler. We adopt a working model $\Delta(Y) = \text{expit}(\beta_0 + \beta_2^T Y) - \text{expit}(\beta_0 + \beta_1 + \beta_2^T Y + \beta_3^T YT)$ for risk difference based on a GLM. Let $\hat{\Delta}(Y)$ be the MLE of $\Delta(Y)$, and $h(Y) = \hat{\Delta}(Y) - \delta$. An optimal δ can be identified by maximizing the empirical (3.3) or doubly robust (3.4) estimator of EB using this $h(Y)$ and a grid search. Note EB(c) in a randomized trial can be represented as

$$P\{D = 1, A(Y) = 1|T = 0\} - P\{D = 1, A(Y) = 1|T = 1\} - P\{A(Y) = c\} \times c - [\rho_0 - \rho_1 - c]_+ \tag{3.6}$$
$$= [P\{D = 1|A(Y) = 1, T = 0\} - P\{D = 1|A(Y) = 1, T = 1\} - c] \times P\{A(Y) = 1\}$$
$$- [\rho_0 - \rho_1 - c]_+. \tag{3.7}$$

The IPWE estimator (3.3) is an empirical estimator for estimating EB as represented in (3.6), whereas an alternative empirical estimator for estimating EB as represented in (3.7) can be constructed by maximizing δ over

$$\left\{\frac{\sum_{i=1}^N D_i \times I(\hat{\Delta}_i > \delta) \times (1-T_i)}{\sum_{i=1}^N I(\hat{\Delta}_i > \delta) \times (1-T_i)} - \frac{\sum_{i=1}^N D_i \times I(\hat{\Delta}_i > \delta) \times T_i}{\sum_{i=1}^N I(\hat{\Delta}_i > \delta) \times T_i} - c\right\} \times \frac{1}{N}\sum_{i=1}^N I(\hat{\Delta}_i > \delta)$$

$$- \left[\frac{\sum_{i=1}^N D_i \times (1-T_i)}{\sum_{i=1}^N (1-T_i)} - \frac{\sum_{i=1}^N D_i \times (T_i)}{\sum_{i=1}^N T_i} - c\right]_+ . \tag{3.8}$$

The empirical estimator (3.8) uses data more efficiently compared with the IPWE estimator (3.3) by taking into account the condition: $P\{A(Y)=1|T=0\} = P\{A(Y)=1|T=1\} = P\{A(Y)=1\}$, ensured by randomization. We use this estimator together with the augmented estimator in our simulation study and data example for identifying δ based on $\hat{\Delta}(Y)$.

Expected benefits of these "robust treatment rules" can be estimated using cross-validation. These robust methods target scenarios where the model for risk and/or risk difference is prone to misspecification. The bounds for PEB in (2.4) and (2.5) rely heavily on correct specification of the risk model. Therefore, we do not consider robust estimation techniques for the bounds.

3.3 *Asymptotic theory for the model-based estimator of expected benefit*

When $\rho_0 - \rho_1 \neq c$, the following theorem holds as proved in supplementary material available at *Biostatistics* online, Appendix D.

THEOREM 1 Under the regularity conditions specified in supplementary material available at *Biostatistics* online, Appendix D, $\widehat{\mathrm{EB}}(c)$, $\widehat{\mathrm{PEB}}^{\mathrm{l}}(c)$, $\widehat{\mathrm{PEB}}^{\mathrm{u}}(c)$, $\widehat{\mathrm{SEB}}^{\mathrm{l}}(c)$, and $\widehat{\mathrm{SEB}}^{\mathrm{u}}(c)$ as defined in Section 3.1 are asymptotically normal as $N \to \infty$ for $c \neq \rho_0 - \rho_1$.

When $c = \rho_0 - \rho_1$, $\sqrt{N}\{(\sum_{i=1}^N \hat{\Delta}_i/N - c)_+ - (\rho_0 - \rho_1 - c)_+\}$ converges to a mixture of 0 and a truncated normal distribution (supplementary material available at *Biostatistics* online, Appendix E). As a result, asymptotic normality of $\widehat{\mathrm{EB}}(c)$, $\widehat{\mathrm{PEB}}(c)$, or $\widehat{\mathrm{SEB}}(c)$ 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. When $c \approx \rho_0 - \rho_1$, the standard bootstrap percentile confidence interval (CI) can lead to undercoverage. Therefore, we adopt an adaptive bootstrap CI 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 CI 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. For any $r \in \mathbb{R}$ defines $\widehat{\mathrm{EB}}_r^b(c) = \sum_{i=1}^N (\hat{\Delta}_i^b - c)_+/N - (\sum_{i=1}^N \hat{\Delta}_i^b/N - c) \times I(r>0)$, $\widehat{\mathrm{PEB}}_r^{\mathrm{l}b}(c) = \sum_{i=1}^N (\hat{\Delta}_i^b)_+ \times (1-c) - (\sum_{i=1}^N \hat{\Delta}_i^b/N - c)_+ \times I(r>0)$, and $\widehat{\mathrm{PEB}}_r^{\mathrm{u}b}(c) = \sum_{i=1}^N \{\widehat{\mathrm{Risk}}_{0i}^b - (\widehat{\mathrm{Risk}}_{0i}^b + \widehat{\mathrm{Risk}}_{1i}^b - 1)_+\} \times (1-c)/N - (\sum_{i=1}^N \hat{\Delta}_i^b/N - c) \times I(r>0)$. Let $\zeta_{\mathrm{EB}(c),\eta}(r)$, $\zeta_{\mathrm{PEB}^l(c),\eta}(r)$, and $\zeta_{\mathrm{PEB}^u(c),\eta}(r)$ denote $(1-\eta) \times 100\%$ percentile bootstrap CIs formed by taking empirical percentiles of $\widehat{\mathrm{EB}}_r^b(c)$, $\widehat{\mathrm{PEB}}_r^{lb}(c)$, and $\widehat{\mathrm{PEB}}_r^{ub}(c)$ over bootstrap samples, respectively. Let $\Gamma_\alpha(c)$ denote an asymptotically valid $(1-\alpha) \times 100\%$ CI for

$\rho_0 - \rho_1 - c$. The $(1-\eta-\alpha)\times 100\%$ projection intervals for $\mathrm{EB}(c)$, $\mathrm{PEB}^l(c)$, and $\mathrm{PEB}^u(c)$ are given by $\bigcup_{r\in\Gamma_\alpha(c)}\zeta_{\mathrm{EB}(c),\eta}(r)$, $\bigcup_{r\in\Gamma_\alpha(c)}\zeta_{\mathrm{PEB}^l(c),\eta}(r)$, and $\bigcup_{r\in\Gamma_\alpha(c)}\zeta_{\mathrm{PEB}^u(c),\eta}(r)$, respectively. Let P^b denotes probability taken with respect to the bootstrap sampling algorithm, conditional on the observed data. The following Theorem 2 is proved in supplementary material available at *Biostatistics* online, Appendix F.

THEOREM 2 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 $|\hat\rho_0-\hat\rho_1-c|\leqslant\tau_N$ and $\{\sum_{i=1}^N \hat\Delta_i/N - c\}$ otherwise. The $(1-\eta-\alpha)\times 100\%$ adaptive bootstrap CIs for $\mathrm{EB}(c)$, $\mathrm{PEB}^l(c)$, and $\mathrm{PEB}^u(c)$ based on pre-specified τ_N are given by $\bigcup_{r\in\mathfrak{A}(c)^c}\zeta_{\mathrm{EB}(c),\eta}(r)$, $\bigcup_{r\in\mathfrak{A}(c)}\zeta_{\mathrm{PEB}^l(c),\eta}(r)$, and $\bigcup_{r\in\mathfrak{A}(c)}\zeta_{\mathrm{PEB}^u(c),\eta}(r)$, respectively. Assume $\Delta(Y)$ has a continuous and bounded density function. For $\alpha,\eta\in(0,1)$, we have

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

If $E\Delta(Y)\neq c$, then the right-hand side of the foregoing inequalities can be replaced with equalities.

REMARK 1 Berger and Boos (1994) recommend choosing α to quite small in which case $1-\eta\approx 1-\eta-\alpha$. Consequently, the proposed projection CI is nearly exact in large samples provided $E\{\Delta(Y)\}\neq c$, but potentially conservative otherwise. However, Theorem 2 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 supplementary material available at *Biostatistics* online, Appendix G, we provide a locally consistent projection confidence interval that does not require $\Delta(Y)$ to 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.

4. SIMULATION STUDIES

Consider a two-arm 1 : 1 randomized trial where a biomarker Y following a standard normal distribution is measured. Suppose the risk of a binary disease D conditional on Y and T follows a linear logistic model: $\mathrm{logit}\{P(D=1|Y,T)\}=-0.158+3.385T-0.5Y-4YT$, with disease prevalences $\rho_0=0.25$ and $\rho_1=0.125$. We consider cost ratios $c=$0, 0.105, 0.125, 0.145, and 0.175, which correspond to expected benefit values of 0.043, 0.059, 0.063, 0.048, and 0.029. The pairs of lower and upper bounds for expected benefit from perfect treatment selection are {0.043, 0.098}, {0.130, 0.180}, {0.147, 0.196}, {0.144, 0.191}, and {0.139, 0.184}, respectively.

Tables 1 and 2 show performance of the model-based estimators for $\mathrm{EB}(c)$, $\mathrm{PEB}(c)$, and $\mathrm{SEB}(c)$ for sample sizes 200, 500, and 2000 based on 5000 simulations and 1000 bootstrap samples. At $N=200$, model-based estimators have minimal bias for each measure. Coverage of 95% percentile bootstrap CI is close to the nominal level when c is away from $\rho_0-\rho_1$, whereas undercoverage is observed when $c=\rho_0-\rho_1$, which is not alleviated with an increase in sample size. The adaptive bootstrap CI fixes the undercoverage where we adopt the projection interval (with $\alpha=0.01$) when $\rho_0-\rho_1$ is close to c (defined as $|\hat\rho_0-\hat\rho_1-c|\leqslant\max\{N^{0.05},\hat{SE}(\hat\rho_0-\hat\rho_1)\times\Phi^{-1}(0.95)\}$ in the simulation study and data example).

Table 3 presents performance of various robust estimators of treatment-selection rules and robust estimators of expected benefit. These simulations explore performance of various treatment-selection rules

Table 1. *Performance of the model-based estimator of 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
			SD×$\sqrt{N}$		
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
			SE×$\sqrt{N}$		
200	0.28	0.27	0.27	0.27	0.26
500	0.29	0.29	0.30	0.32	0.32
2000	0.29	0.31	0.32	0.39	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	95.22	97.04	95.56	96.48	97.02
500	95.12	97.42	95.88	96.82	95.98
2000	94.62	96.68	95.88	96.72	95.78

under the correctly specified risk model, so that one can examine the penalty associated with the use of the robust methods in terms of increased variability and suboptimal expected benefit. For each rule derived from a simulated dataset, expected benefit in the population is computed through numerical integration. With $A(Y) = I\{\hat{\Delta}(Y) > \delta\}$, estimating the optimal δ rather than using $\delta = c$ leads to smaller expected benefit and larger variability for the small sample size of 200, but the difference is minimal when the sample size is as large as 2000. Performance of the treatment-selection rule derived by minimizing weighted classification errors through a grid search is comparable with that based on $\hat{\Delta}(Y)$ and estimated δ. In general, using the augmented version for δ estimation or for weight computation leads to a small increase in expected benefit and decreased variability. Between the two algorithms for weighted classification, the grid search in general does better than weighted logistic regression, especially for large cost ratio; the latter has performance worse than the optimal strategy absent the marker at cost ratio 0.175. Note that the advantage of the robust approaches is expected to become apparent when the model of risk difference is misspecified (an example is presented in supplementary material available at *Biostatistics* online, Table S3).

In Table 3, we also compare naive estimates of expected benefit using the same data where a treatment-selection rule is derived, and estimates based on random cross-validation. For the latter, we randomly split the data into 2/3 training and 1/3 test, estimate a treatment-selection rule from the training set, then compute the expected benefit for this rule using the test set; an average of expected benefit estimates is computed over 500 splits. From Table 3, we see that naive estimates of expected benefit can have severe overestimation even with a sample size as large as 2000, for all estimators. An exception is the model-based

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
					$SD \times \sqrt{N}$					
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
					$SE \times \sqrt{N}$					
200	0.32	0.31	0.32	0.34	0.38	0.28	0.27	0.29	0.32	0.37
500	0.33	0.31	0.32	0.37	0.42	0.29	0.28	0.29	0.36	0.42
2000	0.33	0.32	0.33	0.43	0.45	0.29	0.29	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	94.94	95.98	94.54	94.98	95.86	95.22	96.26	94.22	95.72	96.62
500	94.52	96.08	94.34	95.42	95.50	95.12	97.04	94.86	96.56	95.64
2000	94.82	96.64	95.06	96.42	95.18	94.62	96.74	95.14	96.72	95.58
			$SEB^l(c)$					$SEB^u(c)$		
	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
					$SD \times \sqrt{N}$					
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
					$SE \times \sqrt{N}$					
200	1.92	1.26	1.25	1.25	1.22	0.00	1.50	1.51	1.53	1.49
500	2.06	1.31	1.31	1.45	1.49	0.00	1.48	1.50	1.70	1.82
2000	2.15	1.34	1.31	1.66	1.70	0.00	1.45	1.44	1.90	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

continued.

Table 2. *continued.*

Cost ratio c	0.000	0.105	0.125	0.145	0.175	0.000	0.105	0.125	0.145	0.175
					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

estimator whose overfitting bias is minimal for sample sizes >500. The overfitting bias is corrected by cross-validation.

Bootstrap standard errors and CI coverage for robust estimates of EB are presented in supplementary material available at *Biostatistics* online, Table 1. Undercoverage of ordinary bootstrap CIs happens in some cases when the cost ratio equals $\rho_0 - \rho_1$; the adaptive bootstrap CI alleviates the problem.

5. DATA EXAMPLE

In this section, we illustrate the approaches using the Diabetes Control and Complications Trial (DCCT) (DCCTRG, 1993), 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 early vascular and neurologic complications of diabetes. Overall, 1441 patients with insulin-dependent diabetes mellitus were enrolled beginning in 1983 and followed through 1999. One outcome significantly impacted by intensive therapy in DCCT is microalbuminuria, a sign of kidney damage, defined as albumin excretion rate >40 mg/24 h. Our analysis here consists of 579 subjects in the secondary prevention cohort of DCCT, defined as patients with mild preexisting retinopathy or other complications, who did not have microalbuminuria and neuropathy at baseline. We consider baseline hemoglobin A1C (HBA1C) as a biomarker for selecting treatment: a linear logistic regression model of microalbuminuria developed during the study versus treatment and baseline HBA1C and their interaction shows a significant interaction between treatment and HBA1C.

We estimate the (standardized) expected benefit of HBA1C. The curve of model-based estimator of EB(c) versus c is presented in Figure 1(a), together with estimated lower and upper bounds of PEB. Corresponding bounds for standardized expected benefit of HBA1C are displayed in Figure 1(b). For a set of chosen cost ratios, the model-based estimates and their 95% CI are shown in Table 4. For example, at cost ratio $c = 0$, i.e. equal treatment cost between the two diabetes therapies, HBA1C has an EB of 0.005, while the PEB can range from 0.005 to 0.206, implying that standardized EB of HBA1C is $>2.3\%$. If $c = 0.05$, i.e. the additional cost by intensive therapy compared with conventional therapy is 5% the cost of developing microalbuminuria, HB1AC has an EB of 0.019, which corresponds to 8–38.8% of PEB. Supplementary material available at *Biostatistics* online, Table 2 presents cross-validated EB for the model-based estimator and robust estimators. In general, we see a reduction in EB resulted from CV. Treatment-selection rules based on $\hat{\Delta}(Y)$ and the estimated threshold δ or based on linear marker combinations that minimize weighted classification errors using augmented weights have slightly better CV estimates of EB compared with the model-based estimator.

6. DISCUSSION

We developed an expected benefit measure for characterizing the capacity of biomarkers for treatment selection, and developed the concept of a perfect treatment-selection rule that correctly identifies subjects

Table 3. *MEAN* (*SD*) *of expected benefit of a derived treatment-selection rule in the population and MEAN*(*SD*) *of naive and cross-validated estimate of corresponding expected benefit*

N	TYPE	PAR⋆	NPAR1⋆	NPAR2⋆	WLOGIS1⋆	WLOGIS2⋆	WCLASS1⋆	WCLASS2⋆
				Cost ratio $c = 0$				
200	True	0.0406 (0.0035)	0.0337 (0.011)	0.0363 (0.0084)	0.0329 (0.0105)	0.0362 (0.0085)	0.0335 (0.0122)	0.0377 (0.0072)
	Naive	0.0454 (0.0208)	0.0586 (0.0299)	0.0552 (0.0298)	0.0404 (0.0368)	0.0432 (0.0339)	0.0602 (0.0331)	0.0559 (0.033)
	CV	0.0391 (0.0325)	0.0303 (0.0329)	0.0328 (0.031)	0.0316 (0.036)	0.0336 (0.0336)	0.0307 (0.0381)	0.0358 (0.0351)
500	True	0.042 (0.0012)	0.0385 (0.006)	0.0398 (0.004)	0.0345 (0.0069)	0.0401 (0.004)	0.038 (0.0071)	0.04 (0.0038)
	Naive	0.0439 (0.0133)	0.0513 (0.0198)	0.0499 (0.019)	0.0374 (0.0244)	0.0426 (0.0218)	0.0524 (0.0218)	0.0499 (0.0211)
	CV	0.041 (0.0209)	0.0364 (0.022)	0.0381 (0.0198)	0.0336 (0.0239)	0.0382 (0.0217)	0.0357 (0.0251)	0.0383 (0.0222)
2000	True	0.0427 (3e−04)	0.0415 (0.002)	0.0418 (0.0015)	0.0359 (0.0034)	0.0422 (0.001)	0.0412 (0.0025)	0.0417 (0.0016)
	Naive	0.0433 (0.0065)	0.0468 (0.0102)	0.0463 (0.0095)	0.0368 (0.0125)	0.0432 (0.0109)	0.0472 (0.0114)	0.0462 (0.0106)
	CV	0.0428 (0.0105)	0.0413 (0.0108)	0.0417 (0.0099)	0.0359 (0.0124)	0.0421 (0.0109)	0.0409 (0.0123)	0.0416 (0.0111)
				Cost ratio $c = 0.105$				
200	True	0.0542 (0.0067)	0.0474 (0.0136)	0.049 (0.0119)	0.0366 (0.0221)	0.0403 (0.0173)	0.0472 (0.0181)	0.0515 (0.0119)
	Naive	0.0518 (0.0194)	0.0653 (0.0273)	0.0629 (0.0262)	0.0398 (0.0322)	0.0374 (0.0298)	0.064 (0.0294)	0.0598 (0.0287)
	CV	0.0379 (0.0306)	0.0299 (0.031)	0.0318 (0.0287)	0.0178 (0.0344)	0.0217 (0.0314)	0.0287 (0.0383)	0.0343 (0.0333)
500	True	0.0572 (0.002)	0.0536 (0.0066)	0.0543 (0.0055)	0.0477 (0.0127)	0.0471 (0.0092)	0.0538 (0.007)	0.0553 (0.0045)
	Naive	0.0559 (0.0127)	0.0633 (0.0185)	0.0623 (0.0175)	0.0474 (0.0224)	0.0442 (0.0188)	0.0627 (0.0202)	0.0606 (0.019)
	CV	0.0501 (0.0197)	0.0454 (0.0206)	0.0464 (0.0186)	0.0373 (0.0241)	0.0387 (0.0193)	0.0455 (0.0238)	0.0478 (0.0205)
2000	True	0.0585 (5e−04)	0.0571 (0.0022)	0.0573 (0.002)	0.0543 (0.0044)	0.0487 (0.0049)	0.0569 (0.0024)	0.0573 (0.0017)
	Naive	0.0588 (0.0069)	0.0622 (0.0103)	0.0619 (0.0096)	0.0547 (0.0106)	0.0488 (0.0097)	0.0618 (0.0114)	0.061 (0.0104)
	CV	0.0575 (0.0106)	0.0557 (0.0109)	0.056 (0.0099)	0.0524 (0.0111)	0.0478 (0.0096)	0.0555 (0.0122)	0.0561 (0.0109)
				Cost ratio $c = 0.125$				
200	True	0.0577 (0.0072)	0.0514 (0.0136)	0.0529 (0.0119)	0.0356 (0.0233)	0.0397 (0.0185)	0.0489 (0.0203)	0.0535 (0.0143)
	Naive	0.0485 (0.0197)	0.0625 (0.0265)	0.0605 (0.0255)	0.0318 (0.0309)	0.0293 (0.0294)	0.0595 (0.0284)	0.0552 (0.0278)
	CV	0.0327 (0.0298)	0.0257 (0.0302)	0.0275 (0.0282)	0.009 (0.0328)	0.0131 (0.0306)	0.0214 (0.0379)	0.0275 (0.0334)
500	True	0.0611 (0.0022)	0.0577 (0.0067)	0.0583 (0.0056)	0.0465 (0.0148)	0.0457 (0.011)	0.0569 (0.0079)	0.0584 (0.0051)
	Naive	0.0531 (0.013)	0.0608 (0.0177)	0.06 (0.0169)	0.0389 (0.0224)	0.0355 (0.02)	0.0587 (0.0194)	0.0567 (0.0183)
	CV	0.0463 (0.0191)	0.0419 (0.0199)	0.0428 (0.0186)	0.0285 (0.0238)	0.03 (0.0201)	0.0406 (0.0235)	0.0432 (0.0204)
2000	True	0.0626 (6e−04)	0.0611 (0.0024)	0.0613 (0.0021)	0.052 (0.0069)	0.047 (0.0057)	0.0602 (0.0024)	0.0606 (0.0018)
	Naive	0.0575 (0.0068)	0.0607 (0.0094)	0.0604 (0.0089)	0.0464 (0.0114)	0.0411 (0.0105)	0.0592 (0.0102)	0.0585 (0.0095)
	CV	0.0555 (0.0096)	0.0537 (0.01)	0.0539 (0.0094)	0.0445 (0.0114)	0.0401 (0.0104)	0.0528 (0.011)	0.0534 (0.01)

continued.

Table 3. *continued.*

N	TYPE	PAR⋆	NPAR1⋆	NPAR2⋆	WLOGIS1⋆	WLOGIS2⋆	WCLASS1⋆	WCLASS2⋆
					Cost ratio $c = 0.145$			
200	True	0.042 (0.008)	0.0368 (0.0127)	0.038 (0.0111)	0.0164 (0.0221)	0.0205 (0.0171)	0.0309 (0.021)	0.0356 (0.0152)
	Naive	0.0429 (0.0205)	0.058 (0.0261)	0.0562 (0.0254)	0.0227 (0.0293)	0.0199 (0.0287)	0.0529 (0.0277)	0.0484 (0.0272)
	CV	0.0255 (0.0293)	0.0199 (0.0294)	0.0217 (0.028)	−3e−04 (0.0301)	0.0038 (0.0286)	0.0121 (0.0372)	0.018 (0.0337)
500	True	0.0459 (0.0025)	0.0426 (0.0071)	0.0432 (0.0061)	0.0246 (0.0151)	0.0245 (0.0113)	0.0399 (0.0092)	0.0415 (0.0063)
	Naive	0.0462 (0.0148)	0.0549 (0.018)	0.0542 (0.0176)	0.0253 (0.0236)	0.0227 (0.0218)	0.0504 (0.0199)	0.0485 (0.019)
	CV	0.0389 (0.0195)	0.0348 (0.0202)	0.0357 (0.0194)	0.0155 (0.0235)	0.0174 (0.0211)	0.0307 (0.0248)	0.0336 (0.0221)
2000	True	0.0476 (7e−04)	0.0463 (0.0024)	0.0464 (0.0022)	0.0275 (0.0081)	0.0251 (0.0059)	0.0437 (0.0024)	0.0441 (0.0019)
	Naive	0.048 (0.009)	0.0519 (0.0104)	0.0516 (0.0104)	0.0277 (0.0147)	0.025 (0.0128)	0.0485 (0.0112)	0.0477 (0.0109)
	CV	0.0462 (0.0107)	0.0444 (0.011)	0.0446 (0.011)	0.0262 (0.0144)	0.024 (0.0127)	0.042 (0.0119)	0.0425 (0.0114)
					Cost ratio $c = 0.175$			
200	True	0.0214 (0.0075)	0.0179 (0.0106)	0.0191 (0.009)	−0.0044 (0.0183)	−8e−04 (0.012)	0.0054 (0.0194)	0.01 (0.0139)
	Naive	0.0318 (0.0209)	0.0489 (0.0257)	0.0474 (0.0254)	0.0094 (0.0247)	0.007 (0.0242)	0.0404 (0.0261)	0.0354 (0.0259)
	CV	0.013 (0.0274)	0.0097 (0.0275)	0.0113 (0.0272)	−0.0117 (0.0242)	−0.0079 (0.0231)	−0.004 (0.0341)	0.0014 (0.0318)
500	True	0.0257 (0.0035)	0.0227 (0.007)	0.023 (0.0067)	−0.0011 (0.0094)	−8e−04 (0.0065)	0.0135 (0.0107)	0.0154 (0.0077)
	Naive	0.0313 (0.0158)	0.0417 (0.0184)	0.0412 (0.0184)	0.0052 (0.0191)	0.0032 (0.0183)	0.0326 (0.0201)	0.0305 (0.0197)
	CV	0.0234 (0.0198)	0.0204 (0.0202)	0.0211 (0.0201)	−0.0029 (0.0177)	−0.0015 (0.0172)	0.0093 (0.0254)	0.0124 (0.0237)
2000	True	0.028 (8e−04)	0.0266 (0.0032)	0.0266 (0.0031)	−0.003 (0.0041)	−0.0024 (0.0035)	0.0189 (0.0038)	0.0193 (0.0034)
	Naive	0.0293 (0.0085)	0.0339 (0.0103)	0.0337 (0.0104)	−0.0014 (0.0107)	−0.0012 (0.0103)	0.0254 (0.0116)	0.0246 (0.0114)
	CV	0.0274 (0.0106)	0.0257 (0.0111)	0.0258 (0.0112)	−0.0025 (0.0101)	−0.0022 (0.01)	0.0177 (0.0133)	0.0183 (0.0129)

PAR⋆: corresponds to the rule $A(Y) = I(\hat{\Delta} > c)$, where $\hat{\Delta}$ is the estimated risk difference based on the GLM risk model

NPAR1⋆, NPAR2⋆: correspond to $A(Y) = I(\hat{\Delta} > \delta)$ with δ chosen to maximize the empirical or augmented estimate of EB;

WLOGIS1⋆, WLOGIS2⋆: correspond to rule $A(Y) = I(\alpha_0 + \alpha_1 Y > 0)$, where α_0 and α_1 are estimated based on converting the problem to a weighted classification problem, which is solved using weighted logistic regression with empirical weight or augmented weight, respectively;

WCLASS1⋆, WCLASS2⋆: correspond to rule $A(Y) = I(\alpha_0 + \alpha_1 Y > 0)$, where α_0 and α_1 are estimated based on converting the problem to a weighted classification problem, which is solved using a grid search with empirical weight or augmented weight, respectively;

True⋆: indicates population performance of a treatment-selection rule derived from a training data.

Table 4. *Estimate and* 95% *adaptive CI of model-based estimator 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)
$\text{PEB}^{\text{l}}(c)$	0.005 (0, 0.166)	0.05 (0.031, 0.157)	0.086 (0.029, 0.149)	0.084 (0.029, 0.149)
$\text{PEB}^{\text{u}}(c)$	0.206 (0.157, 0.352)	0.242 (0.192, 0.335)	0.267 (0.216, 0.329)	0.261 (0.211, 0.343)
$\text{SEB}^{\text{l}}(c)$	0.023 (0, 0.498)	0.08 (0, 0.382)	0.131 (0.003, 0.334)	0.107 (0.001, 0.366)
$\text{SEB}^{\text{u}}(c)$	1 (1, 1)	0.388 (0, 0.802)	0.408 (0.026, 0.809)	0.333 (0.005, 0.823)

Note: the adaptive CI was constructed using $\alpha = 0.01$; the optimal treatment-selection rule at cost ratio c is to recommend intensive treatment if $\{1 + \exp(8.779 - 3.589 \times \log(\text{HBA1C})\}^{-1} - \{1 + \exp(2.942 - 0.727 \times \log(\text{HBA1C})\}^{-1} > \text{c}$.

who will benefit from treatment. Expected benefit of the latter is in general not identifiable and we developed bounds for it based on disease risk model. The idea of generating bounds can be readily applied to other summary measures such as the reduction in the population disease rate under marker-based treatment (Song and Pepe, 2004). An interesting observation regarding the model-based bounds is that their width depends on how well the risk model used to construct the bounds can identify the percent of subjects "benefited" by treatment. A model that better predicts heterogeneity in treatment benefit in terms of larger variability in $\Delta(Y)$ tends to move up the lower bound for PEB by increasing $E\{\Delta(Y)\}_+$. When we have several risk models in a population, e.g. risk given Y_1 and risk given Y_1 and Y_2, tighter bounds can be constructed combining bounds derived from individual risk models. Specifically, at a given cost ratio, the lower bound of PEB can be constructed as the maximum among individual lower bounds, and the upper bound can be constructed as the minimum among individual upper bounds. However, the variance may be difficult to calculate in this case and will require further investigation.

We developed a GLM model-based approach for deriving treatment-selection rules to maximize the expected benefit. In addition, we considered robust approaches that combine the risk difference from the GLM model and an estimated threshold, or that find marker combinations that directly maximize the estimate of expected benefit. The latter can be computationally intensive if a grid search is used to identify model parameters; the GLM model, in contrast, can be easily implemented with standard statistical software. The model-based approach can lead to treatment-selection rules with better performance as well as increased efficiency in estimating the expected benefit when the risk model is correctly specified, while the robust approaches are less affected by misspecification of the working risk model and can be used for sensitivity analysis.

For inference about expected benefit, we proposed an adaptive bootstrap procedure to handle non-regularity when the cost ratio is close to the average treatment effect. This idea of using data to adaptively construct a bootstrap CI has 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 the treatment–disease cost ratio c to be constant in the population, and vary c using a sensitivity analysis. In practice, one needs to put the burdens of disease and treatment on the same scale to determine the value of c. For example, Gail (2009) evaluated the benefit of using the Gail model for recommending tamoxifen for breast cancer prevention. Tamoxifen has been shown to increase the risk of "secondary" events such as stroke and endometrial cancer. Gail (2009) assumed that burden per secondary event is the same as the burden per breast cancer event, and chose c to be the increased rate of having any secondary event due to tamoxifen. Alternatively, if one can associate a monetary cost with each disease event, and a monetary cost with treatment (potentially including the cost of the treatment itself and the cost due to secondary events), then c can be computed as the rate of the latter relative to that of the former. In some settings, the cost ratio might be a function of the 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 to allow $c = C(Y)$ to be a function of the biomarker 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 estimation and inference methods developed here can be readily generalized to handle continuous outcomes.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

This work was supported by NIH grant R01 GM106177-01, P01 CA142538, R01 CA152089 and U01 CA086368.

REFERENCES

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.

BERGER, R. L. AND BOOS, D. 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, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* **12**(2), 270–282.

DCCTRG. (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, G. L., IYER, H. K. AND ALBERT, J. 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.

HUANG, Y., GILBERT, P. B. AND JANES, H. (2012). Assessing treatment-selection markers using a potential outcomes framework. *Biometrics* **68**(3), 687–696.

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.

JANES, H., BROWN, M. D., PEPE, M. S. AND HUANG, Y. (2014). Statistical methods for evaluating and comparing biomarkers for patient treatment selection. *International Journal of Biostatistics*, in Press.

JANES, H., PEPE, M. S. AND HUANG, Y. (2013). A framework for evaluating markers used to select patient treatment. *Medical Decision Making* **34**(2), 159–167.

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

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, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*. New York: 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, B., Tsiatis, A. A., Davidian, M., Zhang, M. and Laber, E. B. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat* **1**(1), 103–114.

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

Zhang, Z., Wang, C., Nie, L. and Soon, G. (2013). Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**(5), 687–704.

Zhao, L., Tian, L., Cai, T., Claggett, B. and Wei, L. J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association* **108**(502), 527–539.

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

[*Received October 22, 2013; revised July 18, 2014; accepted for publication July 23, 2014*]