Published in final edited form as:

Biometrics. 2012 September ; 68(3): 687–696. doi:10.1111/j.1541-0420.2011.01722.x.

Assessing Treatment-Selection Markers using a Potential Outcomes Framework

Ying Huang^{1,*}, Peter B. Gilbert^{1,2}, and Holly Janes^{1,2}

¹Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109, U.S.A.

²Department of Biostatistics, University of Washington, Seattle, Washington, 98109, U.S.A.

Summary

Treatment-selection markers are biological molecules or patient characteristics associated with one's response to treatment. They can be used to predict treatment effects for individual subjects and subsequently help deliver treatment to those most likely to benefit from it. Statistical tools are needed to evaluate a marker's capacity to help with treatment selection. The commonly adopted criterion for a good treatment-selection marker has been the interaction between marker and treatment. While a strong interaction is important, it is, however, not sufficient for good marker performance. In this paper, we develop novel measures for assessing a continuous treatment-selection marker, based on a potential outcomes framework. Under a set of assumptions, we derive the optimal decision rule based on the marker to classify individuals according to treatment benefit, and characterize the marker's performance using the corresponding classification accuracy as well as the overall distribution of the classifier. We develop a constrained maximum-likelihood method for estimation and testing in a randomized trial setting. Simulation studies are conducted to demonstrate the performance of our methods. Finally, we illustrate the methods using an HIV vaccine trial where we explore the value of the level of pre-existing immunity to Adenovirus serotype 5 for predicting a vaccine-induced increase in the risk of HIV acquisition.

Keywords

Classification accuracy; Constrained maximum likelihood; Monotone treatment effect; Potential outcomes; Sensitivity analysis; Treatment-selection marker

1. Introduction

In many clinical contexts today, the hope is to improve the effectiveness of interventions by targeting therapy to those most likely to benefit. Some therapies are thought to only benefit a subset of individuals, while others are thought to be particularly harmful for some subjects. Markers that are predictive of the effect of one treatment over another, termed treatment-selection markers, can potentially be used to deliver a treatment only to individuals likely to benefit, and to avoid the burden, cost, and potential toxic effects of providing it to others. These markers may be biological measures, other patient or disease characteristics, or combinations of the above. Well-known examples of existing markers are K-RAS gene expression for selecting against anti-epidermal growth factor receptor treatment of colorectal

cancer (Karapetis et al., 2008; Allegra et al., 2009); estrogen-receptor (ER) positivity for selecting endocrine therapy for breast cancer treatment (Harris et al., 2007); and Oncotype-DX, a multi-gene score for selecting adjuvant chemotherapy for the treatment of ER-positive breast cancer (Paik et al., 2006; Harris et al., 2007; Albain et al., 2010).

A fundamental question in the development of treatment-selection markers is what statistical measures should be used for marker evaluation. Answering this question is essential for understanding the clinical utility of a given marker and for comparing candidate markers, as well as for performing more complex tasks such as developing marker combinations, assessing the incremental value of a marker, and evaluating covariate effects. Most of the methodologic literature on treatment-selection markers assumes that the statistical interaction between marker value and treatment assignment in the context of a randomized trial is the primary measure of marker performance (eg Sargent et al. (2005); Freidlin and Simon (2005); Buyse (2007); Simon (2008); Simon et al. (2009)). However, a strong interaction is important but not sufficient for adequate marker performance (Janes et al., 2011). Specifically, two markers can have the same interaction but very different performance. In Web supplementary figure 1, we present an example where two markers have the same interaction coefficient but different capacity in terms of classifying a subject according to treatment effectiveness. Details on our approach to assessing a marker's treatment-selection capacity will be presented in Section 2.

Additional limitations of the interaction approach include the fact that the scale of the interaction coefficient depends on the functional form of the model relating marker and treatment to outcome. The interaction approach also does not facilitate more sophisticated tasks such as deriving marker combinations.

In this paper, our goal is to develop a measure for assessing the treatment-selection performance of a marker (or risk model) which puts different markers (models) on the same scale to facilitate their comparison. We focus on the setting of a randomized trial, and use a potential outcomes framework to develop novel measures of the performance of a continuous treatment-selection marker. While most of the existing methodologic literature focuses on binary markers (Mandrekar et al., 2005; Lijmer and Bossuyt, 2009; Wang and Zhou, 2010), commonly markers are continuous. Under a set of assumptions, we derive the optimal decision rule for using the marker to classify individuals according to the benefit of one treatment over another. We then characterize the accuracy with which this classifier distinguishes between individuals who do and do not benefit, as well as the overall distribution of the classifier. We develop a constrained maximum-likelihood method for estimation and testing. Simulation studies are conducted to demonstrate the finite-sample performance of the estimators. Finally, we illustrate the methods using an HIV vaccine trial where we explore the value of the level of pre-existing immunity to Adenovirus serotype 5 for identifying individuals with a vaccine-induced increase in the risk of HIV acquisition.

2. Methods

We consider a study setting where individuals are randomly assigned to a placebo arm or an active treatment arm. Let S denote the continuous candidate baseline biomarker of interest, which could be a vector. Let Z be a binary treatment indicator, with values 0 and 1 indicating placebo and active treatment arm respectively. Let Y be the binary disease outcome of interest, with value 0 and 1 indicating nondiseased and diseased respectively. Without loss of generality (WLOG), suppose every subject has S and Y measured. Moreover, we adopt a potential outcomes framework and use $Y(0)$ and $Y(1)$ to indicate the potential outcome for Y if an individual is assigned to placebo or active treatment respectively.

2.1 Assumptions for Identifiability and Estimation

We make the following assumptions required for identifying and estimating the measures that will be defined later:

(A1) The stable unit treatment value and consistency assumption (SUTVA): $Y(1), Y(0)$ for a subject are independent of the treatment assignments of other subjects, and a subject's potential outcome $Y(z)$ under the assigned treatment $Z = z$ equals the observed outcome Y

(A2) The ignorable treatment assignments assumption: $Y(0), Y(1), S \perp Z$

(A3) The monotone treatment effect assumption: treatment either won't do any harm or won't have any benefit. WLOG here we assume $Y(0) \leq Y(1)$. That is, treatment will not do any harm to an individual where this particular clinical outcome is concerned

(A4) We assume the probability of $Y(Z)$ given S and Z follows a generalized linear structural model: $Risk_{(Z)}(S) = P\{Y(Z) = 1|S\} = G(S, Z, \theta)$.

A1 is commonly adopted in causal inference literature, and is plausible when participants do not interact with one another in a trial. A2 is satisfied by randomization. The risk model assumption A4 provides interpretable parameters for inference and can be tested using the study data with standard statistical tools. For example, a Hosmer-Lemeshow test can be performed to compare the expected and observed risk quantiles. More powerful empirical likelihood based techniques can also be utilized (Qin and Zhang, 1997; Gilbert, 2003). The monotonicity assumption A3 is the key for the definition and identifiability of our classification accuracy measure defined next. Although it cannot be confirmed, it can be rejected based on statistical tests in a randomized trial. Throughout this manuscript, A1, A2 and A4 are assumed; we assume A3 in developing our framework and estimation methods and relax it later in section 2.4 in developing a sensitivity analysis method.

2.2 Measures of Treatment-Selection Capacity

Let $D = Y(0) - Y(1)$ denote the individual treatment effect on the disease outcome. Under the monotone treatment effect assumption A4, D takes values 0 or 1. For those subjects with $D = 0$, $Y(0) = Y(1)$, having the active treatment or not does not change their disease outcome, whereas for those with $D = 1$, $Y(0) > Y(1)$, the active treatment prevents the disease. Therefore, we can divide individuals into a treatment-effective group ($D = 1$) and a treatment-ineffective group ($D = 0$).

We characterize the marker's treatment-selection ability in two ways. First, researchers will be interested in the discriminatory power of a marker in terms of correctly identifying an individual's treatment-effectiveness status. Suppose in practice a decision rule is constructed based on $Q(S)$, a function of S , with larger value of $Q(S)$ indicating better chance of benefitting from the active treatment. Given a particular threshold value q , suppose an individual is classified as treatment-effective if $Q(S) \geq q$, and treatment-ineffective if $Q(S) < q$. We define the true positive fraction as the probability of correctly identifying a treatment-effective individual, $TPF(q) = P\{Q(S) \geq q | D = 1\}$, and the false positive fraction as the probability of incorrectly classifying a treatment-ineffective individual to be treatment-effective, $FPF(q) = P\{Q(S) \geq q | D = 0\}$. An ROC curve for varying thresholds can be constructed using $ROC(t) = TPF\{FPF^{-1}(t)\}$.

While TPF and FPF tell us about the marker's underlying accuracy of classifying a subject according to treatment response, health policy makers may also be interested in gauging the population impact of the marker in helping make informed treatment decisions. This can be addressed by examining the distribution of $Q(S)$, namely the CDF of $Q(S)$ where $CDF_Q(q) = P\{Q(S) \leq q\}$. In practice suppose we adopt a low and a high thresholds q_L and q_H such that

the active treatment is believed to be warranted for an individual with $Q(S)$ exceeding the high threshold q_H and not warranted for one with $Q(S)$ below the low threshold q_L , while the best treatment is unclear if $Q(S)$ is in-between q_L and q_H . Then a marker that has large $\text{CDF}_{Q(q_L)}$ and $1 - \text{CDF}_{Q(q_H)}$ is preferred since it assigns more people into the definite high and low $Q(S)$ range where treatment decisions are clear. Note that while the classification accuracy measures are well-defined only under assumption A3, the distribution of $Q(S)$ is well-defined even without this assumption. However, incorporating A3 during estimation is advantageous for gaining efficiency when the assumption holds.

2.2.1 Classification Accuracy and Risk—For a related task of evaluating a diagnostic testing marker, the accuracy of correctly classifying a subject according to disease status is typically defined conditional on the observed disease status Y . In the treatment-selection problem, TPF and FPF defined in section 2.2 under assumption A3 are defined conditional on treatment-effectiveness status $D = Y(0) - Y(1)$, which is not completely observable. Hence these classification accuracy measures cannot be estimated directly conditional on D . However, under A3 they both can be represented as functions of disease risk:

$$\begin{aligned} \text{TPF}(q) &= \frac{P\{Y(0)-Y(1)=1, Q(S)>q\}}{P\{Y(0)-Y(1)=1\}} = \frac{E[\{Y(0)-Y(1)\}I\{Q(S)>q\}]}{E\{Y(0)-Y(1)\}} \\ &= \frac{E[E\{Y(0)-Y(1)|S\}I\{Q(S)>q\}]}{E\{Y(0)-Y(1)|S\}} = \frac{E[\{Risk_{(0)}(S)-Risk_{(1)}(S)\}I\{Q(S)>q\}]}{E[Risk_{(0)}(S)-Risk_{(1)}(S)]}, \\ \text{FPF}(q) &= \frac{P\{Y(0)-Y(1)=0, Q(S)>q\}}{P\{Y(0)-Y(1)=0\}} = \frac{E[\{1-Y(0)+Y(1)\}I\{Q(S)>q\}]}{E\{1-Y(0)+Y(1)\}} \\ &= \frac{E[E\{1-Y(0)+Y(1)|S\}I\{Q(S)>q\}]}{E\{1-Y(0)+Y(1)|S\}} = \frac{E[\{1-Risk_{(0)}(S)+Risk_{(1)}(S)\}I\{Q(S)>q\}]}{E[1-Risk_{(0)}(S)+Risk_{(1)}(S)]}. \end{aligned} \quad (1)$$

Therefore, as will be shown later, TPF and FPF can be calculated on the basis of a risk model estimated from the observed data.

2.2.2 Optimal choice of the decision rule—Having introduced measures of treatment-selection capacity using the decision rule $Q(S)$, we now consider the choice of $Q(S)$. Examples of $Q(S)$ could simply be S itself, the difference between $Risk_{(0)}(S)$ and $Risk_{(1)}(S)$, or the ratio of $Risk_{(1)}(S)$ over $Risk_{(0)}(S)$. Under A3, there does in fact exist an optimal choice in terms of maximizing classification accuracy. Let H_0 and H_1 be the point hypotheses that a subject is treatment-ineffective ($D=0$) and treatment-effective ($D=1$). According to the Neyman Pearson Lemma, the most powerful test rejecting H_0 favoring H_1 [i.e., the test with smallest type-II error ($1-\text{TPF}$) at a fixed type-I error rate (FPF)] is based on comparing the likelihood ratio $P(S|D=1)/P(S|D=0)$ with a threshold. By Bayes' theorem, the likelihood ratio

$$\frac{P(S|D=0)}{P(S|D=1)} = \frac{P(D=0|S)P(D=1)}{P(D=1|S)P(D=0)} = \frac{P(D=1|S)}{1-P(D=1|S)} \frac{P(D=1)}{P(D=0)},$$

is a monotone increasing function of $P(D=1|S)$. Therefore a decision rule comparing $P(D=1|S) = Risk_{(0)}(S) - Risk_{(1)}(S)$ to a threshold maximizes classification accuracy for identifying treatment effect. It follows that $P(D=1|S)$ has the highest ROC curve among all functions of S for classifying subjects according to their treatment-effectiveness.

An argument similar in flavor was made by McIntosh and Pepe (2002), where they pointed out that for classification with respect to a binary disease outcome, the risk of disease as a function of marker value is the optimal function in terms of maximizing the ROC curve at every point. In our setting, the risk difference between treatment arms, $Risk_{(0)}(S) - Risk_{(1)}(S)$, can be viewed as the 'risk' of the binary outcome $D=1$ under A3.

We therefore employ $Q(S) = Risk_{(0)}(S) - Risk_{(1)}(S)$ in simulation studies and in the data example. Next we investigate the identifiability of the performance measures and propose methods for estimation and testing.

2.3 Cumulative Link Model for the Distribution of Potential Outcomes

In this section, we demonstrate the identifiability of the distribution of $Y(0)$ and $Y(1)$ based on risk modeling, under assumption A3. Given $P\{Y(0) = 0, Y(1) = 1|S\} = 0$, there are only three possible combinations of $Y(0)$ and $Y(1)$. Let $C = k$ indicate the event that an individual's pair of potential outcomes falls into the k^{th} category as shown below, $k = 1, 2, 3$:

$$\begin{aligned} C=1: & Y(0)=1 \& Y(1)=1 \\ C=2: & Y(0)=1 \& Y(1)=0 \\ C=3: & Y(0)=0 \& Y(1)=0. \end{aligned} \quad (2)$$

For categorical outcomes, alternative models can be used for the distribution of C conditional on S . Here we adopt a cumulative link model because its parameters can be interpreted in terms of a generalized linear risk model, as will be shown next.

For $k = 1, 2$, suppose we model the cumulative probabilities of C conditional on S by

$$g\{P(C \leq k|S)\} = \alpha_k + \beta_k^T \gamma(S), \quad (3)$$

where g is a link function such as logit or normal CDF and $\gamma(S)$ is a vector of pre-specified functions of S such that $P(C \leq 2|S) - P(C \leq 1|S)$ holds for a practical range of S . We take $\gamma(S) = S$ later in the simulation studies and in the data example, but $\gamma(S)$ can be made flexible by including transformed versions of S such as polynomial terms and regression spline.

Note that subjects with $Y(1) = 1$ belong to category $C = 1$ and subjects with $Y(0) = 1$ belong to either category $C = 1$ or category $C = 2$. As a result, there is a one-to-one relationship between the risk model and the joint distribution of $Y(0)$ and $Y(1)$, here based on the cumulative link model. Specifically, a risk model including main effects for Z and $\gamma(S)$ and their interaction,

$$g\{P\{Y(Z)=1|S\}\} = \theta_0 + \theta_1 Z + \theta_2^T \gamma(S) + \theta_3^T \gamma(S) Z, \quad (4)$$

corresponds to a cumulative link model (3) with $\alpha_2 = \theta_0$, $\beta_2 = \theta_2$, $\alpha_1 = \theta_0 + \theta_1$, and $\beta_1 = \theta_2 + \theta_3$.

With a little abuse of notation, let $Risk(S, Z)$ denote the probability of disease conditional on S and the assigned treatment Z . Under (A2) the ignorable treatment assignment assumption, $Risk(S, Z) \equiv P\{Y=1|Z, S\} = P\{Y(Z)=1|S\} \equiv Risk_{(Z)}(S)$. Thus parameters in (3) can be uniquely identified based on observed data. Next we propose methods for estimating and comparing biomarkers' treatment-selection capacity. Estimation can be carried out in two steps, the estimation of parameters in model (3) and the estimation of measures of interest as defined in section 2.2.

2.3.1 Estimation of the Parameters in the Risk Model—As shown above, with the one-to-one correspondence between (3) and (4), applying an ordinary risk model such as a logistic regression model to the data will lead to consistent estimates for parameters in the cumulative link model. However, this procedure does not explicitly exploit the monotonicity

assumption A3 during estimation. Note that $P(C=1|S) - P(C=2|S) \Leftrightarrow Risk_{(0)}(S) - Risk_{(1)}(S)$ for every S value, which is not guaranteed by fitting of an ordinary risk model.

To address this issue, we propose to maximize a constrained observed likelihood. Suppose there are n_0 subjects in placebo arm and n_1 subjects in active treatment arm with $n = n_0 + n_1$. Let i be the subject indicator. We maximize

$$l = \log L = \sum_{i=1}^n Y_i \log \{P(Y_i=1|Z_i, S_i)\} + (1 - Y_i) \log \{P(Y_i=0|Z_i, S_i)\} \quad (5)$$

subject to the constraint that $Risk(S_i, Z=0) - Risk(S_i, Z=1)$ for $i = 1, \dots, n$. In other words, we ensure that the second category in (2), $\{Y(0)=1, Y(1)=0\}$, has a non-negative probability, even if this term does not appear explicitly in (5) as we cannot observe (or identify) any individual in this category. As we will show next, this constraint guarantees that the estimates of classification accuracy fall into a reasonable range.

2.3.2 Estimation of Measures of Treatment-Selecting Capacity—Let $\hat{\theta}$ be the constrained MLE for θ in risk model (4). We calculate estimates for $Risk_{(0)}$ and $Risk_{(1)}$ given each biomarker value observed in the sample as $\widehat{Risk}_{(0)}(S_i) = G(S_i, Z=0, \hat{\theta})$ and $\widehat{Risk}_{(1)}(S_i) = G(S_i, Z=1, \hat{\theta})$. Assumption A2 implies the equal distribution of marker within each treatment arm: $S|Z=0 \stackrel{d}{=} S|Z=1$. Thus $F(S)$, the common cumulative distribution function of S , can be estimated empirically with $\widehat{F}(s) = \sum_{i=1}^n I(S_i \leq s) / n$. Then we estimate TPF(q) and FPF(q) by entering $\widehat{Risk}_{(z)}$, $\widehat{F}$, and $\widehat{Q}(s)$ into (2):

$$\begin{aligned} \widehat{TPF}(q) &= \frac{\int \{\widehat{Risk}_{(0)}(s) - \widehat{Risk}_{(1)}(s)\} I\{\widehat{Q}(s) > q\} d\widehat{F}(s)}{\int \{\widehat{Risk}_{(0)}(s) - \widehat{Risk}_{(1)}(s)\} d\widehat{F}(s)} \\ \widehat{FPF}(q) &= \frac{\int \{1 - \widehat{Risk}_{(0)}(s) + \widehat{Risk}_{(1)}(s)\} I\{\widehat{Q}(s) > q\} d\widehat{F}(s)}{\int \{1 - \widehat{Risk}_{(0)}(s) + \widehat{Risk}_{(1)}(s)\} d\widehat{F}(s)}, \end{aligned}$$

where for $Q(S)$ being a pre-specified function of S , $\widehat{Q}(s)$ denotes the estimator of $Q(s)$ based on $\hat{\theta}$ when $Q(S)$ involves θ (e.g. when $Q(S) = Risk_{(0)}(S) - Risk_{(1)}(S)$), and $\widehat{Q}(s)$ equals $Q(s)$ otherwise (e.g. when $Q(S) \equiv S$). Observe that because of the constraints guaranteeing that $\widehat{Risk}_{(0)}(S_i) \geq \widehat{Risk}_{(1)}(S_i)$ for $i = 1, \dots, n$, $\widehat{TPF}(q)$ and $\widehat{FPF}(q)$ will always fall within (0,1), whereas estimates for TPF and FPF in (1) could go beyond (0,1) if an ordinary risk model were applied.

Finally, the ROC curve can be estimated by $\widehat{ROC}(t) = \widehat{TPF}\left\{\widehat{FPF}^{-1}(t)\right\}$ and the distribution of $Q(S)$ by $\widehat{CDF}_Q(q) = \int Q(s) d\widehat{F}(s)$.

Under standard regularity conditions, asymptotic normality of the constrained MLE for risk model coefficients and measures of treatment-selection performance can be established based on standard techniques (Web Supplementary Appendix B). However, in practical sample sizes, applying the constraints leads to a substantial drop in the variances, which are over-estimated by asymptotic formulas (results omitted). Instead we use bootstrap resampling for variance estimation and construction of confidence intervals.

Using estimates of the measures developed above such as a point on the ROC curve, area under the ROC curve (AUC), or the distribution of $Q(S)$, we can characterize and test for the treatment-selection performance of a marker (or risk model). For example, whether or not $AUC > 0.5$ in our setting corresponds to whether the biomarker distinguishes patients who benefit from the treatment versus patients who do not. A more relevant question in biomarker research is to compare treatment-selection performance between markers (or risk models), which can also be addressed by comparing these summary measures using a Wald-test, as will be demonstrated in the simulation studies.

2.4 Sensitivity Analysis for Departures from the Monotonicity Assumption

So far we have assumed the monotonicity assumption (A3), which is key to the definition and identification of TPF and FPF defined in section 2.2. In settings where A3 is likely to hold, e.g., when the outcome is specific to a particular disease such as HIV and the treatment cannot plausibly cause the disease, adopting this assumption simplifies the problem and allows efficiency gain in estimating CDF_Q by restricting the parameter space during estimation. There are other occasions, however, where A3 could be violated. For example, if the outcome of interest is survival, even if the treatment is effective at treating the disease of interest, it might increase the risk of other adverse events and therefore result in a nonzero probability of $Y(1) > Y(0)$. When this happens, $D = Y(0) - Y(1)$ takes three values, 1, 0, -1, which we define as the *Benefitted*, *Unaffected*, and *Harmed* groups respectively. We can define the probability of $Q(s) > q$ conditional on $D = 1, 0, -1$ as the true positive fraction (*Benefitted*), false positive fraction (*Unaffected*), and false positive fraction (*Harmed*) respectively. That is: $TPF_B(q) = P\{Q(S) > q | D = 1\}$, $FPF_U(q) = P\{Q(S) > q | D = 0\}$, and $FPF_H(q) = P\{Q(S) > q | D = -1\}$. Then two ROC curves can be constructed as

$ROC_U(t) = TPF_B\{FPF_U^{-1}(t)\}$ and $ROC_H(t) = TPF_B\{FPF_H^{-1}(t)\}$. Alternatively, we can categorize an individual according to whether $D = 1$ or $D < 1$. The quantities of interest would be the true positive fraction (*Benefitted*) $TPF_B(q)$ as defined before and the false positive fraction (*Non-benefitted*) $FPF_N(q) = P\{Q(S) > q | D < 1\}$, with the corresponding ROC curve being $ROC_N(t) = TPF_B\{FPF_N^{-1}(t)\}$. In Web Supplementary Appendix C, we relate the quantities TPF_B , FPF_U , FPF_H , and FPF_N to the TPF and FPF defined in (1). When the probability of $Y(0) < Y(1)$ or $D = -1$ is small, TPF approximates TPF_B , and FPF approximates FPF_U or FPF_N .

$$\begin{aligned} TPF(q) - TPF_B(q) &= [P\{Q(S) > q | D = 1\} + P\{Q(S) > q | D = -1\}] P(D = -1) / E(D), \\ FPF(q) - FPF_U(q) &= [P\{Q(S) > q | D = -1\} - 2P\{Q(S) > q | D = 0\}] P(D = -1) / \{1 - E(D)\}, \\ FPF(q) - FPF_N(q) &= [P\{Q(S) > q | D = -1\} - P\{Q(S) > q | D < 1\}] P(D = -1) / \{1 - E(D)\}. \end{aligned}$$

This is intuitive since small $P(D = -1)$ corresponds to small violation of A3. Therefore, under a small violation of A3, our proposed ROC curve defined under A3 still approximates some ROC curves with meaningful interpretations (ROC_U and ROC_N).

In practice, although the monotonicity assumption (A3) cannot be confirmed, it does have testable implications in a randomized trial. That is, evidence for a greater probability of $Y(1)=1$ than of $Y(0)=1$ within any subgroup rejects the monotonicity assumption. In addition we can conduct a sensitivity analysis in order to relax A3.

For a sensitivity analysis, we use a similar type of selection model as used by Gilbert et al. (2003) and Shepherd et al. (2008), and specify the following two model assumptions:

S.1: $P\{Y(0)=0|Y(1)=1, S\} = w(S; \delta, \gamma)$, where $w(S; \delta, \gamma) = \Phi\{\delta + h(\gamma, S)\}$, γ is fixed and known, $\Phi(\cdot)$ is a known cumulative distribution function, σ is an unknown parameter, and for each γ , $h(S; \gamma)$ is a known function of S .

S.2: $P\{Y(0)=0|Y(1)=1\} = \phi$ for a fixed constant ϕ .

The parameters γ and ϕ are unknown and not identified by the observed data; they are sensitivity parameters, treated as fixed and known in a single analysis and then varied over a plausible range of values to form a sensitivity analysis (Scharfstein et al., 1999). The parameter ϕ is constrained between $\max(0, \rho_1 - \rho_0)/\rho_1$ and $\min(\rho_1, 1 - \rho_0)/\rho_1$ for $\rho_z = P\{Y(z)=1\}$, $z=0, 1$. The parameter γ varies between $-\infty$ and ∞ . Setting γ to either extreme and also setting ϕ to its extremes provides the nonparametric bounds for the sensitivity analysis.

Following Gilbert et al. (2003), we specify $w(S; \delta, \gamma)$ with the anti-logit function, $w(S; \delta, \gamma) = \{1 + \exp(-\delta - \gamma S)\}^{-1}$, under which γ is the change in the log odds that a subject who experiences disease under active treatment $Z=1$ would be harmed by treatment – i.e., would have been spared disease had placebo been assigned – per unit increase of the marker. The parameter ϕ is interpreted as the conditional probability that the active treatment causes harm given that a subject gets disease under assignment to active treatment.

S.1 together with A1 and A2 imply the following relationships:

$$\begin{aligned} P\{Y(0)=0, Y(1)=1|S\} &= w(S; \delta, \gamma) Risk_{(1)}(S) \\ P\{Y(0)=1, Y(1)=1|S\} &= \{1 - w(S; \delta, \gamma)\} Risk_{(1)}(S) \\ P\{Y(0)=1, Y(0)=1|S\} &= Risk_{(0)}(S) - \{1 - w(S; \delta, \gamma)\} Risk_{(1)}(S) \\ P\{Y(0)=0, Y(1)=0|S\} &= 1 - w(S; \delta, \gamma) Risk_{(1)}(S) - Risk_{(0)}(S). \end{aligned} \quad (6)$$

In addition, straightforward calculation shows that

$$\int w(S; \delta, \gamma) Risk_{(1)}(S) dF(S) = P\{Y(0)=0, Y(1)=1\} = \phi \rho_1. \quad (7)$$

Equation (7) determines δ given fixed γ and ϕ . In earlier work (Gilbert et al., 2003), estimation of $Risk_{(z)}(S)$ and estimation of δ were performed separately. That is, $Risk_{(1)}(S)$ was estimated based on an ordinary MLE method, and together with estimates of F and ρ_1 , (7) was used to estimate δ . However, this separate estimation procedure will not guarantee that the estimates of $P\{Y(0)=1, Y(0)=0|S\}$ and $P\{Y(0)=0, Y(0)=0|S\}$ in (6) are nonnegative. We propose to extend the constrained MLE method under A3 to estimate the risk model and δ together. Specifically for given γ and ϕ , we maximize the log-likelihood (5) under the constraints, $Risk(S_i; Z=0) - \{1 - w(S_i; \delta, \gamma)\} Risk(S_i; Z=1)$, and $1 - w(S_i; \delta, \gamma) Risk(S_i; Z=1) - Risk(S_i; Z=0) \geq 0$ for $i=1, \dots, n$. Note that the constrained MLE under A3 is a special case where $\phi=0$ and $w=0$. For an application, one might vary γ between $-\log(10)$ and $\log(10)$, assuming that selection odds ratios greater than 10 are unlikely, and vary ϕ between $\max(0, \widehat{\rho}_1 - \widehat{\rho}_0)/\widehat{\rho}_1$ and $\min(\widehat{\rho}_1, 1 - \widehat{\rho}_0)/\widehat{\rho}_1$, where $\widehat{\rho}_z$ is an empirical estimate based on disease outcomes in treatment arm $Z=z$. Note that once ϕ in S.1-S.2 are specified, assumptions (A1, A2, A4) and the observed data identify all of the conditional probabilities in (6).

3. Simulation

We consider a simulation setting where we compare two markers with respect to their treatment-selection capacity. Two markers, S_1 and S_2 , are sampled from truncated bivariate normal distribution with correlation ρ equal to 0.0 or 0.5. For each marker, the potential outcomes follow the three-category cumulative link model (3) with logit link and $\gamma(S) = S$.

Consider a randomized trial with 1:1 randomization to $Z=0$ or $Z=1$. The risk model parameters are chosen such that the disease prevalence is 0.12 and 0.06 in the placebo and active arm respectively. For total sample size n varying from 200 to 1,000, we evaluate performance of the constrained MLE for $\text{TPF}(q)$, $\text{FPF}(q)$, and $\text{CDF}_Q(q)$ for $q_L = 0.035$ and $q_H = 0.085$, $\text{ROC}(t)$ for $t = 0.1, 0.3, 0.5, 0.7, 0.9$, and AUC. Overall 5,000 Monte Carlo simulations are performed, with 250 bootstrap samples for each dataset.

We present results for a setting where S_1, S_2 are each simulated from a normal distribution truncated at zero with mean 3, with standard deviation 0.5 and 0.8 respectively. The AUCs for marker 1 and 2 are 0.61 and 0.73 respectively. Estimates of the risk model parameters (Table 1), TPF , FPF , CDF_Q , $\text{ROC}(t)$, and AUC (Table 2 and Web Supplementary Table 1) have minimal bias and good converge using percentile bootstrap confidence intervals. Note that in our setting, estimates of $\text{ROC}(t)$ and AUC have much smaller variability compared to estimates of TPF , FPF , or CDF_Q . This is intuitive, considering the small difference in disease prevalence between treatment arms. The curves of $\text{TPF}(q)$, $\text{FPF}(q)$ or $\text{CDF}_Q(q)$ versus q thus are very steep, making it hard to estimate points on these curves precisely.

Enforcing the constraints on risk model estimation tends to increase efficiency. We looked at the efficiency of the constrained MLE relative to the MLE from ordinary logistic model fitting (i.e. ratio of variance for unconstrained MLE over variance of constrained MLE) for estimation of risk model parameters and CDF_Q . Enforcing the constraints appears to have a larger effect on model parameter estimates than on CDF_Q estimates. For example, with a sample size 1,000, the efficiency gains achieved by constrained model are around 46% for marker 1 and 50% for marker 2 in estimating α_1 and β_1 , and around 32% for marker 1 and 28% for marker 2 in estimating α_2 and β_2 ; for $\text{CDF}_Q(q_L)$ the efficiency gains are 10% and 18% for markers 1 and 2 respectively.

Our method has a small to medium power for comparing the two markers with an AUC difference of around 10% as sample size varies from 200 to 1000 (Table 3). Power increases when the two markers are positively correlated. In another example, we compare marker 1 with another marker that comes from a normal distribution truncated at zero with mean 3, variance 1, and AUC value 0.81. With a 20% difference in AUC between the two markers, we achieve reasonable power: 58.4, 81.3, and 94.8 respectively for $n = 200, 500, 1,000$ for $\rho = 0.0$, and 74.3, 92.0, and 98.9 for $\rho = 0.5$ (results not shown). Thus comparing treatment-selection performance between markers is a challenging task. Satisfactory power requires either a large sample size or a big difference in marker performance.

Table 4 presents the results of a sensitivity analysis based on a sample size of 500. We compare the area under ROC estimated under A3 with the areas under ROC_U and ROC_N for different ranges of the sensitivity parameter ϕ and for a range of values for γ from $-\log(10)$ to $\log(10)$. For ROC_U (and similarly for ROC_N), as deviation from A3 increases, we see an increase in bias of the AUC estimated under A3 and an increase in length of the sensitivity interval (the union of confidence intervals across the range of ϕ and γ) as expected.

4. Example

We illustrate our methods using data from the ‘Step’ HIV vaccine efficacy trial, a study with the primary objective of evaluating the ability of the MRKAd5 HIV-1 gag/pol/nef vaccine to prevent HIV infection. In this study 3000 HIV-seronegative participants were 1:1 randomized to receive the candidate vaccine or placebo, pre-stratified by sex, baseline adenovirus type 5 (Ad5) neutralization titer, and study site (Buchbinder et al., 2008). Since only one female was infected during the study, we restricted our analysis to men with 922 in the placebo arm and 914 in the vaccine arm.

Among male participants, HIV acquisition rates within 3 years of randomization were similar in vaccine (0.060) and placebo (0.051) recipients with baseline Ad5 titers < 18 , but for those with Ad5 ≥ 18 , the infection rate in vaccinees (0.058) appeared to be more than twice as high as that in placebo recipients (0.030). Since this particular vaccine appears to have either a harmful effect or no effect in various subgroups, we adopt the model assuming the vaccine has no benefit to an individual, i.e., $Y(0) = Y(1)$. Log-transformed baseline Ad5 is evaluated as a treatment-selection marker for the vaccine-induced increase in risk of HIV infection using the constrained MLE method with logistic risk model.

Let $Y(0)$ and $Y(1)$ be potential HIV infection status at the end of the study for a subject assigned to placebo and vaccine respectively. The estimated probabilities of $Y(0)$, $Y(1)$ conditional on Ad5 and their 95% percentile bootstrap confidence intervals are presented in figure 1. Specifically, $P\{Y(0) = 0, Y(1) = 0\}$ is the probability of avoiding infection under both assignments, $P\{Y(0) = 0, Y(1) = 1\}$ is the probability of being harmed by vaccination, and $P\{Y(0) = 1, Y(1) = 1\}$ is the probability of being infected no matter what assignment a subject receives. Note that the probability of being harmed by vaccine appears to increase with baseline Ad5 titer, which is consistent with the result based on dichotomizing baseline Ad5 at value 18.

On the basis of a decision rule $Q(S) = Risk_{(1)}(S) - Risk_{(0)}(S)$, plots of CDF_Q , TPF and FPF at various risk difference thresholds for baseline Ad5 are displayed in figure 2(a), with the ROC curve displayed in figure 2(c). For comparison purposes, the corresponding curves for log-transformed age are presented in figures 2(b) and (c). The sample correlation between $\log(\text{Ad5})$ and $\log(\text{age})$ is -0.15 . Variability in risk difference appears to be slightly larger for baseline Ad5 compared to age. Baseline Ad5 seems to have better discriminatory power in terms of the ROC curve for classifying an individual into treatment-effectiveness groups compared to age, with AUC estimate 0.72 and 0.63 respectively.

Web Supplementary Table 2 shows the treatment-selection capacity of Ad5 and age with respect to TPF, FPF, and CDF_Q for a given low risk difference threshold $q_L = 0.005$ and a high risk difference threshold $q_H = 0.025$, as well as the points on the ROC curve and the AUC. Not surprisingly, the variability in the TPF, FPF, or CDF_Q estimates is much larger compared to the estimates of the ROC curve given that HIV infection is a rare event. The difference in estimated performance between the two candidate markers is small, and not statistically significant for any of the measures.

We also conducted sensitivity analysis under violation of the monotonicity assumption. Without monotonicity, baseline Ad5 could have substantially lower classification accuracy. For example, estimates of the AUC under the ROC_N curve (based on $D < 1$ versus $D = 1$) range from 0.64 to 0.75 for $\phi = 0.1$, from 0.58 to 0.75 for $\phi = 0.2$, and from 0.53 to 0.75 for $\phi = 0.3$ when γ varies from $-\log(10)$ to $\log(10)$.

5. Discussion

In this paper we propose characterizing a marker's capacity for predicting individual treatment effect using a potential-outcomes based framework. Under this framework with a monotone treatment effect assumption, we can identify the joint distribution of the pair of the potential outcomes under the two treatment assignments, which is meaningful to individual patients. We propose measures that quantify a marker's ability to classify a subject according to potential treatment-effectiveness — the ROC measure, as well as a marker's ability to inform treatment decisions — the CDF_Q measure. While the CDF_Q measure depends on the scale of response rate, the ROC measure can be thought of an intrinsic property of a biomarker since it does not depend on disease prevalence within

treatment group and thus is potentially useful to compare markers across different populations or studies. Under a randomized trial design, we develop a constrained maximum likelihood method for estimation of these measures. The procedure is closely related to common risk model estimation and is easy to implement. R code is available for downloading at <http://labs.fhcrc.org/huang/index.html>.

In practice, we recommend the method with monotonicity assumption be used when its plausibility is supported by data and ideally also by the scientific knowledge. If no major subgroups of interest have an overall treatment effect estimate in the ‘wrong direction’, then monotonicity is supported. When this assumption holds, the method we use to incorporate it is desirable for the simple interpretation and the efficiency gain. We have shown through simulation studies that when violation of the monotonicity assumption is minor, we still get meaningful estimates that approximate the measures defined in a general setting that allows for a non-monotone treatment effect. We further propose a sensitivity analysis as a way to relax the monotonicity assumption.

Our proposed approach facilitates a more-informed assessment of a marker’s value for treatment selection than does the classical approach of testing for an interaction between marker and treatment in an ordinary risk model. As we have demonstrated in the example of Web Supplementary Appendix A, a strong interaction coefficient is important for a marker to have value for treatment selection but is not useful for summarizing performance because it depends on other coefficients in the risk model as well as the functional form of the model. Therefore the interaction coefficient is not directly comparable between markers (models). In practice, a common approach is to plot disease risk versus marker value for each treatment group. This provides useful information to individual patients who have marker results in hand about their expected benefit of treatment given their marker measure. In this manuscript, we address a different question which is more relevant to biomarker researchers and policy makers: i.e., how to characterize and compare markers (risk models) with respect to their treatment-selection capacity. The ROC and CDF_Q measures proposed here are suitable for this task. For a univariate marker, a plot of q versus CDF_Q(q) is actually just the difference in risk between treatment groups as a function of marker percentile.

We want to point out that good treatment selection performance as characterized by our proposed measures corresponds to large variability in risk difference as a function of marker value. The actual impact of measuring the marker on treatment decisions depends on the risk difference threshold. Our measures provide an overview of treatment-selection capacity allowing the risk difference threshold to vary. This is helpful in situations where there does not exist a well-established decision threshold and the choice relies on other factors such as the cost and side-effects of the active treatment. This is oftentimes true in practice.

Song and Pepe (2004) proposed two measures to characterize a univariate marker’s treatment-selection performance: (i) the population response rate (θv) given a treatment policy that assigns a subject to active treatment if one’s marker value exceeds the v^{th} quantile in the population; (ii) individual difference in risk with and without treatment, $d(v)$, conditional on the v^{th} quantile of the marker value. For the special case with $Q(S)$ equal to the risk difference, our proposed CDF_Q measure is the inverse function of $d(v)$ and gauges the population impact of the marker in helping make informed treatment decisions.

We studied a univariate marker in our example, but the methodology is general and applies to models with multiple biomarkers by entering the joint distribution of markers into (1). Our measures can be used to compare two general risk models. For example, we can compare two biomarkers with respect to their treatment-selection capacity as demonstrated in this paper, and we can look at the gain by adding a new marker into a baseline model.

The methods we propose summarize the treatment-selection capacity of a marker in the whole population. In practice, a marker's performance might vary with other baseline covariates such as age or gender and its performance conditional on those covariates is often of interest. Our framework can be easily extended to allow the assessment of a marker's treatment-selection capacity within subpopulations defined by other covariates. Specifically, for estimation of the proposed performance measures conditional on specific covariates value, we need to estimate the risk model with the additional covariates included and estimate the marker's distribution conditional on covariates. Again constraints can be enforced on the risk model. This is currently under investigation.

In this manuscript we focus on a binary endpoint, but the ROC and CDF_Q measures proposed can be naturally extended to handle an event time endpoint. Specifically, for given time t of interest, these measures can be constructed from the risk difference between treated and untreated groups, where risk is defined as the probability of developing the disease before time t . Existing techniques for estimating risk using censored data can be employed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to reviewers for constructive comments. This work was supported by NIH grants P30CA015704, R01CA152089, 2R37AI05465-08, R01GM054438, P01CA053996, and U24CA086368.

References

- Albain KS, Barlow WE, Shak S, Hortobagoy GN, Livingston RB, Yeh I-T, et al. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: A retrospective analysis of a randomised trial. *Lancet Oncol.* 2010; 11:55–65. [PubMed: 20005174]
- Allegra CJ, Jessup M, Somerfield MR, Hamilton SR, Hammond EH, Hayes DF, et al. American Society of Clinical Oncology Provisional Clinical Opinion: Testing for KRAS Gene Mutations in Patients With Metastatic Colorectal Carcinoma to Predict Response to AntiEpidermal Growth Factor Receptor Monoclonal Antibody Therapy. *J Clin Onc.* 2009; 27:2091–6.
- Buchbinder SP, Mehrotra DV, Duerr A, Fitzgerald DW, Mogg R, Li D, et al. Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *Lancet.* 2008; 372:1881–1893. [PubMed: 19012954]
- Buyse M. Towards validation of statistically reliable biomarkers. *EJC Supp.* 2007; 5(5):89–95.
- Duffy MJ. Predictive markers in breast and other cancers: a review. *Clinical Chemistry.* 2005; 51(3): 494–503. [PubMed: 15637130]
- Freidlin B, Simon R. Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Research.* 2005; 11:7872–8.
- Gilbert P. Goodness-of-fit tests for semiparametric biased sampling models with application to AIDS vaccine trials. *JSPI.* 2003; 118:51–81.
- Gilbert PB, Bosch RJ, Hudgens MG. Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics.* 2003; 59:531–541. [PubMed: 14601754]
- Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, Taube S, et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *Journal of Clinical Oncology.* 2007; 25:5287–312. [PubMed: 17954709]
- Janes H, Pepe MS, Bossuyt PB, Barlow WE. Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine.* 2011; 154(4):253–259. [PubMed: 21320940]

- Karapetis CS, Khambata-Ford S, Jonker DJ, O-Callaghan CH, Tu D, Tebbutt NC, et al. K-ras Mutations and Benefit from Cetuximab in Advanced Colorectal Cancer. *NEJM*. 2008; 359:1757–65. [PubMed: 18946061]
- Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. *Journal of Clinical Epidemiology*. 2009; 62:364–73. [PubMed: 18945590]
- Mandrekar SJ, Grothey A, Goetz MP, Sargent DJ. Clinical trial designs for prospective validation of biomarkers. *American Journal of Pharmacogenomics*. 2005; 5:317–25. [PubMed: 16196501]
- McIntosh MW, Pepe MS. Combining several screening tests: optimality of the risk score. *Biometrics*. 2002; 58(3):657–664. [PubMed: 12230001]
- Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *Journal of Clinical Oncology*. 2006; 24:3726–34. [PubMed: 16720680]
- Qin J, Zhang J. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*. 1997; 84(3):609–618.
- Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*. 2005; 23:2020–7. [PubMed: 15774793]
- Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *JASA*. 1999; 94:1096–1146.
- Shepherd BE, Redman MW, Ankerst DP. Does finasteride affect the severity of prostate cancer? A causal sensitivity analysis. *JASA*. 2008; 484:1392–1404. [PubMed: 20526381]
- Simon R. The use of genomics in clinical trial design. *Clin Cancer Res*. 2008; 14(19):5954–8. [PubMed: 18829473]
- Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *JNCI*. 2009; 21:1446–52. [PubMed: 19815849]
- Song X, Pepe MS. Evaluating markers for selecting a patient's treatment. *Biometrics*. 2004; 60(4): 874–883. [PubMed: 15606407]
- Wang X, Zhou H. Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling. *Biometrics*. 66(2):502–511. [PubMed: 19508239]

1. Probability of potential outcomes conditional on log(Ad5).

2.

Plot of $CDF_Q(q)$, $TPF(q)$, $FPF(q)$ versus $q = Risk_{(1)} - Risk_{(0)}$ for baseline Ad5 (a) and age (b), and corresponding ROC curves for classifying a subject into treatment effective or ineffective groups (c).

Table 1
Bias, SE, and Coverage of bootstrap CI of risk model parameter estimates for each marker

<i>n</i>	Marker 1				Marker 2			
	α_1	β_1	α_2	β_2	α_1	β_1	α_2	β_2
True	1.24	-1.4	1.476	-1.2	1.843	-1.8	2.272	-1.6
Bias								
200	-0.054	-0.074	0.165	-0.065	0.089	-0.153	0.237	-0.107
500	-0.081	0.001	0.099	-0.039	0.002	-0.033	0.125	-0.055
1000	-0.059	0.007	0.05	-0.019	-0.026	-0.007	0.068	-0.03
SE								
200	2.11	0.85	1.52	0.54	1.44	0.75	1.12	0.46
500	1.24	0.45	1	0.35	0.81	0.36	0.69	0.28
1000	0.93	0.34	0.73	0.26	0.6	0.26	0.51	0.21
Coverage of 95% bootstrap percentile CI								
200	92.3	91	90.3	90.3	92.2	88.7	90	89.9
500	94.2	93.9	91.4	91.6	95.2	94	91.8	92.1
1000	94.8	94.7	92.6	92.9	95	94.4	93.3	93.2

Table 2
Bias, SE, and Coverage of bootstrap CI of estimates for $\text{TPF}(qL)$, $\text{FPF}(qL)$, and $\text{CDF}_{\mathcal{Q}}(qL)$ at $qL = 0.035$, and of estimates for $\text{ROC}(t)$ for $t_1 = 0.1$, $t_2 = 0.3$, and AUC for each marker

n	Marker 1					Marker 2				
	TPF	FPF	$\text{CDF}_{\mathcal{Q}}$	ROC	AUC	TPF	FPF	$\text{CDF}_{\mathcal{Q}}$	ROC	AUC
True	0.94	0.86	0.14	t_1 0.18	t_2 0.45	d_L 0.88	0.58	0.4	t_1 0.29	t_2 0.63
200	-0.14	-0.17	0.17	0.03	0.04	0.02	-0.07	-0.06	0.05	0.04
500	-0.08	-0.12	0.11	0.02	0.02	0.02	-0.03	-0.03	0.03	0.03
1000	-0.05	-0.07	0.07	0.01	0.01	0.01	-0.01	-0.02	0.02	0.02
Bias										
200	0.31	0.33	0.33	0.09	0.12	0.08	0.25	0.24	0.24	0.12
500	0.23	0.26	0.26	0.08	0.11	0.07	0.15	0.17	0.17	0.09
1000	0.16	0.21	0.21	0.06	0.09	0.06	0.08	0.12	0.12	0.08
SE										
200	94.4	92.7	92.7	95.3	94.2	93.7	92.5	92.6	93	92.7
500	94.9	94.2	94.2	95.5	95	94.7	93.4	93.4	93.6	94.8
1000	95.2	94.4	94.5	94.6	94.5	94.4	93	94.7	94.8	95.3
Coverage of 95% bootstrap percentile CI										
200	94.4	92.7	92.7	95.3	94.2	93.7	92.5	92.6	93	92.7
500	94.9	94.2	94.2	95.5	95	94.7	93.4	93.4	93.6	94.8
1000	95.2	94.4	94.5	94.6	94.5	94.4	93	94.7	94.8	95.3

Table 3
Performance for estimating difference in AUC between the two markers with true AUC difference 0.12

n	ρ	Bias	SE	Coverage*	Power
200	0	-0.01	0.10	95.8	19.3
	0.5	-0.009	0.09	96.7	24.8
	0	-0.003	0.08	95.8	27.8
500	0	-0.002	0.07	97.0	38.6
	0	-0.004	0.07	96.0	36.2
	0.5	-0.001	0.06	96.5	52.5

* coverage of 95% percentile bootstrap CI

Table 4

Sensitivity Analyses for estimating the AUC for each marker under varying degrees of violation of the monotonicity assumption. $P(D = -1)$ is less than 6% in all scenarios

		range of φ	Bias ^a	Median Statistics ^b				MSIW ^c
				Min ^b	Med ^b	Max ^b	Range ^b	
AUC under ROC _U								
Marker 1	1	$\varphi = 0$	0.015	0.621	0.621	0.621	0	0.261
	φ	0.1	0.003 – 0.032	0.598	0.622	0.648	0.039	0.353
	φ	0.6	−0.012 – 0.04	0.585	0.633	0.694	0.099	0.553
Marker 2	$\varphi = 0$		0.011	0.736	0.736	0.736	0	0.214
	φ	0.1	0.000 – 0.027	0.701	0.736	0.753	0.047	0.309
	φ	0.6	−0.025 – 0.029	0.695	0.738	0.78	0.079	0.38
AUC under ROC _N								
Marker 1	$\varphi = 0$		0.015	0.621	0.621	0.621	0	0.261
	φ	0.1	0.004 – 0.032	0.598	0.621	0.646	0.038	0.347
	φ	0.6	−0.005 – 0.04	0.585	0.63	0.684	0.089	0.514
Marker 2	$\varphi = 0$		0.011 – 0.011	0.736	0.736	0.736	0	0.214
	φ	0.1	0.002 – 0.027	0.701	0.736	0.751	0.045	0.306
	φ	0.6	−0.013 – 0.03	0.694	0.735	0.771	0.073	0.366

^aBias: difference between mean AUC estimated under A3 and true AUC under ROC_U or ROC_N

^b(Min,Med, Max,Range): medians of the three order statistics and median range over 5,000 simulations for AUC under estimated ROC_U or ROC_N

^cMSIW: median width of the 95% sensitivity interval, which is the union of the 95% Wald confidence intervals for all the γ and ϕ considered, based on Monte Carlo standard error computed over 5,000 simulations.