Published in final edited form as:
Biometrics. 2009 September ; 65(3): 894–902. doi:10.1111/j.1541-0420.2008.01125.x.

Identifying subjects who benefit from additional information for better prediction of the outcome variables

L. Tian[1,*], **T. Cai**[2], and **L.J. Wei**[2]

[1]Department of Preventive Medicine, Northwestern University, Chicago, IL 60611, USA.

[2]Department of Biostatistics, Harvard University, Boston, MA 02115, USA.

SUMMARY

Suppose that we are interested in using new bio- or clinical-markers to improve prediction or diagnosis of the patient's clinical outcome in addition to the conventional markers. The incremental value from the new markers is typically assessed by averaging across patients in the entire study population. However, when measuring the new markers is costly or invasive, an overall improvement does not justify measuring the new markers in all patients. A more practical strategy is to utilize the patient's conventional markers to decide whether the new markers are needed for improving prediction of his/her health outcomes. In this article, we propose inference procedures for the incremental values of new markers across various subgroups of patients classified by the conventional markers. The resulting point and interval estimates can be quite useful for medical decision makers seeking to balance the predictive or diagnostic value of new markers against their associated cost and risk. Our proposals are theoretically justified and illustrated empirically with two real examples.

Keywords

Biomarker; Cardiovascular events; Diagnosis; *K*-fold crossvalidation; Prediction accuracy; Subgroup analysis

1. Introduction

Biological and technological advances continually generate promising new clinical- and biomarkers with the potential to improve medical care by providing more accurate, personalized predictions of health outcomes and diagnoses of clinical phenotypes. However, extensive use of new markers may provide only negligible improvements in prediction or diagnosis, while subjecting patients to additional risks and costs. It is therefore important to develop statistical methods that can quantify for individual patients the value of new markers over conventional ones, especially when measuring these markers is costly or invasive. As an example from a recent study, Wang et. al. (2006) examined extensively the incremental values of ten new biomarkers in predicting first major cardiovascular events or death in a Framingham cohort. There were 3209 participants in the study. They were followed for a median of 7.4 years, during which 207 participants died and 169 had a first major cardiovascular event. Based on various prediction precision criteria, the study

[*]lutian@northwestern.edu.

Supplementary Materials
Web Appendices and figure referenced in Sections 2 and 3 are available under the Paper Information link at the Biometrics website http://www.tibs.org/biometrics.

investigators found that these ten contemporary biomarkers added only moderate *overall* predictive value to the classical risk factors including gender, age, total cholesterol, HDL cholesterol, blood pressure, smoking status, and diabetes mellitus. In contrast, other investigators studying different populations with different prediction precision measures demonstrated that certain biomarkers, for example, the high-sensitivity C-reactive protein, provide clinically useful prognostic information on top of the traditional Framingham risk score for heart diseases (Ridker et. al., 2002, 2007, Blumenthal et. al., 2007).

Despite these often controversial findings in the literature, clinical practitioners would generally not change their recommendation for the patient's care with the extra marker information if the patient, for example, has either *high* or *low* conventional risk score. Therefore, a practically important question is how to *systematically* identify patients who would benefit from the additional markers instead of evaluating these markers based only on their average incremental value across the entire population (D'Agostino, 2006). In this article, we propose procedures to estimate the incremental values of new markers for diagnosis or prediction in various subgroups of patients classified by conventional markers. These, coupled with the sampling variations of the estimates, provide a useful tool for researchers and practitioners to decide when, after observing the conventional risk factors, the new markers are needed. In Section 2, we describe in detail the new procedure and provide theoretical justification. In Section 3, we illustrate our methods with two examples, one with a continuous response and the other with a binary outcome.

There are quite a few procedures in the literature for evaluating the over-all incremental value of new markers for an entire population of interest. For example, Pepe et. al. (2004) compared the ROC curves among models with and without an additional marker. Recently, Tian et. al. (2007) and Uno et. al. (2007) proposed robust inference procedures for evaluating prediction rules. Prediction or diagnostic precision measures, which may be used for comparing different prediction procedures, have also been proposed and utilized, for example, by Brier (1950), Breiman et. al. (1984), Speigelhalter (1986), Korn and Simon (1990), McLachlan (1992), Mittlböck and Schemper (1996), Ripley (1996), Zhou et. al. (2002), Pepe (2003) and Pencina et. al. (2008).

2. Estimating Subject-Specific Prediction Error Based on Risk Score Constructed from Conventional Markers

Let Y be a continuous or binary response variable, U be the set of its conventional marker values, and V be the corresponding counterpart from the new markers. Our data consist of n independent copies $\{(Y_i, U_i, V_i), i = 1, \cdots, n\}$ from (Y, U, V). The problem is how to use the data to identify future subjects via U, which would benefit from the new markers for better prediction of their responses Y. Suppose that there are no well-established rules for classifying subjects based on U for predicting Y. First, we may estimate a center value of Y given U nonparametrically and use this estimate to construct a predictor for Y. We then estimate the average prediction error, the "distance" between the observed response and its predicted value over all subjects which have the same marker value U. Next, we estimate the center of Y given U and V, and estimate the corresponding average prediction error conditional only on U. Inferences about the improvement from the new markers can be made via these functional estimates over U. Unfortunately, in general, we can only construct nonparametric functional estimates, which behave reasonably well, when the dimension of U is very small and the sample size n is quite large.

A practically feasible alternative to handle this problem is to consider a parametric or semi-parametric approach. To this end, let X be a p-dimensional vector, a function of U. Assume that the conditional mean of Y given U can be approximated by the following *working* model

$$E(Y \mid U)=g_1(\beta' X), \tag{2.1}$$

where $g_1(\cdot)$ is a smooth, strictly increasing, known function and β is an unknown vector of parameters. Note that the first component of X is one. In this article, we deal with the interesting and challenging case that $\beta' X$ is a continuous variable.

To estimate the regression parameters for model (2.1) which, most likely, is an approximation to the true conditional mean of Y given U, one may use the estimator $\hat{\beta}$ based on the simple estimating function

$$S_1(\beta)=\sum_{i=1}^{n} X_i\,\{Y_i - g_1(\beta' X_i)\}, \tag{2.2}$$

where $\{(Y_i, X_i), i = 1, \cdots, n\}$ are n-independent copies of (Y, X) (Tian et. al., 2007). Note that even when (2.1) is not the true model, $\hat{\beta}$ converges to a constant vector β_0, as $n \to \infty$. It is not clear, however, that other standard estimators for β in (2.1) would be convergent as n gets large.

Now, consider a future independent subject with $(Y, X) = (Y^0, X^0)$. For a given β in (2.1), let $\hat{Y}_1(\beta' X^0)$ be the predictor for Y^0. For example, when Y^0 is continuous, one may let $\hat{Y}_1(\beta' X^0) = g_1(\beta' X^0)$ and when Y^0 is binary, one may predict Y^0 by a binary variable $\hat{Y}_1(\beta' X^0) = I\{g_1(\beta' X^0) \geq 0.5\}$, where $I(\cdot)$ is the indicator function. Other prediction rules for the binary case will be discussed in the Example Section. To evaluate the performance of $\hat{Y}_1(\hat{\beta}' X^0)$, we first need to quantify its prediction accuracy based on a "distance" between the true Y^0 and the predicted $\hat{Y}_1(\hat{\beta}' X^0)$, denoted by $D\{Y^0, \hat{Y}_1(\hat{\beta}' X^0)\}$. For example, a simple, physically interpretable function is the absolute prediction error with $D(a, b) = |a - b|$. When Y is binary, this distance function is simply the overall mis-classification rate. Choosing an appropriate distance measure is a crucial yet often difficult step in evaluating the incremental value for the cost-benefit decision. It is important to note that such quantification should have a practical/clinical interpretation. We discuss this issue in great details with real examples in Section 3 and also in the Remarks Section.

Since clinical practitioners almost always group subjects with a "risk scoring system" for medical decision making, we consider an average prediction error over a set of X's which have "similar" $g_1(\hat{\beta}' X)$ to evaluate $\hat{Y}_1(\cdot)$. To be specific, let $J_z = (c_z, d_z)$ be a data-independent interval centered about z, where z ranges over a set of possible values of $g_1(\beta_0' X)$. The average prediction error over J_z is $\mathscr{D}_1^*(z)=E[D\{Y^0, \widehat{Y}_1(\widehat{\beta}' X^0)\}|\, g_1(\widehat{\beta}' X^0) \in J_z]$, where the conditional expectation is taken with respect to (Y^0, X^0) and $\hat{\beta}$. As $n \to \infty$, $\mathscr{D}_1^*(z)$ converges to

$$\mathscr{D}_1(z)=E\left[D\{Y^0, \widehat{Y}_1(\beta_0' X^0)\}|\, g_1(\beta_0' X^0) \in J_z\right]. \tag{2.3}$$

As a process of z, this moving average process $\{\mathscr{D}_1(z)\}$ provides a performance profile of $\hat{Y}_1(\cdot)$ over all possible values of $g_1(\beta_0' X)$. The proper choice of interval J_z would be made on a case-by-case basis and is illustrated with two examples in the next section.

Now, let W be a q × 1 vector, a function of U and V. Assume that a working model for the conditional mean of Y given U and V is

$$\mathrm{E}(Y \mid U, V)=g_2(\theta' W), \tag{2.4}$$

where $g_2(\cdot)$ is a smooth, strictly increasing, known function and θ is an unknown vector of parameters. The first component of W is one. Again, we assume that $g_2(\theta' W)$ is a continuous variable. Let $\hat{\theta}$ be the estimator for θ obtained from the following simple estimating function

$$S_2(\theta)=\sum_{i=1}^{n} W_i\,\{Y_i - g_2(\theta' W_i)\}, \tag{2.5}$$

where W_i, $i = 1, \cdots, n$, are n independent copies of W. Let θ_0 be the limit of $\hat{\theta}$. Consider a future independent $(Y, X, W) = (Y^0, X^0, W^0)$. Let $\hat{Y}_2(\theta' W^0)$ be the predictor constructed from (2.4) with parameter value θ, the counterpart of $\hat{Y}_1(\beta' X^0)$. For the aforementioned interval J_z, let the average prediction error for $\hat{Y}_2(\cdot)$ over J_z be

$$\mathscr{D}_2(z)=E\left[D\{Y^0, \widehat{Y}_2(\theta_0' W^0)\}\middle| \quad g_1(\beta_0' X^0) \in J_z\right], \tag{2.6}$$

where the expectation is taken with respect to (Y^0, X^0, W^0). Then, as a processes in z, $\mathscr{D}_1(z)$, $\mathscr{D}_2(z)$ and

$$\Delta_0(z)=\mathscr{D}_1(z) - \mathscr{D}_2(z) \tag{2.7}$$

provide a global picture for identifying subgroups of patients who would benefit from the additional markers.

To estimate $\mathscr{D}_1(z)$ and $\mathscr{D}_2(z)$, one may use $\mathscr{D}_1(z) = \tilde{\mathscr{D}}_1(z, \hat{\beta})$ and $\mathscr{D}_2(z) = \tilde{\mathscr{D}}_2(z, \hat{\beta}, \hat{\theta})$, respectively, where

$$\tilde{\mathscr{D}}_1(z,\beta)=\frac{\sum_{i=1}^{n} D\{Y_i, \widehat{Y}_1(\beta' X_i)\} I\{g_1(\beta' X_i) \in J_z\}}{\sum_{i=1}^{n} I\{g_1(\beta' X_i) \in J_z\}} \tag{2.8}$$

and

$$\tilde{\mathscr{D}}_2(z,\beta,\theta)=\frac{\sum_{i=1}^{n} D\{Y_i, \widehat{Y}_2(\theta' W_i)\} I\{g_1(\beta' X_i) \in J_z\}}{\sum_{i=1}^{n} I\{g_1(\beta' X_i) \in J_z\}}. \tag{2.9}$$

We then let $\hat{\Delta}(z) = \mathscr{D}_1(z) - \mathscr{D}_2(z)$ to estimate $\Delta_0(z)$. In Appendix A of the web-based supplementary material available at http://www.tibs.org/biometrics, we show that with the distance function $D(a, b) = |a - b|$ or a function thereof, the above three estimators are uniformly consistent over an interval Ω consisting of all z's whose intervals J_z's are properly in the support of $g_1(\beta_0' X)$. Similar arguments may be used for cases with other distance functions.

To make further inferences about the added value from the new markers for predicting the response, in Appendix A of the web-based supplementary material, we show that the limiting distributions of the processes $\widehat{\mathscr{W}_1}(z)=n^{1/2}\{\widehat{\mathscr{D}}_1(z)-\mathscr{D}_1(z)\}$, $\widehat{\mathscr{W}_2}(z)=n^{1/2}\{\widehat{\mathscr{D}}_2(z)-\mathscr{D}_2(z)\}$ and $\widehat{\mathscr{W}}(z)=n^{1/2}\{\widehat{\Delta}(z)-\Delta_0(z)\}$, are the same as those of the Gaussian processes $\mathscr{W}_1^*(z)$, $\mathscr{W}_2^*(z)$ and $\mathscr{W}^*(z)$, respectively, for $z \in \Omega$, where $\mathscr{W}^*(z)=\mathscr{W}_1^*(z)-\mathscr{W}_2^*(z)$,

$$\mathscr{W}_1^*(z)=n^{\frac{1}{2}}\left\{\frac{\sum_{i=1}^n[D\{Y_i,\widehat{Y}_1(\widehat{\beta}'X_i)\}-\widehat{\mathscr{D}}_1(z)]I\{g(\widehat{\beta}'X_i)\in J_z\}G_i}{\sum_{i=1}^n I\{g(\widehat{\beta}'X_i)\in J_z\}}+\tilde{\mathscr{D}}(z,\widehat{\theta}^*)-\widehat{\mathscr{D}}_1(z)\right\},$$

$$\mathscr{W}_2^*(z)=n^{\frac{1}{2}}\left\{\frac{\sum_{i=1}^n[D\{Y_i,\widehat{Y}_2(\widehat{\theta}'W_i)\}-\widehat{\mathscr{D}}_2(z)]I\{g(\widehat{\beta}'X_i)\in J_z\}G_i}{\sum_{i=1}^n I\{g(\widehat{\beta}'X_i)\in J_z\}}+\tilde{\mathscr{D}}(z,\widehat{\beta}^*,\widehat{\theta}^*)-\widehat{\mathscr{D}}_2(z)\right\},$$

$$\widehat{\beta}^*=\widehat{\beta}+\left\{\sum_{i=1}^n \dot{g}_1(\widehat{\beta}'X_i)X_iX_i'\right\}^{-1}\sum_{i=1}^n X_i\{Y_i-g_1(\widehat{\beta}'X_i)\}G_i$$

$$\widehat{\theta}^*=\widehat{\theta}+\left\{\sum_{i=1}^n \dot{g}_2(\widehat{\theta}'W_i)W_iW_i'\right\}^{-1}\sum_{i=1}^n W_i\{Y_i-g_2(\widehat{\theta}'W_i)\}G_{i,}$$

and $\{G_1, \ldots, G_n\}$ are independent standard normal random variables that are independent of the data. Here, realizations from three Gaussian processes $\mathscr{W}_1^*(z)$, $\mathscr{W}_2^*(z)$ and $\mathscr{W}^*(z)$ given above can be generated easily for any interval of z, where $\mathscr{D}_1(z)$ and $\mathscr{D}_2(z)$ are well-defined. In practice, one may not be able to construct reasonably well-behaved interval estimators for $\mathscr{D}_l(z)$, $l = 1, 2$, for z is the tail parts of Ω. To this end, let $\hat{\Omega}$ be a set of z such that $J_z \subset [\eta_1, \eta_2]$, where $n^{-1}\sum_{i=1}^n I\{g(\widehat{\beta}'X_i)\le\eta_1\}>d_1, n^{-1}\sum_{i=1}^n I\{g(\widehat{\beta}'X_i)\ge\eta_2\}>d_2$, and d_1 and d_2 are given positive numbers. Then, with the above large sample approximations, for $z \in \hat{\Omega}$, a $(1 - \alpha)$, $0 < \alpha < 1$, point-wise confidence interval for $\mathscr{D}_l(z)$, $l = 1, 2$, is

$$\widehat{\mathscr{D}}_l(z)\pm n^{-\frac{1}{2}}\xi_{\alpha/2}\sigma_{\mathscr{W}_l^*(z)}. \tag{2.10}$$

Here, $\sigma^2_{\mathscr{W}_l^*(z)}$ is the variance of the random variable $\mathscr{W}_l^*(z)$ and ξ_α is the upper 100αth percentage point of the standard normal. Furthermore, a $(1 - \alpha)$ simultaneous confidence band for $\{\mathscr{D}_l(z), z \in \hat{\Omega}\}$ is

$$\widehat{\mathscr{D}}_l(z)\pm n^{-\frac{1}{2}}\tau_{l\alpha}\sigma_{\mathscr{W}_l^*(z)}, \tag{2.11}$$

where

$$\mathrm{pr}\,\{\sup_{z\in\widehat{\Omega}}|\mathscr{W}_l^*(z)/\sigma_{\mathscr{W}_l^*(z)}|<\tau_{l\alpha}\}=1-\alpha.$$

It is important to note that in contrast to the standard subgroup analysis, our proposal takes care of the multiple comparison problems with such simultaneous confidence interval estimates via the scoring system indexed by z.

To construct interval estimators for $\Delta_0(z)$, it is important to note that $\hat{\Delta}(z)$ has a degenerate limiting distribution when $\widehat{Y}_1(\beta_0'X)=\widehat{Y}_2(\theta_0'W)$ for all $g_1(\beta_0'X)\in J_{\underset{\sim}{z}}$. Therefore, to obtain reasonable interval estimators in practice, we consider the set $\Omega \subset \hat{\Omega}$ such that for

$z \in \tilde{\Omega}$, $\sum_{i=1}^{n} I\{\widehat{Y}_1(\widehat{\beta'}X_i) \neq \widehat{Y}_2(\widehat{\theta'}W_i), g_1(\widehat{\beta'}X_i) \in J_z\} / \sum_{i=1}^{n} I(g_1(\widehat{\beta'}X_i) \in J_z) > d_3$, where d_3 is a given positive number. Then, for $z \in \tilde{\Omega}$, a $(1 - \alpha)$, $0 < \alpha < 1$, point-wise confidence interval for $\Delta_0(z)$, is

$$\widehat{\Delta}(z) \pm n^{-\frac{1}{2}} \xi_{\alpha/2} \sigma_{\mathscr{W}^*(z)}. \tag{2.12}$$

Here, $\sigma^2_{\mathscr{W}^*(z)}$ is the variance of the random variable $\mathscr{W}^*(z)$. Moreover, a $(1 - \alpha)$ simultaneous confidence band for $\{\Delta_0(z), z \in \tilde{\Omega}\}$ is

$$\widehat{\Delta}(z) \pm n^{-\frac{1}{2}} \tau_{\alpha} \sigma_{\mathscr{W}^*(z)}, \tag{2.13}$$

where

$$\text{pr}\{\sup_{z \in \tilde{\Omega}} |\mathscr{W}^*(z)/\sigma_{\mathscr{W}^*(z)}| < \tau_{\alpha}\} = 1 - \alpha.$$

Note that for the case with a continuous response Y, $\hat{\Omega} = \tilde{\Omega}$.

Now, since we use the entire data set to estimate the parameters in (2.1) and (2.4) and also to estimate the average prediction errors (2.3) and (2.6), $\mathscr{D}_1(\cdot)$ and $\mathscr{D}_2(\cdot)$ may be significantly underestimated. To reduce such potential bias, one may consider the commonly used K-fold crossvalidation scheme. Specifically, we randomly split the data into K disjoint subsets of about equal size and label them as $\mathscr{I}_k$, $k = 1, \cdots, K$. For each k, we use all the observations, which are *not* in $\mathscr{I}_k$, to estimate parameters in (2.1) and (2.4) via estimating functions (2.2) and (2.5), and then use the observations in $\mathscr{I}_k$ to estimate prediction errors $\mathscr{D}_1(\cdot)$ and $\mathscr{D}_2(\cdot)$ with (2.8) and (2.9). Let the resulting estimators be denoted by $\widehat{\mathscr{D}}_{1k}(\cdot)$ and $\widehat{\mathscr{D}}_{2k}(\cdot)$, respectively. The crossvalidated estimators for $\mathscr{D}_1(\cdot)$, $\mathscr{D}_2(\cdot)$ and $\Delta_0(\cdot)$ are $\tilde{\mathscr{D}}_1(\cdot) = K^{-1}\sum_{k=1}^{K} \tilde{\mathscr{D}}_{1k}(\cdot)$, $\tilde{\mathscr{D}}_2(\cdot) = K^{-1}\sum_{k=1}^{K} \widehat{\mathscr{D}}_{2k}(\cdot)$ and $\tilde{\Delta}(\cdot) = \tilde{\mathscr{D}}_1(\cdot) - \tilde{\mathscr{D}}_2(\cdot)$, respectively. Again, these estimators are uniformly consistent if K is relatively small with respect to n.

In Appendix B of the web-based supplementary material, we show that for large n, the distributions of the processes $\mathscr{W}_1(\cdot) = n^{1/2}\{\tilde{\mathscr{D}}_1(\cdot) - \mathscr{D}_1(\cdot)\}$, $\mathscr{W}_2(\cdot) = n^{1/2}\{\tilde{\mathscr{D}}_2(\cdot) - \mathscr{D}_2(\cdot)\}$ and $\mathscr{W}\ (\cdot) = n^{1/2}\{\tilde{\Delta}(\cdot) - \Delta_0(\cdot)\}$ can also be approximated well by those of $\mathscr{W}_1^*(\cdot)$, $\mathscr{W}_2^*(\cdot)$ and $\mathscr{W}^*(\cdot)$, respectively. Point-wise and simultaneous confidence intervals for $\mathscr{D}_1(\cdot)$, $\mathscr{D}_2(\cdot)$, and $\Delta_0(\cdot)$ can then be constructed based on the crossvalidated estimates and their large sample distributions accordingly.

3. Examples

We use two examples to illustrate the new proposals. The first example is from a clinical trial conducted by the AIDS Clinical Trials Group, ACTG 320 (Hammer et. al., 1997). The study demonstrates that for various response endpoints, on average the three-drug combination therapy consisting of indinarvir, zidovudine and lamivudine, is much better than the two drug combination without indinarvir for treating HIV infected patients. Unfortunately, even with this potent combination, some patients may not respond to treatment, but suffer from non-trivial toxicity. Therefore, for future patients' management, it is important to have a reliable model for predicting patient's treatment responses based on certain "baseline" markers. A general conception is to use the baseline CD4 count and HIV RNA, a measure of viral load, and the early changes of these two markers after initiation of therapy for treatment guidance (Demeter et. al., 2001). For resource-limited regions, however, the cost of obtaining RNA is relatively expensive. Therefore, a challenging question is when we need RNA in addition to CD4 for better prediction of patient's response.

Recently Tian et. al. (2007) demonstrated that, on a population average sense, neither the baseline nor early RNA change (from baseline to week 8) would add a clinically meaningful value for predicting the long term change of CD4 (from baseline to Week 24), an important measure of the patient's immune response. Here, we try to locate a subgroup of patients, if any, who would benefit from the expensive marker RNA. To this end, let the response Y be the change of CD4 cell counts from Week 0 to 24, let U consist of age, baseline CD4 and the early change in CD4, and let V consist of the baseline RNA and the early change in RNA. For our analysis, in Models (2.1) and (2.4), we let $X = (1, U')'$, $W = (1, U', V')'$, and $g_1(\cdot)$ and $g_2(\cdot)$ be the identity function. Also, we let $\hat{Y}_1(\beta'X) = \beta'X$, $\hat{Y}_2(\theta'W) = \theta'W$, $D(a, b) = |a - b|$ and interval J_z be $[z - 10, z + 10]$ for $z \in \hat{\Omega} = [15, 165]$. Note that the intra-patient standard deviation of the CD4 count is about 60. Therefore, a choice of J_z whose length of the similar magnitude to 60 would be appropriate from a practical point of view. Moreover, in our analysis, we let $d_1 = d_2 = 0.01$ discussed in Section 2 for choosing $\hat{\Omega}$. With $n = 392$ sets of complete observations of (Y, U, V), the regression parameter estimates for Models (2.1) and (2.4) are reported in Table 1. Note that the short term changes of CD4 and RNA are statistically highly significant.

For both working models, we utilized 5-fold crossvalidation scheme discussed in Section 2 to obtain the regression parameters and then $\tilde{\mathscr{D}}_1(\cdot)$, $\tilde{\mathscr{D}}_2(\cdot)$, and $\tilde{\Delta}(\cdot)$. In Figure 1, we present these estimated prediction errors and their differences with the corresponding 0.95 point-wise and simultaneous confidence intervals given in (2.10)–(2.13). The values of $\{\tilde{\mathscr{D}}_1(z)\}$ based on the model with age, baseline CD4 and early change in CD4 range from 37 to 74. The values of $\{\tilde{\mathscr{D}}_2(z)\}$ based on the model with additional RNA information range from 36 to 73. The estimated differences $\{\tilde{\Delta}(z)\}$ range from −1.7 to 6.0. These indicate that there is no clinically meaningful gain from RNA for any subgroup of patients classified by $\hat{\beta}'X$. One may draw further statistical inference about the $\Delta_0(\cdot)$. For example, for subjects whose score $g_1(\hat{\beta}'X) \in J_z = [40, 60]$, the estimated $\tilde{\Delta}(50) = 0.45$ with 0.95 point-wise interval of (−3.25, 4.15) and simultaneous interval of (−7.48, 8.38). Note that the results reported here are based on J_z with interval length of 20, which is well within the intra-patient variation of CD4 measures. Various analyses have also been done with J_z's whose lengths range from 30 to 60. All the results lead to the same conclusion. That is, statistically or clinically, we cannot identify a subgroup of patients who would benefit from the extra information of RNA for prediction of the long term CD4 change.

The data for the second example is from a population of patients screened for a clinical study, called TRACE, for treating heart failure or acute myocardial infraction (MI) (Kober et. al., 1995). There were 6676 patients screened. Each patient had six routine clinical covariates: age, creatine (CRE), occurrence of heart failure (CHF), history of diabetes (DIA), history of hypertension (HYP), and cardiogenic shock after MI (KS). Moreover, each patient had an echocardiographic assessment of left ventricular systolic function which was quantified by a measure called the wall motion index (WMI). Compared with the above six covariates, the WMI is relatively expensive to obtain. Although not every screened patient entered the clinical trial, all patients screened were followed closely for mortality.

Recently, Thune et. al. (2005) studied the prognostic importance of left ventricular systolic function in patients diagnosed with either heart failure or acute MI in addition to the patient's medical history. It would be interesting to identify subpopulations that can benefit from the extra WMI measure for predicting clinical outcomes such as mortality. Here, we let the outcome Y be a binary variable, which is one if the patient died within five years. The five-year survival rate for this data set is approximately 42%. To evaluate the incremental value of WMI, we first fit the data using Model (2.1) with $X = (1, \text{AGE}, \text{CRE}, \text{CHF}, \text{DIA}, \text{HYP}, \text{KS})$, and $g_1(s) = \exp(s)/\{1 + \exp(s)\}$. With the extra variable WMI, we fit a second logistic regression model with $W = (X', \text{WMI})'$. A total of 5921 subjects have complete predictor information. The estimates for the regression parameters with their standard errors are reported in Table 2. Note that the WMI is highly statistically significant.

Here, we consider the prediction rules

$$\widehat{Y}_1(\beta' X) = I\{g_1(\beta' X) \geq c\}, \tag{3.1}$$

and

$$\widehat{Y}_2(\theta' W) = I\{g_2(\theta' W) \geq c\}. \tag{3.2}$$

Note that we have also fitted the data with more complicated models, for example, by including various interaction terms. The results for the present case, however, are almost identical to that based on the above two additive models. For binary response variables, we consider the distance function $D(a, b) = |a - b|$, where a is the observed response and b is a binary predicted response based on the working model with the assumed event rate of $g(\cdot)$. This distance function is a conventionally used metric for evaluating the binary prediction rules. One may use other possible distance functions, for instance, by letting b be $g(\cdot)$ from (2.1) and (2.4) and $D(a, b) = |a - b|$ or $(a - b)^2$. Moreover, one may consider a likelihood-based criterion to evaluate Models (2.1) and (2.4). In general, the metric comparing Y and the estimated $g(\cdot)$ can efficiently discriminate two prediction models. On the other hand, for the present case, we are more interested in evaluating the practical performance of the specific prediction rules, not the adequacy of the model fitting (although these two are closely related). Thus, considering distance functions between a practically applicable prediction rule $\hat{Y}$ and the true response Y seems more relevant. The distance function, $|a - b|$, consists of two discordance rates or two types of error rates. Specifically, $\mathscr{D}_1(z)$ in (2.3) is $\mathscr{D}_{11}(z) + \mathscr{D}_{10}(z)$, where $\mathscr{D}_{11}(z) = E[Y^0 D\{1, \widehat{Y}_1(\beta_0' X^0)\} | \, g_1(\beta_0' X^0) \in J_z]$ and $\mathscr{D}_{10}(z) = E[(1 - Y^0) D\{0, \widehat{Y}_1(\beta_0' X^0)\} | \, g_1(\beta_0' X^0) \in J_z]$ are the discordance rates for false negative and false positive errors, respectively. Similarly, $\mathscr{D}_2(z) = \mathscr{D}_{20}(z) + \mathscr{D}_{21}(z)$. Let $\Delta_0(z) = \mathscr{D}_{10}(z) - \mathscr{D}_{20}(z)$ and $\Delta_1(z) = \mathscr{D}_{11}(z) - \mathscr{D}_{21}(z)$. Oftentimes, a false negative error is more serious than a false positive one. Therefore, one may consider a weighted sum $\Delta_0(w, z) = w_0 \Delta_0(z) + w_1 \Delta_1(z)$ to evaluate the prediction rules. Here, $w = (w_0, w_1)'$ and the weights w_0 and w_1 reflect the "cost" of making these two types of errors. It is interesting to note that the corresponding distance function for $\Delta_0(w, z)$ is $w_0^{1-Y} w_1^Y |Y - \widehat{Y}|$. For a given w, the crossvalidated point estimates $\tilde{\Delta}(w, z)$ and their interval estimates for $\underset{\sim}{\Delta}_0(w, z)$ can be constructed as for $\Delta_0(z)$ in Section 2. The large sample properties for $\Delta(w, z)$ are derived in Appendix A of the web-based supplementary material.

We first considered the most commonly used prediction rule with $c = 0.5$ and $w_0 = w_1 = 1$. The 5-fold crossvalidated estimates, obtained by letting J_z be the entire real line in (2.8) and (2.9), for the overall prediction errors $E[D\{Y^0, \hat{Y}_1(\hat{\beta}' X^0)\}]$ and $E[D\{Y^0, \hat{Y}_2(\hat{\beta}' W^0)\}]$ are 0.28 and 0.26, respectively, a modest overall incremental gain from the extra information of WMI for the entire population of interest. To identify which subgroup of patients who would benefit with WMI, we let $J_z = [z - 0.1, z + 0.1]$, for $z \in \hat{\Omega} = [0.15, 0.82]$. Here, the scale of z is the probability of developing the event within five years based on the conventional risks factors in the model. The choice of the interval length of J_z is not obvious as that for the HIV example and should be made by the entire research team (not only from the statistical point of view). For example, it may depend on the cost of obtaining the WMI measure, the distribution of the initial predicted risk score, and the clinical interpretation of the scale of such a scoring system. In our analysis, $\hat{\Omega}$ is chosen by letting $d_1 = d_2 = 0.01$ discussed in Section 2. To estimate $\mathscr{D}_l(z)$, $l = 1, 2$, and $\Delta_0(z)$, we used the 5-fold crossvalidation to obtain $\tilde{\mathscr{D}}_1(\cdot)$, $\tilde{\mathscr{D}}_2(\cdot)$ and $\tilde{\Delta}(\cdot)$. In Figure 2, we present these point estimates and their corresponding 0.95 point-wise and simultaneous confidence intervals. For the interval estimation, we let $d_3 = 0.01$. This results in $\tilde{\Omega} = [0.26, 0.76]$. Note that the point estimates $\tilde{\Delta}(z)$ for z outside $\hat{\Omega}$ are not reliable, and $\tilde{\Delta}(z)$ is pretty at around 0 for $z \in \hat{\Omega} - \tilde{\Omega}$, indicating that there is no evidence that WMI has a meaningful gain outside the interval $\tilde{\Omega}$. On the other hand, with the point and interval estimates displayed in Figure 2(c), one may conclude that WMI is likely to be beneficial for patients with conventional risk scores $g_1(\hat{\beta}'X)$ ranging from 0.16 to 0.74. If WMI is relatively affordable to the population of interest, then one may consider using the upper bound of the simultaneous confidence intervals to identify the subpopulation based on $\tilde{\Delta}(z)+n^{-\frac{1}{2}}\tau_\alpha\sigma_{\mathscr{W}^*(z)} \geq 0$ 0 and thus conclude that patients with $g_1(\hat{\beta}'X) \in [0.16, 0.86]$ are likely to benefit from the WMI. On the other hand, when WMI is not quite affordable, then one may select the region conservatively and use the lower bound of the simultaneous confidence intervals based on $\tilde{\Delta}(z) - n^{-\frac{1}{2}}\tau_\alpha\sigma_{\mathscr{W}^*(z)} \geq 0$ and thus conclude that patients with $g_1(\hat{\beta}'X) \in [0.29, 0.63]$ are likely to benefit from the WMI.

To illustrate the effect of the weighting parameter $w = (w_0, w_1)$ on the incremental value of WMI, we present in Figure 3(a),(b),(c) and (d) the point and interval estimates of $\Delta_0(w, z)$ for the predictors (3.1) and (3.2) with $c = 0.5$ and various choices of w.

Note that when $w_0 \neq w_1$, even if the working model is correctly specified, the prediction rule in (3.1) or (3.2) with $c = 0.5$ is not optimal with respect to the weighted error rate. Furthermore, with the unequal weighting criterion, for some subgroups of patients, inclusion of the extra information of WMI may significantly decrease the prediction precision. For a given w, with the weighted sum prediction precision measure, $w_0\mathscr{D}_{10}(z) + w_1\mathscr{D}_{11}(z)$, it is straightforward to show that the optimal prediction rule based on X that minimizes the above criterion is $\hat{Y} = I\{\text{pr}(Y = 1 \mid X) \geq c_w\}$, where $c_w = w_0/(w_0 + w_1)$. Therefore, for the present example, if $g_1(\hat{\beta}'X)$ and $g_2(\hat{\theta}'W)$ are reasonably good approximations to $E(Y|U)$ and $E(Y|U, V)$, the predictors $I(g_1(\hat{\beta}'X) \geq c_w)$ and $I(g_2(\hat{\theta}'W) \geq c_w)$ are almost optimal. In Figure 4, we present the crossvalidated point estimates along with the 0.95 interval estimates of $\Delta_0(w, z)$ with $w = (1, 4)'$ and $(1, 9)'$ when “optimal” prediction rules are used for both models. It appears that there is minimal gain from WMI across all sub-populations indexed by $g(\hat{\beta}'X) \in J_z$ for both cases. These findings underscore the importance of selecting the cut-off value as well as the distance measure for quantifying the incremental predictive value of new biomarkers. In any event, we highly recommend to perform such sensitivity analyses to provide an over-all picture of the incremental value from the new biomarkers when there is no consensus about the weights used in the binary classification.

As suggested by a reviewer of the paper, we have also applied our method to quantify the incremental value of new biomarkers based on the integrated discrimination improvement (IDI) index (Pencina et. al., 2008, Pepe et. al., 2008). The IDI index is defined as the integrated difference in Youden’s indices (Youden, 1950). It can be viewed as an improvement of the average sensitivity and specificity with the new markers. The crossvalidated point estimates and the corresponding 0.95 simultaneous interval estimates are presented in Figure 1 of the web-based supplementary material. With this utility function, it appears that the WMI would be beneficial for all patients whose conventional risk scores are between 0.10 and 0.81 for predicting subject’s five year survival. It is important to note that there are no specific prediction rules attached to this approach (similar to the area under curve as an average measure over a class of prediction rules). Therefore, at the end it is not clear which prediction rule(s) one would recommend for practical usage based on the model with the additional markers. Utilizing an over all measure of the incremental value for evaluating new markers over various subgroups of patients can be useful at the initial stage of the investigation.

4. Remarks

As for any scientific investigation, we need to define the endpoint of the study at the very first step. For the HIV example in Section 3, instead of using the change of the CD4 counts as the endpoint, one may be interested in the percent of change in CD4 counts from the baseline level. This is equivalent to considering the change in the log transformed CD4 counts as the endpoint. For the TRACE study example in Section 3, we dichotomized the patient's survival time to make the response variable being the binary survival status at 5 years in our analysis. Naturally we may consider prediction rules for other time points. A good prediction rule for five-year survival may not be appropriate for ten-year survival. Therefore, when evaluating prediction rules for the patient's survival time, a global distance function, for example, based on the L_1 norm, between the observed and predicted survival times may not be desirable. Moreover, in practice, often the survival time cannot be observed completely and the support of its censoring variable is generally shorter than that of the survival time. As a result, it is difficult to evaluate prediction rules efficiently with, for example, the L_1 distance without artificial extrapolation. On the other hand, using the approach taken by Uno et. al. (2007), one may identify patients who would benefit from the additional biomarkers via prediction of t-year survival.

After we select the endpoint of the study, we should make every effort to find the "best" models (2.1) and (2.4), which fit the data well. Model (2.1) would thus provide us a reasonable scoring system with which we can classify future patients into different subgroups via the conventional markers. Subsequently, we fit the response variable with the conventional and new marker values jointly to build model (2.4). Such an elaborate joint model may include interactions between the conventional and new markers, which would be potential contributing factors to the varying incremental values from the new markers.

The next crucial step is to choose a proper prediction precision measure to quantify the incremental value of the new markers. Different distance functions between the predicted and observed may result in quite different conclusions regarding the selection of subset of patients as illustrated by two real examples in Section 3. Since the final decision of using the new biomarkers would be based on the trade off between the risk/cost and benefit, the distance function needs to be "clinically" or heuristically interpretable. In analyzing the data from the HIV and TRACE trials, we proposed several distance functions for illustration. The choice of such functions is by no means restricted to those discussed in Section 3. For example, one may use a theoretically interesting metric such as the conventional likelihood ratio statistic or the mean square error loss to differentiate two prediction models (one with and the other without new markers). However, the magnitude of gain under such a metric is often difficult to interpret when the cost or risk is involved for decision making.

In practice the choice of the distance function even for the simple case with a binary response is rather complex. In the cardiovascular disease arena, conventionally a patient with more than 10% risk for having a serious cardiovascular event within ten years is generally regarded as having a high risk, and usually would be recommended for certain preventive treatments. However, the utility function may vary across individuals and hence different patients may have different optimal cutoff points for predicting patient-level outcomes. The weighted sum of prediction error rates presented in this article is an attempt to cope with this complicated cost-benefit issue. The complexities of choosing a distance function extend to the case with continuous responses. For example, weighting absolute prediction errors according to the observed response may lead to a more meaningful penalty in some applications compared to the un-weighted counterpart.

In this paper, we provide a useful tool for making valid inferences on the incremental value of new markers simultaneously over subsets of patients with well-defined endpoint, prediction models, utility (distance) function and the study population. We propose the use of the simultaneous confidence band for the incremental values to control the type one error and determine whether the new markers have a positive incremental value in a subset of patients. As the sample size increases, the confidence bands become tighter and one would be able to more accurately identify all subsets of interest.

Lastly, we may want to identify subsets where the incremental value is greater than a positive threshold to incorporate the cost of measuring the new markers. As such, subsets where the new markers have a positive, yet very small added predictive value would be excluded. If the new markers become less costly or invasive in the future, we may construct a new scoring system to index patients. It is likely that some old markers may not be needed on top of the new ones.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to Drs. Jens Jakob Thune and James Signorovitch for helpful comments on the paper. The authors also very grateful to two reviewers and an associate editor for their insightful comments on the paper. This research was partially supported by the US NIH grants for I2B2, AIDS and General Medical Sciences.

References

Blumenthal R, Michos E, Nasir K. Further improvements in CHD risk prediction for women. JAMA 2007;297:641–643. [PubMed: 17299201]

Breiman, L.; Friedman, J.; Stone, C.; Olshen, R. Classification and regression trees. Chapman & Hall/CRC; 1984.

Brier G. Verification of forecasts expressed in terms of probability. Monthly Weather Review 1950;78:1–3.

D'Agostino RB. Risk prediction and finding new independent prognostic factors. Journal of Hypertension 2006;24:643–645. [PubMed: 16531791]

Demeter L, Hughes M, Coombs R, Jackson J, Grimes J, Bosch R, Fiscus S, Spector S, Squires K, Fischl M, Hammer S. Predictors of virologic and clinical outcomes in HIV-1-infected patients receiving concurrent treatment with indinavir, zidovudine, and lamivudine. AIDS Clinical Trials Group Protocol 320. Annals of Internal Medicine 2001;135:954–964. [PubMed: 11730396]

Hammer S, Squires K, Hughes M, Grimes J, Demeter L, Currier J, Eron J, Feinberg J, Balfour H, Deyton L, Chodakewitz J, Fischl M. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. New England Journal of Medicine 1997;337:725–733. [PubMed: 9287227]

Kober L, Torp-Pedersen C, Carlsen J, Bagger H, Eliasen P, Lyngborg K, Videbak J, Cole D, Auclert L, Pauly N, Aliot E, Persson S, Camm A. A clinical trial of the angiotensin-converting-enzyme inhibitor trandolapril in patients with left ventricular dysfunction after myocardial infarction. New England Journal of Medicine 1995;333:1670–1676. [PubMed: 7477219]

Korn E, Simon R. Measures of explained variation for survival data. Statistics in Medicine 1990;9:487–503. [PubMed: 2349402]

McLachlan, J. Discriminant analysis and statistical pattern recognition. John Wiley & Sons; 1992.

Mittlböck M, Schemper M. Explained Variation for Logistic Regression. Statistics in Medicine 1996;15:1987–1997. [PubMed: 8896134]

Pencina J, D'Agostino R Sr, D'Agostino R Jr, Vasan R. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in Medicine 2008;27:157–172. [PubMed: 17569110]

Pepe, MS. The statistical evaluation of medical tests for classification and prediction. Oxford University Press; 2003.

Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. American Journal of Epidemiology 2004;159:882–890. [PubMed: 15105181]

Pepe MS, Feng Z, Gu JW. Comments on Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond by M. J. Pencina et al. Statistics in Medicine 2008;27:173–181. [PubMed: 17671958]

Ridker P, Rifai N, Rose L, Buring J, Cook N. Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. New England Journal of Medicine 2002;347:1557–1565. [PubMed: 12432042]

Ridker P, Buring J, Rifai N, Cook N. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds risk score. JAMA 2007;297:611–619. [PubMed: 17299196]

Ripley, B. Pattern recognition and neural networks. Cambridge University Press; 1996.

Speigelhalter D. Probabilistic predictin in patient management and clinical trials. Statistics in Medicine 1986;5:421–433. [PubMed: 3786996]

Thune J, Carlsen C, Buch P, Seibak M, Burchardtb H, Torp-Pedersen C, Kober L. Different prognostic impact of systolic function in patients with heart failure and/or acute myocardial infarction. European Journal of Heart Failure 2005;7:852–858. [PubMed: 15923139]

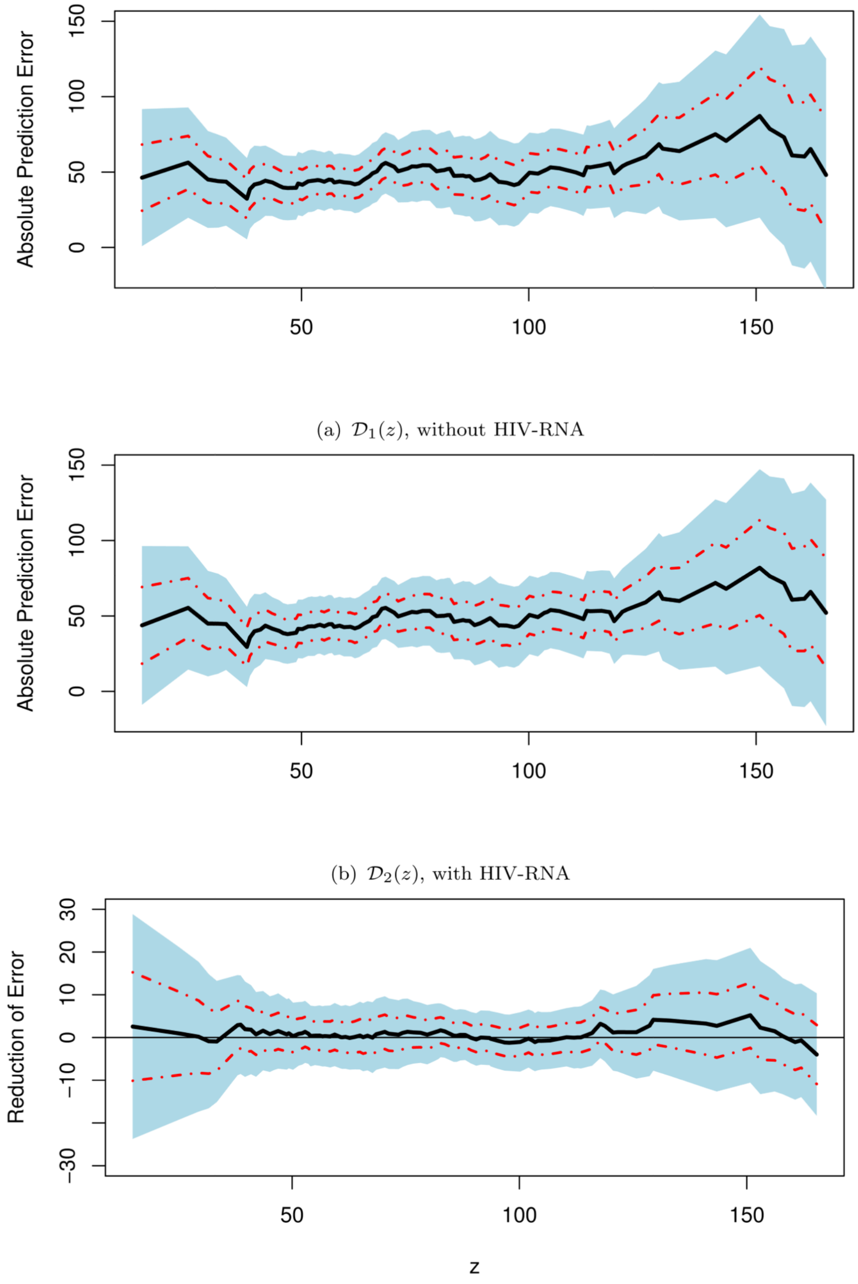
Tian L, Cai T, Goetghebeur E, Wei LJ. Model evaluation based on the distribution of estimated absolute prediction error. Biometrika 2007;94:297–311.

Uno H, Cai T, Tian L, Wei LJ. Evaluating prediction rules for t-Year survivors with censored regression models. Journal of American Statistical Association 2007;102:527–537.

Wang T, Gona P, Larson M, Tofler G, Levy D, Newton-Cheh C, Jacques P, Rifai N, Selhub J, Robins S, Benjamin E, D'Agostino R, Vasan R. Multiple biomarkers for the prediction of first major cardiovascular events and death. New England Journal of Medicine 2006;355:2631–2639. [PubMed: 17182988]

Youden W. An index for rating diagnostic tests. Cancer 1950;3:32–35. [PubMed: 15405679]

Zhou, XH.; Obuchowski, NA.; McClish, DK. Statistical methods in diagnostic medicine. John Wiley & Sons; 2002.



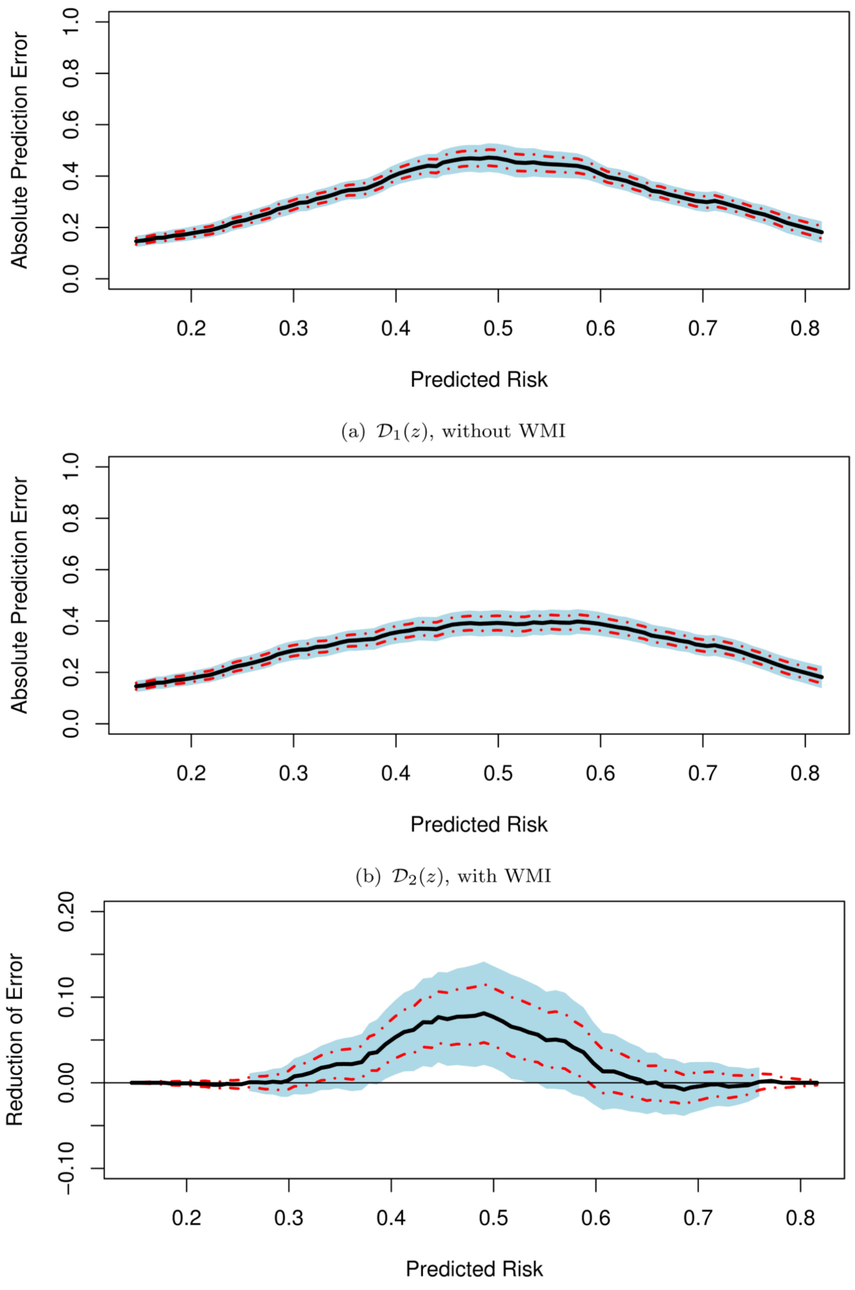
(a) $\mathcal{D}_1(z)$, without HIV-RNA

(b) $\mathcal{D}_2(z)$, with HIV-RNA

(c) $\Delta_0(z)$

Figure 1.

Point estimates for $\mathscr{D}_1(\cdot)$, $\mathscr{D}_2(\cdot)$ and $\Delta_0(\cdot)$ with corresponding 0.95 point-wise (dashed lines) and simultaneous (shaded regions) confidence intervals for the HIV example.



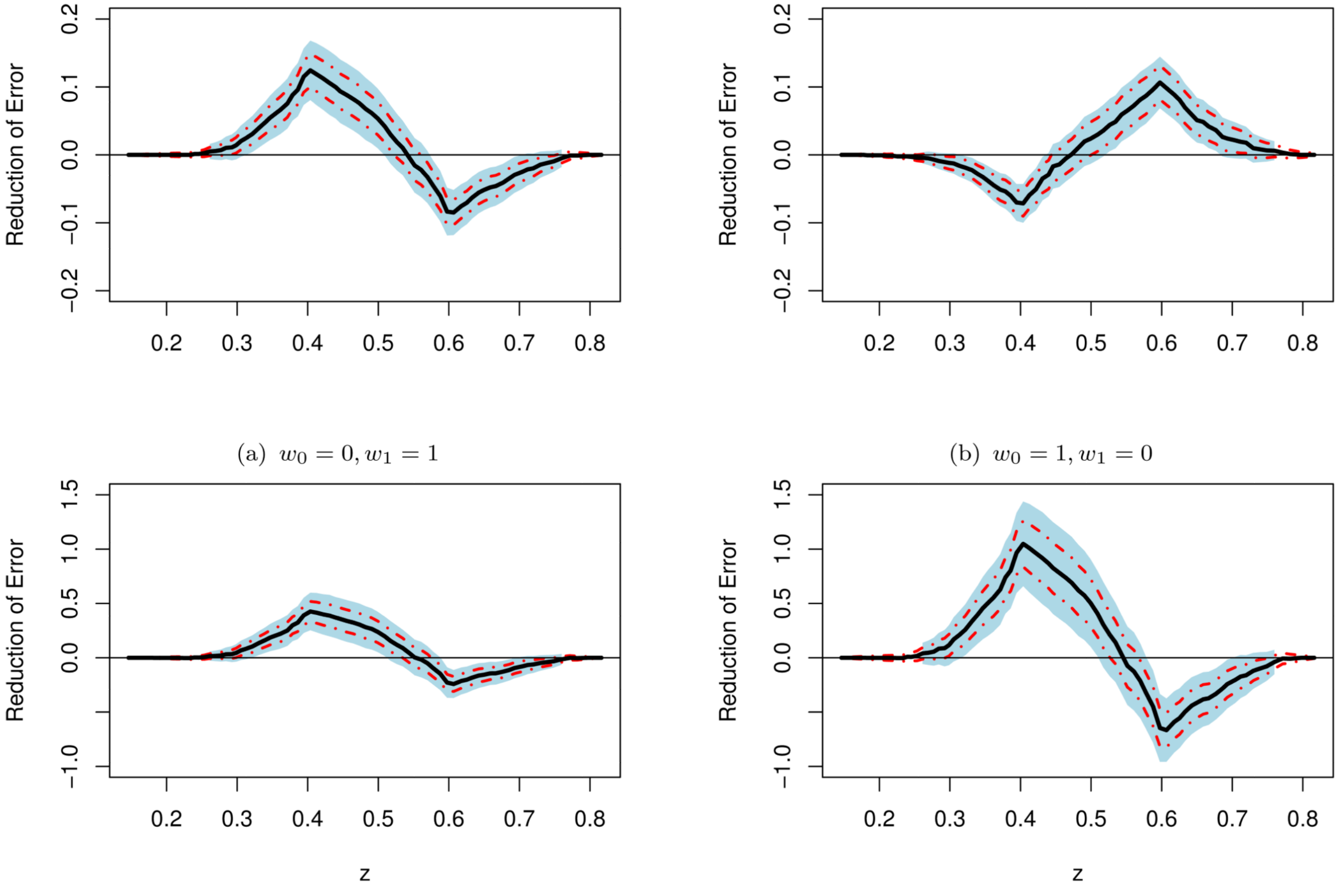
(a) $\mathcal{D}_1(z)$, without WMI

(b) $\mathcal{D}_2(z)$, with WMI

(c) $\Delta_0(z)$

Figure 2.

Point estimates for $\mathscr{D}_1(\cdot)$, $\mathscr{D}_2(\cdot)$ and $\Delta_0(\cdot)$ with corresponding 0.95 point-wise (dashed lines) and simultaneous (shaded regions) confidence intervals for the screened population of the TRACE study (the prediction with $c = 0.5$).



(a) $w_0 = 0, w_1 = 1$

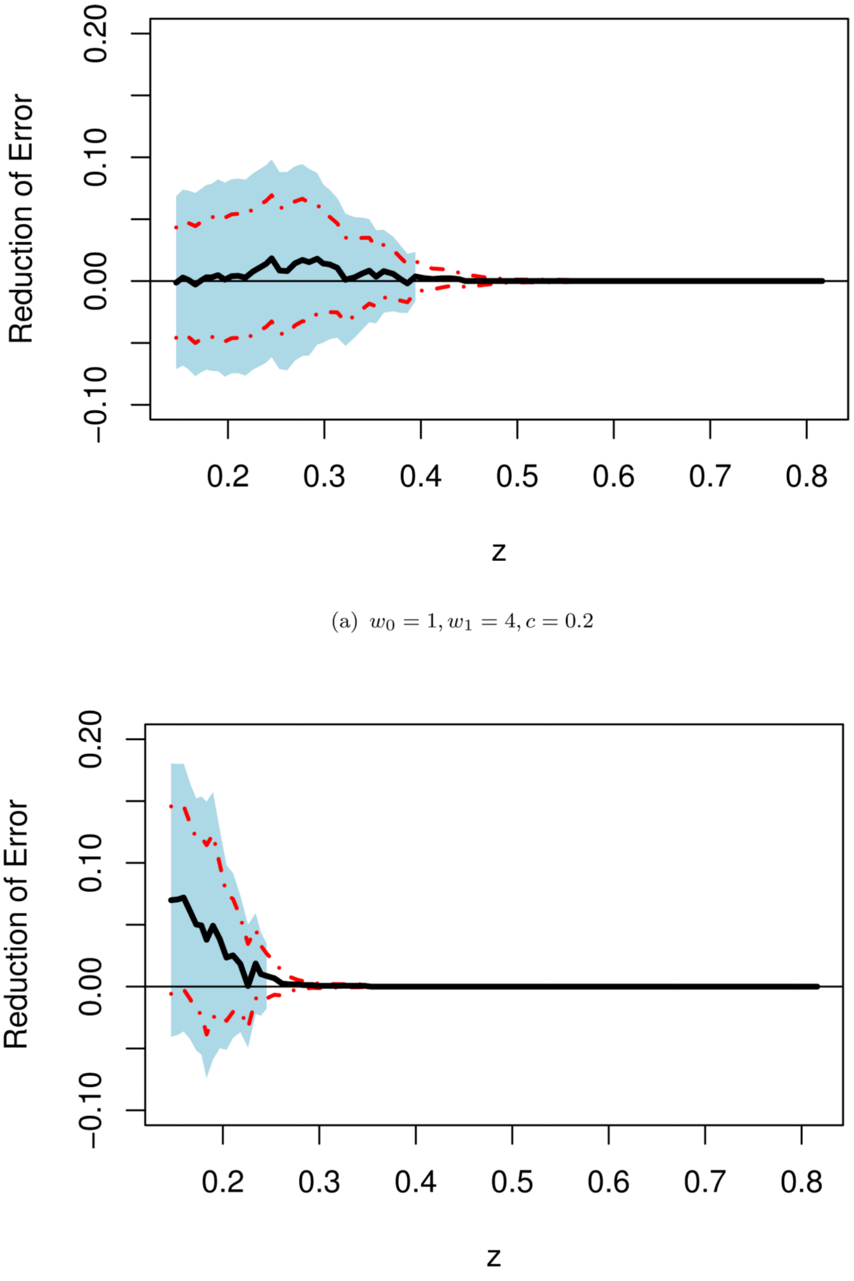
(b) $w_0 = 1, w_1 = 0$

(c) $w_0 = 1, w_1 = 4$

(d) $w_0 = 1, w_1 = 9$

Figure 3.

Point estimate $\tilde{\Delta}(w, \cdot)$ for $\Delta_0(w, \cdot)$ with various weights and the corresponding 0.95 pointwise (dashed lines) and simultaneous (shaded regions) confidence intervals for the screened population of the TRACE study (the prediction with $c = 0.5$).



(a) $w_0 = 1, w_1 = 4, c = 0.2$

(b) $w_0 = 1, w_1 = 9, c = 0.1$

Figure 4.

Point estimate $\tilde{\Delta}(w, \cdot)$ for $\Delta_0(w, \cdot)$ with the "optimal" weights and the corresponding 0.95 point-wise (dashed lines) and simultaneous (shaded regions) confidence intervals for the screened population of the TRACE study.

Table 1

Estimates of the regression parameters with their standard errors and corresponding p-values for testing zero covariate effects for the AIDS example

	Age	Baseline RNA	RNA Change	Baseline CD4	CD4 Change
Estimate	−0.55	0.08	−12.06	0.03	0.68
Std Error	0.35	5.53	2.80	0.07	0.10
P-value	0.12	0.99	0.00	0.72	0.00

Table 2

Estimated Regression Coefficients for Model (2.1) with AGE, CRE, CHF, DIA, HYP, KS and WMI for the screened population of TRACE study

	AGE	CRE	CHF	DIA	HYP	KS	WMI
Estimate	0.055	−0.010	0.759	0.718	0.187	1.153	−1.097
Std. Error	0.004	0.002	0.067	0.101	0.073	0.163	0.083
P-value	0.000	0.000	0.000	0.000	0.010	0.000	0.000