EVALUATING SUBJECT-LEVEL INCREMENTAL VALUES OF NEW MARKERS FOR RISK CLASSIFICATION RULE

T. Cai,
Department of Biostatistics, Harvard University, Boston, MA 02115, USA

L. Tian,
Department of Health Research and Policy, Stanford University, Palo Alto, CA 94305, USA

D. Lloyd-Jones, and
Department of Preventive Medicine, Northwestern University, Chicago, IL 60611, USA

L.J. Wei
Department of Biostatistics, Harvard University Boston, MA 02115, USA

Abstract

Suppose that we need to classify a population of subjects into several well-defined ordered risk categories for disease prevention or management with their "baseline" risk factors/markers. In this article, we present a systematic approach to identify subjects using their conventional risk factors/ markers who would benefit from a new set of risk markers for more accurate classification. Specifically for each subgroup of individuals with the same conventional risk estimate, we present inference procedures for the reclassification and the corresponding correct re-categorization rates with the new markers. We then apply these new tools to analyze the data from the Cardiovascular Health Study sponsored by the US National Heart, Lung, and Blood Institute (NHLBI). We used Framingham risk factors plus the information of baseline anti-hypertensive drug usage to identify adult American women who may benefit from the measurement of a new blood biomarker, CRP, for better risk classification in order to intensify prevention of coronary heart disease for the subsequent ten years.

Keywords

Coronary Heart Disease; Nonparametric Functional Estimation; Risk Factors/Markers and Pointwise; Simultaneous Confidence Intervals; Subgroup Analysis

1 Introduction

An integral part of evidence-based guidelines for disease prevention or management is a well-defined risk classification rule, which assigns each subject from a population of interest to one of several ordered risk groups. The assignment is based on the individual risk: the chance that the subject will experience a specific type of event during a given time period. Appropriate interventions will then be offered to subjects in each risk category. The risk is

corresponding author tcai@hsph.havard.edu.

estimated using a scoring system with the subject's "conventional" baseline risk factors/ markers. For example, in the United States more than a half million women die of cardiovascular diseases (CVD) each year and the majority of such deaths are due to coronary heart disease (CHD). Recently, the American Heart Association (AHA) issued guidelines for the prevention of CHD in adult women (Mosca et al., 2004, 2007). Specifically, risk categories are defined by the individual's predicted risk of having a CHD event in the next ten years. Adult American women are classified as being low risk ($< 10\%$), intermediate risk (between 10% and 20%), or high risk (20%). Such risk threshold values have also been employed by the Adult Treatment Panel III (ATP III) of the National Cholesterol Education Program to develop an evidence-based set of guidelines on cholesterol management (Grundy et al., 2004). For subjects with intermediate or high risk, certain life style and pharmacologic interventions are recommended. The risk is estimated using a modified version of the Framingham Risk Score (FRS), which is a multivariate risk equation based on traditional risk factors such as age, blood cholesterol, HDL cholesterol, blood pressure, smoking status and diabetes mellitus (Wilson et al., 1998).

Now, suppose that new risk markers for such future events are available, and they may potentially improve the conventional risk estimation. Measuring these markers, however, may be invasive or costly. Under various settings, novel procedures have been proposed to quantify the *overall* incremental benefit from the new markers for the entire population of interest (Pepe et al., 2004; Cook, 2007; Tian et al., 2007; Uno et al., 2007; Pepe et al., 2008; Pencina et al., 2008; Chi and Zhou, 2008; Cook, 2008; Greenland, 2008). In a recent paper, Wang et al. (2006) concluded that almost all new contemporary biomarkers for prevention of CHD added rather moderate overall predictive value to FRS.

In general, for subjects whose conventional risk estimates are either extremely low or high, clinical practitioners are less likely to request additional marker measurements for their decision making. Unfortunately, rather little research has been done for developing a systematic procedure to identify a subset of individuals who would benefit significantly from new risk markers (D'Agostino, 2006). With respect to the aforementioned risk classification rule recommended by the AHA and ATP III, recently Cook et al. (2006) and Ridker et al. (2007) examined the incremental value of a new blood biomarker – the C-Reactive protein (CRP) measured by a highly sensitive assay – with a cohort of subjects from the Women's Health Study (Buring and Hennekens, 1992). They concluded that subjects whose traditional risk scores are between 5% and 20% would benefit from this marker. Their claims were based on the reclassification rates with CRP among four discretized subgroups of the entire cohort determined by conventional risk estimates. However, large observed reclassification rates, which include both "good" and "bad" reclassification rates, does not equal to the benefit of a new marker. More recently, Prentice and his colleagues performed a comprehensive analysis on five pooled cohorts to examine the additive value of genetic variants to breast cancer risk models and conclude that "Inclusion of newly discovered genetic factors modestly improved the performance of risk models for breast cancer. The level of predicted breast-cancer risk among most women changed little after the addition of currently available genetic information."(Waholder et al., 2010). One of their main analytic strategies employed to quantify the improvement of

genetic markers is to examine how the addition of genetic data actually affected the risk stratification, i.e., the reclassification rate, for the case subjects with breast cancer and control subjects without breast cancer, separately. This can be viewed as a great improvement of the crude re-classification rate in the direction of detailed quantification of the "good" and "bad" reclassification. On the other hand, the incremental benefit of the "genetic markers" for risk stratification is evaluated for the entire population. It is useful to generalize the approach for identifying subgroup of patients, whose risk stratification may still be substantially improved by using new genetic markers in the risk prediction model (Mealiffe et al., 2010). Moreover, all the existing summary measures related to reclassification rate are simply point estimators lacking rigorous statistical inference procedure to provide appropriate precision measures.

In this article, we propose a systematic approach for evaluating the incremental value of a novel marker in risk stratification across subgroups of subjects and identify subjects who may benefit the most from the information of the new marker. To this end, for each subject in the study we construct two individual risk estimates, one based on the conventional risk markers and the other based on both the conventional and new markers. A subject may be reclassified into a different risk category with the new risk estimate. For the subset D of subjects with the same conventional risk estimate, we obtain consistent nonparametric functional estimates for the reclassification rates and their corresponding standard error estimates. Note the limitation that the large observed reclassification rates do not automatically imply that the new markers are important for given subsets of subjects. Consistent estimates for the proportions of subjects who are reclassified correctly are also needed for cost-benefit decision making. In this paper, we develop point and interval estimation procedures for making inference about such proportions.

As an illustration, consider the set D consisting of individuals who have the same conventional FRS estimate of 9% for experiencing CHD events within ten years. Based on the aforementioned prevention guidelines, individuals with this predicted risk would not be recommended for lipid-lowering drug therapy at present. Now, suppose that with the new markers, the estimate for the proportion of subjects in D reclassified to the next higher risk category is 25%. Furthermore, suppose that among those who are reclassified to the next higher risk category, the CHD events rate is 12%. The potential benefit with the new markers would be preventing $12\% \times 25\% = 3\%$ of subjects in D from having future CHD events. However, the costs consist of measuring new marker values for every subject in this subset, and giving possibly long-term, potentially toxic interventions to 25% of the individuals. Even if we decide that for this scenario the benefit outweighs the cost, we still need to know whether this reclassification scheme is better than a "random allocation" process. If we randomly move 25% of subjects in D to the next higher risk class, on average, 9% of these subjects would have CHD events. The question is whether the difference between the observed 12% and the null value of 9% represents a real gain or is simply due to sampling variation.

The new proposal is illustrated with a data set from the Cardiovascular Health Study (CHS) sponsored by the US National Heart, Lung and Blood Institute (Fried et al., 1991). This study is a prospective, population-based, long term follow-up cohort study to determine risk

factors for predicting future coronary heart disease and stroke in adults 65 years or older. There were 5888 study subjects recruited between 1988 and 1993. For our analysis, we only considered data from the 3393 female participants. For each subject, we utilized the risk factors/markers values at her entry time to the study and her CHD event time by year 2003. The binary response variable is whether the subject experienced a CHD event (non-fatal MI, angina or CHD-related death) during ten year follow-up. For this data set, the median age at the baseline is 72.5 and there is no loss of follow-up for these endpoints. During the first 10 year follow-up, 19.5% of female participants experienced non-fatal MI, angina or CHD-related death. On the other hand, 37% of them died from non-CHD-related diseases. Among 3393 subjects, 52% of them were surviving in year 2003. In our analysis, the conventional risk factors consist of the Framingham risk factors and the anti-hypertensive drug usage. The new marker considered here is CRP. For the risk classification rule recommended by the ATP III, the CRP provides *pointwise* significant incremental values for subjects whose conventional risk estimates are around 10% and 20%. On the other hand, we cannot find any subgroup of individuals who would benefit from the new marker under the *simultaneous* inference setting when controlling the overall family-wise type I error rate of 0.05.

In the next section, we describe our proposed procedures for quantifying the incremental value of new markers. Procedures for making inference about the incremental values were proposed in Section 3. In section 4, we provide details of the analysis of the CHS data and of a simulation study for examining the performance of our procedures. Concluding remarks are given in Section 5.

2 Quantifying Incremental Values from New Markers

Consider a subject randomly drawn from the study population. Let Y be its binary response variable with $Y = 1$ denoting "cases" and $Y = 0$ denoting "controls", U be the set of conventional markers and V be the set of new markers. Let $p(U) = \mathrm{pr}(Y = 1| U)$ and $p(U, V) = \mathrm{pr}(Y = 1| U, V)$, the risks of this subject conditional on U and $\{U, V\}$, respectively. Assume that the classification rule assigns each subject to one of L ordered risk categories $\{C_1, \cdots, C_L\}$. A subject is classified to C_l if its risk for $Y = 1$ is in the interval $[\nu_{l-1}, \nu_l)$, where $l = 1, \cdots, L$, and $0 = \nu_0 < \cdots < \nu_L = 1$. Based on $p(U, V)$, the subject may be reclassified into a higher or lower risk category.

Now, suppose that the data $\{(Y_i, U_i, V_i), i = 1, \cdots, n\}$ consist of n independent copies of (Y, U, V). For any given (u, v), the probabilities $p(u)$ and $p(u, v)$ cannot be consistently estimated well with fully non-parametric procedures unless the dimension of U or V is very low. A practical alternative is to consider a working model for approximating $p(U)$ by a parametric model

$$g_1 \left(\beta' X \right), \quad (2.1)$$

where X, a $p \times 1$ vector, is a function of U, β is an unknown vector of parameters and $g_1(\cdot)$ is a known, smooth, increasing function. Now, suppose that β is estimated by $\hat{\beta}$ via an estimating function $S_1(\beta)$. The risk for a subject with $U = u$, whose $X = x$, is estimated by $\hat{p}_u = g_1\left(\hat{\beta}'x\right)$.

With the additional marker set V, consider the working model for approximating $p(U, V)$ by

$$g_2\left(\gamma'W\right), \quad (2.2)$$

where W, a $r \times 1$ vector, is a function of U and V, $g_2(\cdot)$ is a known, smooth, increasing function, and γ is a vector of unknown parameters. Let γ be estimated by $\hat{\gamma}$ via an estimating function $S_2(\gamma)$. For a subject with $(U, V) = (u, v)$ whose $W = w$; the estimated risk is $\hat{p}\{u, v\} = g_2\left(\hat{\gamma}'w\right)$. Note that when Models (2.1) and (2.2) are correctly specified, with reasonable estimating functions $S_1(\cdot)$ and $S_2(\cdot)$, $\hat{p}_u$ and $\hat{p}\{u, v\}$ are consistent estimators for $p(u)$ and $p(u, v)$, respectively.

Now, consider a random future subject with $(Y, U, V) = (Y^0, U^0, V^0)$. For a given $0 < s < 1$, let D_s be the group of future subjects whose $\hat{p}_U^{\,0} = s$ and $l_s = \sum_{l'=1}^{L} I\left(v_{l'} \leq s\right)$ be the corresponding risk category. For subjects in D_s with additional new marker information, the probability of a subject's being classified to C_l with $Y^0 = q$, is

$$\eta_l^{(q)}(s) = pr\left\{\hat{p}\left\{U^0, V^0\right\} \in (v_{l-1}, v_l), Y^0 = q | \hat{p}_{U^0} = s\right\}, \qquad for\, q = 0, 1; l = 1, \cdots L. \quad (2.3)$$

Here, the conditional probabilities are with respect to the data $\{(Y_i, U_i, V_i)\}$, Y^0, U^0 and V^0. The probability of a subject being classified to C_l is $\eta_l(s) = \eta_l^{(0)}(s) + \eta_l^{(1)}(s)$. For the subgroup with $\hat{p}_{U^0} = s$, the probability of crossing risk boundaries or being reclassified into other risk categories is $\mathscr{P}(s) = \sum_{l \neq l_s} \eta_l(s)$. However, it is important to note that a large value of the reclassification rate $P(s)$ does not automatically imply that the new markers are valuable for D_s. The quantity in (2.3) also plays an important role in cost-benefit decision making. For subjects who are re-assigned to a higher risk category, $\left\{\eta_l^{(1)}(s), l > l_s\right\}$ provides us with the proportion of subjects in D_s who would benefit from the new markers. On the other hand, for subjects who are moved down to a lower risk group, $\{\eta_l^{(0)}(s), l < l_s\}$ would reflect the incremental gain. They are the generalizations of the net reclassification improvement (NRI) of the subpopulation D_s:

$$\sum_{l=1}^{L} \eta_l^{(1)}(s) / \xi(s)\, sign\left(l - l_s\right) - \sum_{l=1}^{L} \eta_l^{(0)}(s) / \{1 - \xi(s)\}\, sign\left(l - l_s\right)$$

and provide detailed information on the improvements by including the new biomarker,

where $\xi\left(s\right)=pr\left(Y^0=1|\hat{p}_{U^0}=s\right)=\sum_{l=1}^{L}\eta_l^{(1)}\left(s\right).$

Moreover, it is important to know whether the above reclassification is more than a purely random allocation process. Specifically, if the new markers contribute nothing to the classification rule, one would expect that conditional on $\hat{p}_{U^0}=s$, $\hat{p}_{\{U^0,V^0\}}$ 0 is independent of Y^0. That is,

$$\eta_l^{(1)}\left(s\right)=\eta_l\left(s\right)\times\xi\left(s\right),\quad l=1\cdots,L.\quad(2.4)$$

The differences

$$\left\{D_l\left(s\right)=\eta_l^{(1)}\left(s\right)-\eta_l\left(s\right)\xi\left(s\right),\quad l=1,\cdots,L\right\}\quad(2.5)$$

also play important roles for quantifying the incremental value of the new markers for future subjects in D_s. If V^0 improves the risk stratification for subjects in D_s, we expect $D_l(s) > 0$ for $l > l_s$ and $D_l(s) < 0$ for $l < l_s$. To summarize the overall gain of V^0 in reclassification for D_s, a simple measure would be

$$D_+\left(s\right)=\sum_{l=1}^{L}sign\left(l-l_s\right)D_l\left(s\right),$$

which represents the extent to which V^0 results in correct reclassifications over and above random reallocations.

3 Estimating $\eta_l^{(q)}\left(s\right)$ with Possibly Censored Event Time Observations

Oftentimes the binary response Y indicates that either the event time of interest is greater or less than a specific time point via a long term follow-up study. To this end, let T_0 be the random event time and t_0 be a prespecified time point. Here, the binary variable $Y = 1$, if $T_0 < t_0$, 0, otherwise. The "response" Y may not be observed directly due to censoring. That is, T_0 may be censored by a random variable C. Let $G(\cdot)$ be the survival function of C. In this article, we assume that $G(\cdot)$ is independent of T_0, U and V. For T_0, one can only observe $T = $ min(T_0, C) and $\Delta = I(T_0 \leq C)$, where $I(\cdot)$ is the indicator function. Our data $\{(T_i, \Delta_i, U_i, V_i), i = 1, \cdots, n\}$ consist of n independent copies of (T, Δ, U, V). Note that if there is no censoring involved, $G(\cdot) = 1$ and $\{Y_i, i = 1, \cdots, n\}$ can be observed completely.

To obtain the estimates $\hat{p}_u$ and $\hat{p}\{u, v\}$, one may assume proportional hazards *working* models for T_0 with U, and then for T_0 with U and V. With the maximum partial likelihood estimates for the regression coefficients and the estimates for the underlying cumulative hazard functions, we can estimate $p(u)$ and $p(u, v)$. However, if a working proportional model is not correctly specified, the regression coefficient estimator converges to a constant

vector, as $n \to \infty$, which may depend on the censoring distribution. Moreover, a good prediction model for short term survivors may not work well for predicting long term survivors. In this article, instead of modeling the entire hazard function of T_0, we use the approach taken by Uno et al. (2007) to model the conditional risks of experiencing an event by t_0 directly via (2.1) and (2.2).

To estimate β in (2.1), we let the estimating function $S_1(\beta)$ be

$$\sum_{i=1}^{n} \hat{\omega}_i X_i \left\{ I\left(T_i < t_0\right) - g_1\left(\beta' X_i\right) \right\} = \sum_{i=1}^{n} \hat{\omega}_i X_i \left(\left\{ Y_i - g_1\left(\beta' X_i\right) \right\} \right), \quad (3.1)$$

where the weighting $\hat{\omega}_i = \Delta_i I\left(T_i < t_0\right) / \hat{G}\left(T_i\right) + I\left(T_i \geq t_0\right) / \hat{G}\left(t_0\right)$, and $\hat{G}\left(\cdot\right)$ is the Kaplan-Meier estimator of $G(\cdot)$. Here, the weighting takes care of the problem due to censoring and $I(T_i < t_0) = Y_i$ when $\hat{\omega}_i \neq 0$. It is shown in Uno et al. (2007) that as $n \to \infty$, the resulting estimator $\hat{\beta}$ converges to a constant which is free of $G(\cdot)$ even when the model (2.1) is misspecified. Similarly, with additional V and Model (2.2), one can use the estimating function

$$S_2\left(\gamma\right) = \sum_{i=1}^{n} \hat{\omega}_i W_i \left\{ Y_i - g_2\left(\gamma' W_i\right) \right\} \quad (3.2)$$

to estimate γ.

Now, since $\eta_l^{(q)}\left(s\right)$ is between 0 and 1, we use a non-parametric local logistic likelihood estimation procedure to obtain consistent estimator $\eta^{(q)}(s)$. Specifically, first consider a kernel-type nonparametric functional estimator based on the local likelihood "score" function for the standard logistic regression which relates the binary response $I\left\{ \hat{p}\left\{U_i, V_i\right\} \in \left(v_{l-1}, v_l\right), Y_i = q \right\}$ to the regressor $\hat{\varepsilon}_i\left(s\right) = \psi\left(\hat{p}_{U i}\right) - \psi\left(s\right)$. Here, we choose a proper transformation $\psi\left(\hat{p}_U\right)$ of $\hat{p}_U$ to implement the smoothing, where $\psi(\cdot)$ is a known, non-decreasing function (Wand et al., 1991; Park et al., 1997). For any given s, l and q, this results in a score function of the intercept parameter a and slope parameter b:

$$\sum_{i=1}^{n} \left[h^{-1} \frac{1}{\hat{\varepsilon}}\left(s\right) \right] K_h \left\{ \hat{\varepsilon} \quad (s) \right\} \hat{\omega}_i \left[I\left\{ \hat{p}_{\{U_i, V_i\}} \in \left(v_{l-1}, v_l\right), Y_i = q \right\} - g_0 \left\{ a + b\hat{\varepsilon}_i\left(s\right) \right\} \right] \quad (3.3)$$

where $g_0(x) = \exp(x) / \{1 + \exp(x)\}$, $K_h(x) = K(x/h)/h$, $K(\cdot)$ is a known smooth symmetric kernel density function with a bounded support, and the bandwidth $h > 0$ is assumed to be $O\left(n^{-v}\right)$, $for 1/5 < v < 1/2$. Let $\hat{\eta}_l^{(q)} = g_0\left(\hat{a}\right)$ where $\hat{a}$ is the resulting estimator of the intercept by solving the equations: (3.3)=0, for $q = 0, 1$ and $l = 1, \cdots, L$. Then $\eta_l(s)$ and $\xi(s)$ can be estimated by $\hat{\eta}_l\left(s\right) = \hat{\eta}_l^{(1)}\left(s\right) + \hat{\eta}_l^{(0)}\left(s\right) and \hat{\xi}\left(s\right) = \sum_{l=1}^{L} \hat{\eta}_l^{(1)}$ respectively.

In Appendix A, we show that $\hat{\eta}_l^{(s)}(s) - \eta_l^{(q)}(s) \to 0$ in probability, uniformly in

$s \in \Omega_h \equiv [\psi(\rho_l - h)]$, where $[\rho_l, \rho_r]$ is a subset of the support of $\psi\left\{g_1\left(\beta_0' X\right)\right\}$ and β_0 is the limit of $\hat{\beta}$. Furthermore, we show in Appendix B that for large n, the joint distribution of $\left\{(nh)^{1/2}\left\{\hat{\eta}_l^{(q)}(s) - \eta_l^{(q)}(s)\right\}, q=0,1; l=1, \cdots, L\right\}$ can be approximated by the distribution of

$$\left\{n^{-\frac{1}{2}}h^{\frac{1}{2}}\sum_{i=1}^{n}\hat{V}_{li}^{(q)}(s)(V_i - 1) + (nh)^{\frac{1}{2}}\left\{\tilde{\eta}_l^{(q)}\left(s, \hat{\Theta}^*\right) - \tilde{\eta}_l^{(q)}\left(s, \hat{\Theta}\right)\right\}, \quad q=0,1; l=1, \cdots, L\right\} \quad (3.4)$$

given the data, where $\mathbf{V} = \{V_i, i = 1, \cdots, n\}$ is a random sample from a given distribution with unit mean and unit variance and is independent of the data,

$\hat{\Theta} = \left(\hat{\beta}, \hat{\gamma}, \hat{G}\right), \hat{\Theta}^* = \left(\hat{\beta}, \hat{\gamma}^*, \hat{G}^*\right)$ is the perturbed version of $\hat{\Theta}$, $\mathscr{V}_{li}^{(q)}(s), \tilde{\eta}_l^{(q)}$ and $\hat{\Theta}^*$ are given in (A·1), (A·2) and (A·3) of Appendix B, respectively. It is important to note that although asymptotically one only needs the first term of (3.4) since $\hat{\Theta}$ is of order root-n, the inclusion of the second term improves the approximation to the distribution of $\hat{\eta}_l^{(q)}(s)$ in finite sample. The independent realizations of (3.4) can easily be obtained by generating realizations from independent random samples of $\mathbf{V}$. With a large number of realizations from (3.4), confidence interval estimates for $\left\{\eta_l^{(q)}(s)\right\}$ and $\{\eta_l(s)\}$ can be constructed via this large sample approximation. This perturbation-resampling method, which is conceptually similar to a wild bootstrapping (Davison and Hinkley, 1997), has been utilized successfully in solving many complicated problems in survival analysis.

As for any nonparametric functional estimation problem, the choice of the smooth parameter h is crucial for making inferences about $\eta^{(q)}(s)$, $\eta_l(s)$ and $D_l(s)$. Here, we propose to use the standard K-fold cross validation procedure to obtain an "optimal" h. Specifically, we randomly split the data into K disjoint subsets of about equal sizes denoted by $\{J_k, k = 1, \cdots, K\}$. For each k, we use all observations which are not in J_k to estimate $p(u)$ and $p(u, v)$ by fitting the working models (2.1) and (2.2) and then with a given h, to estimate $\eta_l^{(q)}(s)$. Let the resulting estimators be denoted by $\hat{p}_u^{(-k)}$, $\hat{p}_{\{u,v\}}^{(-k)}$ and $\hat{\eta}_l^{(q)(-k)}(s)$. We then use the observations from J_k to calculate the prediction error

$$PE^{(q)}(k,l) = \sum_{j \in J_k} \hat{\omega}_j\left[I\left\{\hat{p}_{\{U_j, V_j\}(-k)} \in [v_{l-1}, v_l], Y_j = q\right\} - \hat{\eta}_l^{(q)-k}\left\{\hat{p}_{U_j}^{(-k)}\right\}\right]^2. \quad (3.7)$$

Lastly we sum (3.7) over $k = 1, \cdots, K$, and then choose h by minimizing this sum of K prediction errors. Note that the optimal smooth parameter value may only work for this specific (l, q) of $\hat{\eta}_l^{(q)}(s)$. Alternatively, one may obtain a uniform bandwidth, which minimizes the sum of (3.7) over k, q and l. The order of such a bandwidth estimator is expected to be $n^{-1/5}$ (Wand and Jones, 1995; Fan and Gijbels, 1996) and thus the final

bandwidth for estimation can be obtained by multiplying the estimated bandwidth by n^{-d_0}, for some $d_0 \in (0, 3/10)$.

To correct for the potential over-fitting bias in estimating the reclassification rates, one may also use the K-fold cross-validation. For $k = 1, \ldots, K$, we use all observations not in J_k to fit the regression models with and without the new markers and obtain $\hat{p}_U^{(-k)}$ and $\hat{p}_{\{U,V\}}^{(-k)}$. Then using observations in J_k, $\hat{p}_U^{(-k)}$ and $\hat{p}_{\{U,V\}}^{(-k)}$, we obtain an estimate for $\eta_l^{(q)}(s)$ based on equation (3.3), denoted by $\hat{\eta}_l^{(q)k}(s)$. Then the final cross-validated estimator of $\eta_l^{(q)}(s)$ is

$$\hat{\eta}_l^{(q)CV}(s) = K^{-1} \sum_{k=1}^{K} \hat{\eta}_l^{(q)}(s).$$

The proposed operational procedure can be briefly summarized as:

1. Construct a scoring system based on the conventional biomarkers predicting the binary outcome $Y = I(T < t_0) : \left\{ \hat{p}_{\{U_i\}}, i = 1, \cdots, n \right\}$.

2. Construct a scoring system based on the conventional as well as novel biomarkers predicting the binary outcome $Y = I(T < t_0) : \left\{ \hat{p}_{\{U_i, V_i\}}, i = 1, \cdots, n \right\}$.

3. Nonparametrically estimate

$$\eta_l^{(q)}(s) = pr \left\{ \hat{p}_{U,V} \in (v_{l-1}, v_l), Y = q | \hat{p}_U = s \right\}, \qquad for \quad q = 0, 1$$

   based on $\left\{ Y_i, \hat{p}_{\{U_i, V_i\}}, \hat{p}_{\{U_i\}}, i = 1, \cdots, n \right\}$.

4. Correct the estimation bias by applying cross-validation for steps 1-3.

5. Make inference on $\eta_l^{(q)}(s)$, $P(s)$, $D_l(s)$ and $D_+(s)$ with the proposed re-sampling method.

4 Numerical Studies

We apply the proposed procedures to analyze the data from female participants in the Cardiovascular Health Study (CHS) with respect to the risk classification rule recommended by AHS for the 10-year risk of CHD events. There are two different ways to define the binary response variable Y. First, if a study subject died before her ten-year follow-up from a non-CHD cause, we let $Y = 0$ (no CHD event). If the subject experienced a CHD event (non-fatal MI, angina or CHD-related death), we let $Y = 1$. There are no loss-to-follow-ups for these endpoints, therefore, we observe all Y's in this analysis. For the second analysis, we let the time to death of other causes be an independent censoring variable C for the time to the CHD event. As we discussed in the introduction, we consider three risk categories (0, 10%], (10%, 20%] and (20%, 100%] in the analysis.

For both analyses, we consider an additive model (2.1) with $g_1(\cdot)$ being the anti-logit function. The vector X consists of the usage of hypertensive medication and all variables used in the model for deriving the FRS given in Wilson et al (1998). Specifically, the FRS model includes various dummy variables indicating levels of blood pressure, total cholesterol, and HDL, as well as age, age^2, present smoking status and diabetic status. Next, we fit the data using an additive model (2.2) with $g_2(\cdot)$ being the anti-logit function and W being the above risk factors/markers X and the log-transformed CRP variable. For estimating functions (3.1) and (3.2), the estimator $\hat{G}(\cdot)$ is the standard K-M estimator with all death and CHD event times as censored observation for censoring variable C. For the nonparametric function estimation, we let $K(\cdot)$ be the Epanechnikov kernel, and $\psi(\cdot) = g_1^{-1}(\cdot)$. The smoothing parameter $h = 0.11$ was obtained through the 10-fold cross validation scheme.

Specifically, we let h be the minimizer of $\sum_{l=1}^{3}\sum_{k=1}^{10}\sum_{q=0}^{1} PE^{(q)}(k,l)$ in (3.7) for both $\hat{\eta}_l^{(0)}$ and $\hat{\eta}_l^{(1)}(s)$ multiplied by a normalizing constant $n^{-1/10}$. To approximate the distributions of these estimators, we used the resampling method (3.4) with 500 independent realized samples of $\{V_i, i = 1, \cdots, n\}$. In addition, to correct for the potential overfitting bias, we also obtained cross-validated estimates of $\eta^{(l)}$ with $h = 0.11$. The results shown below are based on the cross-validated estimates although there appears to be a minimal bias for the present case.

For the case where there is no censoring involved, we present the results of our analysis in Figure 1. Here, we choose the 2nd and 98th percentiles of the empirical distribution of $\left\{\hat{\beta}' X_i, i = 1, \cdots, n\right\}$ as the boundary points ρ_l and ρ_r, respectively. Therefore, the results presented here are for $\hat{p}_U = s \in [0.07, 0.46]$. Part (a)-(c) of the Figure are the estimated $\eta_l^{(y)}(s)$ along with their 95% point-wise confidence intervals, for $l = 1, 2, 3$ and $y = 0, 1$. Part (d) of the Figure gives the estimated re-classification rates over $\hat{p}_U$. As expected, there are substantial re-allocations with the CRP around the boundary points for the risk classification rule. Part (e) shows estimates for the overall gain, $D_+(s)$. The results suggest that although substantial reclassifications occurred, most of the reclassifications are not substantially better than random reallocations with estimated D_+ ranging from 0% to 2.1%. Only subjects with conventional risk estimates $\hat{P}_U \in [0.085, 0.088] \cup [0.29, 0.33]$ have lower bounds of the 95% confidence interval slightly above zero. However, no subgroup of individuals would benefit from the additional CRP information based on simultaneous confidence intervals which controls for the overall type I error. That is, we cannot claim that the reclassification to risk categories by the new marker is better than a random allocation process. Part (f) of the Figure is a smoothed density estimate for $\hat{p}_U$, which provides useful information regarding the relative size of the subgroup D_s of subjects such that $\hat{p}_{U0} = s$.

For the second set of analysis by treating non-CHD death as censoring, 18.8% of subjects are censored with respect to the CHD events of interest. For this case, the smoothing parameter is 0.14 via the 10-fold cross validation by minimizing $\sum_{l=1}^{3}\sum_{k=1}^{10}\sum_{q=0}^{1} PE^{(q)}(k,l)$ in (3.7) with normalizing constant $n^{-1/10}$. We present the

results in Figure 2 for $\hat{p}_U = s \in [0.075, 0.66]$ under the same settings as in Figure 1. The regions for which the CRP may be helpful are

$\hat{p}_U = s \in [0.075, 0.076] \cup [0.29, 0.30] \cup [0.35, 0.38]$. However, the magnitude of gain is also very minimal with estimated D_+ ranging from 0% to 2.5% and no region appears to have a significant gain after controlling for the overall family-wise type I error.

To assess the performance of the proposed procedure under practical settings, we conducted an extensive numerical study to examine the bias of our point estimator and the empirical coverage levels of its corresponding interval estimators for $\left\{\eta_l^{(q)}\right\}(s)$. Based on the results of the study, we find that the new proposed estimation procedure has negligible bias and the coverage levels are very close to the nominal counterparts. For example, in one of our simulation studies, we simulated data from a study similar to the CHS setting.

Specifically, we generated the event time with a lognormal model with the vector of covariates that is identical to that of the CHS study. First, for each subject, we generated its discrete covariates X_D, which include levels of blood pressure, total cholesterol, HDL, present smoking status and diabetic status, based on the empirical distribution of the CHS data. Then, for each given value of X_D, we generated the corresponding value of continuous covariates $X_C = (\text{Age}, \log\text{CRP})$, from $N(\mu^{XD}, \Sigma)$, where μ^{XD} is the empirical mean of X_C given X_D and Σ is the empirical covariance matrix of X_C in the CHS dataset. Then, we generated the survival time T from $\log T = a_0 + a' w + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, $w = (X_D, X_C, \text{Age}^2)$, and $(a_0, a_1, \sigma)'$ was obtained from fitting this log-normal model to the CHS data but the coefficient of CRP was multiplied by a factor of two to represent the setting where CRP is indeed helpful in risk reclassification. The censoring was generated from a Weibull with shape and scale parameters obtained by fitting the two-parameter Weibull model to the CHS data. This results in about 78% of censoring and an overall 10-year event rate of 17%. As in the CHS example, we assess the accuracy of CRP for predicting the 10-year risk and reclassifying subjects into three risk categories with $v_1 = 0.1$ and $v_2 = 0.2$.

We considered a moderate sample size of 3000 and a relatively large sample size of 6000. For computational ease, the bandwidth for constructing the non-parametric estimate was fixed at $h = 0.09$ for $n = 3000$ and 0.07 for $n = 6000$. Here, these h's were chosen as the averages of the bandwidths selected based on (3.7) with $d_0 = 0.1$ from 10 simulated datasets. For both settings, the non-parametric estimator has negligible bias, the estimated standard errors are close to their empirical counterparts, and the empirical coverage levels of the 95% confidence intervals are close to the nominal level. In Figure 3, we summarize the results of our findings performance for the point and interval estimation procedures for

$\hat{\eta}_l^{(1)}(s) \, (l=1, 2, 3)$ when $n = 3000$. The coverage levels are slightly lower or higher than the nominal level when s is near the boundaries of the support and the true reclassification rates $\eta_l^{(1)}(s)$ are close to 0. However, the coverage level improves when we increase the sample sizes. Similar performances were observed for $\hat{\eta}_l^{(0)}(s)$. For example, when $s = 0.2$, the coverage levels for $\left\{\hat{\eta}_1^{(0)}(s), l=1, 2, 3\right\}$ are (0.96, 0.93, 0.96) when $n = 3000$ and (0.96, 0.93, 0.95) = 6000. Other choice of the coefficient of CRP reflecting strong/weak association with

the biomarker of interest was also used in the simulation and the corresponding results are fairly close to that reported above.

5 Remarks

In this article, we present a systematic approach to quantify the added value from new risk markers for classifying subjects into pre-specified risk categories. At each estimated conventional risk level, we provide point and interval estimates for the reclassification rates along with the corresponding proportions of accurate re-assignment for this subgroup of individuals. The proposed procedure is nonparametric and does not depend on any parametric assumptions on the underlying model generating the data. On the other hand, as a limitation, when the reclassification rate (when the new biomarker is uninformative) or pr(Y = 1) is low, a large sample size is often required to make reliable inferences. Furthermore, appropriate weighting scheme reflecting the relative importance of cases and controls can be conveniently accommodated. These quantities play vital roles for cost-benefit analysis even when the cost of measuring the new markers is not an issue or the re-categorization via the new markers is not generated from a random allocation process. In general, if the new markers improve the risk prediction and change the risk estimates drastically, one may expect the new marker to be helpful in risk reclassification for the entire population. However, in most practical settings, we expect that subjects whose conventional risk estimates are not around the risk threshold values would not benefit much from the new markers with respect to a better assignment of risk category. This is indeed confirmed with the results from extensive analysis of the data from the Cardiovascular Health Study. Thus, to optimize the usage of the new markers for risk classification, one may consider ascertaining the new markers only for subjects with conventional risk estimates in a certain range. The proposed procedures can be easily be extended to quantify the average gain for a range of subgroups by integrating $D_+(s)$ over the range of interest.

Although the risk score is generally on a continuous scale, clinical practitioners almost always discretize the score for decision making. Furthermore, when there are multiple underlying disease subtypes or subpopulations, such a risk discretization may serve as a proxy for these subgroups. For example, a composite score consisting of several biomarkers, including cyclin E and Ki-67, stratified non-small cell lung cancer (NSCLC) patients into markedly different survival subgroups. Cyclin E expression is involved in the cell growth, highly predictive of early disease recurrence after surgery, and considered a good candidate as a molecular target for treatment. Risk groups determined by these biomarkers may represent underlying disease subtypes associated with different treatment responses. To illustrate how an underlying biological process may be approximated by such risk discretization, we conducted a simulation study where we generated the log survival time from $N(4, 1)$ if $D = 1$ and $N(0, 1)$ if $D = 0$; the "latent" binary group membership D is generated based on a probit model $P(D = 1)X, Z) = (-2 + 2Z)$ and (X, Z) were generated from a multivariate normal. Suppose that we only observe (T, X, Z), the risk classification rule $P(T$ 10)X, Z) 0.20 is a highly accurate approximation to the unobserved D with concordance rate as high as 90%. Here T is the survival time. Thus, if D represents unknown disease subtypes, appropriate binary risk classification rule is essentially separating disease subtypes and may guide more effective treatment.

The determination of an appropriate set of risk thresholds is an important yet difficult problem. Appropriate cut-off values could potentially be chosen by evaluating the threshold effect of risk factors or performing cost-risk-benefit analysis. Most likely the new markers have non-trivial incremental values for patients whose conventional risk scores are in the neighborhood of the threshold values. The new proposal provides some guidance about how large this neighborhood would be. When there are no specific risk categories in classification, we may compare certain characteristics of the risk scores for "cases" and "controls" to evaluate the added value of new markers. For example, given an initial risk estimate of $\hat{p}_U = s$, one may consider estimating the mean value of $\hat{p}_{U,V} - \hat{p}_U$ among these with $Y = 1$ and among those with $Y = 0$. If the new marker is indeed helpful for the group with $\hat{p}_U = s$, one would expect that $E\left\{\hat{p}_{U,V} - \hat{p}_U | \hat{p}_U = s, Y = 1\right\}$ is positive and $E\left\{\hat{p}_{U,V} - \hat{p}_U | \hat{p}_U = s, Y = 0\right\}$ is negative. For the CHS example, as shown in Figure 4, the gain from CRP over the FRS score seems rather small. Further research is needed to justify this nonparametric estimation procedure.

Acknowledgements

Appendix

Throughout, unless otherwise noted, we use the notation $\simeq$ to denote equivalence up to $o_p(1)$ uniformly in s, $\lesssim$ to denote bounded above up to a universal constant, and $\mathscr{F}(x)$ to denote $dF(x)/dx$ for any function F. We use $\mathbb{P}_n$ and $\mathbb{P}$ to denote expectation with respect to the empirical probability measure of $\{(T_i, \quad_i, X_i, W_i), i = 1, \cdots, n\}$ and the probability measure of $(T, \quad, X, W)$, respectively. Similarly $\mathbb{G}_n = n^{\frac{1}{2}}(\mathbb{P}_n - \mathbb{P})$. Let β_0 and γ_0 denote the solution to the equations $E\left[X_i\left\{I(T_{0i} < t_0) - g_1\left(\beta' X_i\right)\right\}\right] = 0$ and $E\left[W_i\left\{I(T_{0i} < t_0) - g_2\left(\gamma' W_i\right)\right\}\right] = 0$, respectively. Let $\bar{p}_U = g_1\left(\beta_0' X\right)$. We assume that $\tau(\cdot)$, the density function of $\psi\left(\bar{p}_U\right)$, is continuously differentiable with bounded derivatives and bounded away from zero on the interval $[\psi^{-1}(\rho_l), \psi^{-1}(\rho_u)] \subset (0, 1)$. We also assume that the marker values are bounded, $\left(\beta_0', \gamma_0'\right)'$ belongs to a compact set Ω. For the bandwidth h, we assume that $h = O(n^{-v})$, $1/5 < v < 1/2$.

We first note that from Uno et al. (2007), we have

$$|\hat{\beta} - \beta_0| + |\hat{\gamma} - \gamma_0| + \sup_{t \leq t_0}|\hat{G}(t) - G(t)| = O_p\left(n^{-\frac{1}{2}}\right). \quad \text{(A.0)}$$

It follows that $|\eta_l^{(q)}(s) - \bar\eta_l^{(q)}(s)| = O_p(1)$, where

$$\eta_l^{(q)}(s) = E\left(\hat{M}_{li}^{(q)} | \hat{p}_U = s\right), \hat{M}_{li}^{(q)}$$

$$= I\left\{\hat{p}_{\{U_i, V_i\}} \in (v_0, v_1), Y_i = q\right\}, \bar\eta_l^{(q)}(s)$$

$$= E\left(M_{li}^{(q)} | \bar{p}_U = s\right), M_{li}^{(q)}$$

$$= I\left\{\bar{p}_{\{U_i, V_i\}} \in (v_0, v_1), Y_i = q\right\} \qquad , \text{ and } \bar{p}\{U, V\} = g_2\left(\gamma_0' W\right). \text{ It is also important}$$

to note that $E\{\omega s(T_0, X, W)\} = E\{s(T_0, X, W)\}$ for any function $s(\cdot, \cdot, \cdot)$, where $\hat\eta_1^{(1)}(\cdot)$.
We next derive the asymptotic theory for $\hat\eta_1^{(1)}(\cdot)$, but note that the same arguments can be used to justify the asymptotic properties of other quantities. For the ease of notation, in Appendix A and B, we suppress the subscript l and supscript (q) from $\hat{M}_{li}^{(q)}$ and $M_{li}^{(q)}$.

A. Uniform Consistency of $\hat\eta_1^{(1)}(\cdot)$

At any given s, let $\left\{\hat{a}(s), \hat{b}(s)\right\}'$ be the root of the estimating equation

$$\sum_{i=1}^n \left[h^{-1} \dot{\hat\epsilon}_i(s)\right] K_h\{\hat\epsilon_i(s)\} \hat\omega_i\left[\hat{M}_i - g_0\{a + b\hat\epsilon_i(s)\}\right] = 0,$$

$$a(s) = g_0^{-1}\left\{\eta_1^{(1)}(s)\right\}, b(s) = d\left[g_0^{-1}\left\{\eta_1^{(1)}(s)\right\}\right]/ds = \dot\eta_1^{(1)}(s)/\dot{g}_0\{a(s)\} \text{ and}$$

$$\hat{d}(s) = \left[\begin{array}{c} \hat{d}_a(s) \\ \hat{d}_b(s) \end{array}\right] = \left[\begin{array}{c} \hat{a}(s) - a(s) \\ h^{-1}\left\{\hat{b}(s) - b(s)\right\} \end{array}\right]$$

Our objective is to show that $\hat{d}(s) \to 0$ in probability as $n \to \infty$. To this end, we note that for any given s, $\hat{d}(s)$ is the solution to the estimating equation $\hat{\mathbb{S}}(d, s) = 0$, where $\mathbf{d} = (d_a, d_b)'$,

$$\hat{\mathbb{S}}(d, s) = \left[\begin{array}{c} \hat{\mathbb{S}}_1(d; s) \\ \hat{\mathbb{S}}_2(d; s) \end{array}\right] = n^{-1}\sum_{i=1}^n \left[h^{-1} \dot{\hat\epsilon}_i(s)\right] K_h\{\hat\epsilon_i(s)\} \omega_i\left[\hat{M}_i - \mathscr{G}\{d, s; \psi(\hat{p}U_i), h\}\right]$$

and $\mathscr{G}(d, s; y, h) = g_0\left[a(s) + b(s)\{y - \psi(s)\} + d_a + d_b h^{-1}\{y - \psi(s)\}\right]$.

The first step is to show that $\hat{\mathbb{S}}(d; s)$ is uniformly consistent for

$$\$\,(\mathrm{d};s) = \left[\begin{array}{c} \mathbb{S}_1\,(\mathrm{d};s) \\ \mathbb{S}_2\,(\mathrm{d};s) \end{array} \right] = \tau\,\{\psi\,(s)\}\left[\eta_1^{(1)(s)} - \int K\,(t)\,g_0\,\{a\,(s) + d_a + d_b t\}\,dt - \int t K\,(t)\,g_0\,\{a\,(s) + d_a + d_t\}\,dt \right]$$

We first show that

$$\overset{s}{\sup}\left| n^{-1}\sum_{i=1}^{n} K_h\,\{\hat{\varepsilon}_i\,(s)\}\,\hat{\omega}_i \hat{M}_i - \tau\,\{\psi\,(s)\}\,\eta_1^{(1)}\,(s) \right|$$

and

$$\underset{\mathrm{d},s}{\sup}\left| n^{-1}\sum_{i=1}^{n} K_h\,\{\hat{\varepsilon}_i\,(s)\}\,\omega_i \mathscr{G}\left\{\mathrm{d},s;\psi\left(\hat{p}_{U_i}\right),h\right\} - \tau\,\{\psi\,(s)\}\int K\,(t)\,g_0\,\{a\,(s) + d_a + d_b t\}\,dt \right|$$

are both $O_p\left\{(nh)^{-\frac{1}{2}}log\,(n)\right\}$. To this end, we note that from (A.0),

$$\left| n^{-1}\sum_{i=1}^{n}(\hat{\omega}_i - \omega_i)\,K_h\,\{\hat{\varepsilon}_i\,(s)\}\,\mathscr{G}\left\{\mathrm{d},s;\psi\left(\hat{p}_{U_i},h\right)\right\} \right| \leq n^{-1}\sum_{i=1}^{n}|\hat{\omega}_i - \omega_i|\,K_h\,\{\hat{\varepsilon}_i\,(s)\} = O_p\left(n^{-\frac{1}{2}}\right)$$

This, together with the convergence of $n^{\frac{1}{2}}\left(\hat{\beta} - \beta_0\right)$, implies that

$$\left| n^{-1}\sum_{i=1}^{n} K_h\,\{\hat{\varepsilon}_i\,(s)\}\,\hat{\omega}_i \mathscr{G}\left\{\mathrm{d},s;\psi\left(\hat{p}_{U_i},h\right)\right\} - \tau\,\{\psi\,(s)\}\int K\,(t)\,g_0\,\{a\,(s) + d_a + d_b t\}\,dt \right|$$

$$\lesssim n^{-\frac{1}{2}}h^{-1}\|\mathbb{G}_n\|_{\mathscr{H}_\delta} + \left[n^{-\frac{1}{2}}\int K_h\,\{y - \psi\,(s)\}\,d\mathbb{G}_n\left[\omega\mathscr{G}\left\{\mathrm{d},s;\psi\left(\bar{p}_U\right),h\right\} I\left\{\psi\left(\bar{p}_U\right) \leq y\right\} \right] \right] + O_p\left(n^{-\frac{1}{2}}\right) + h^2$$

where $\mathscr{H}_\delta = \left\{\omega I\left[\psi\left\{g_1\left(\beta' x\right)\right\} \leq c\right]\omega I\left[\psi\left\{g_1\left(\beta_0' x\right)\right\} \leq c\right]:\|\beta - \beta_0\| \leq \delta, y\right\}$ is a class of functions indexed by β and c. Furthermore, δ is uniformly bounded by an envelop function in the order of $\delta^{\frac{1}{2}}$ with respect to L_2 norm. By the maximum inequality of van der Vaart and Wellner (1996), we have

$E\|\mathbb{G}_n\|_{\mathscr{H}_\delta} \lesssim \delta^{\frac{1}{2}}\left(|log\,(\delta)| + |log\,(h)|\right)\left(1 + \delta^{-\frac{1}{2}}n^{-\frac{1}{2}}\left(|log\,(\delta)| + |log\,(h)|\right)\right)$. This, coupled with $|\hat{\beta} - \beta_0| = O_p\left(n^{-\frac{1}{2}}\right)$ implies that $n^{-\frac{1}{2}}h^{-1}\|\mathbb{G}_n\|_{\mathscr{H}_\delta} \lesssim O_p\left\{(nh)^{-\frac{1}{2}}\left(nh^2\right)^{-\frac{1}{4}}log\,(n)\right\}$.

Secondly, with the standard arguments used in Bickel and Rosenblatt (1973), we have

$$\left[n^{-\frac{1}{2}}\int K_h\,\{x - \psi\,(s)\}\,d\mathbb{G}_n\left[\omega\mathscr{G}\left\{\mathrm{d},s;\psi\left(\bar{p}_U\right),h\right\} I\left\{\psi\left(\bar{p}_U\right) \leq y\right\} \right] \right] + O_p\left\{(nh)^{-\frac{1}{2}}log\,(n)\right\}$$

Therefore, for $h = n^{-v}, 1/5 < v < 1/2$,

$$\sup_{\mathrm{d},s} \left| n^{-1} \sum_{i=1}^{n} K_h \{\hat{\varepsilon}_i(s)\} \hat{\omega}_i \mathscr{G} \left\{ \mathrm{d}, s; \psi \left(\hat{p}_{U_i}, h \right) \right\} \tau \{\psi(s)\} \int K(t) g_0 \{a(s) + d_a + d_b t\} dt \right|$$

is $O_p \left\{ (nh)^{-\frac{1}{2}} log(n) \right\}$. From (A.0), $|\hat{\gamma} - \gamma_0| = O_p \left(n^{-\frac{1}{2}} \right)$, and similar arguments as given above,

$$\sup_{s} \left| n^{-1} \sum_{i=1}^{n} K_h \{\hat{\epsilon}_i(s)\} \hat{\omega}_i \hat{M}_i - \tau \{\psi(s)\} \eta_1^{(1)}(s) \right| = O_p \left\{ (nh)^{-\frac{1}{2}} log(n) \right\},$$

and hence $\sup_{\mathrm{d},s} |\hat{\mathbb{S}}_1(\mathrm{d}, s) - \mathbb{S}_1(\mathrm{d}, s)| = O_p \left\{ (nh)^{-\frac{1}{2}} log(n) \right\} = o_p(1)$. From the same arguments as above, $\sup_{\mathrm{d},s} |\hat{\mathbb{S}}_2(\mathrm{d}, s) - \mathbb{S}_2(\mathrm{d}, s)| = O_p \left\{ (nh)^{-\frac{1}{2}} log(n) \right\} = o_p(1)$. Therefore $\sup_{\mathrm{d},s} |\hat{\mathbb{S}}(\mathrm{d}, s) - \mathbb{S}(\mathrm{d}, s)| = o_p(1)$. This uniform convergence, coupled with the fact that $\mathbf{0}$ is the unique solution to the equation $\mathrm{S}(\mathbf{d}, s) = 0$ with respect to $\mathbf{d}$ and all the eigenvalues of $\mathbb{A}(s) = -\partial \mathbb{A}(\mathrm{d};s)/\partial \mathrm{d}'|_{d=0} = \tau \{\psi(s)\} diag \{1, \int v^2 K(v) dv\}$ are uniformly bounded above zero, suggests that $\sup_{\mathrm{d},s} |\hat{d}(s)| = O_p \left\{ (nh)^{-\frac{1}{2}} log(n) + h \right\} = o_p(1)$, which implies the consistency of $\hat{\eta}_1^{(1)}(s) = g_0 \{\hat{a}(s)\}$.

B. Asymptotic distribution of $\hat{\eta}_1^{(1)}(\cdot)$

It follows from a Taylor series expansion that

$$(nh)^{\frac{1}{2}} \hat{d}_a(s) = (nh)^{\frac{1}{2}} \{\hat{a}(s) - a(s)\} = \hat{\mathbb{A}}_1(s)' (nh)^{\frac{1}{2}} \hat{\mathbb{S}}(0;s) + O_p \left\{ (nh)^{\frac{1}{2}} \left(|\hat{d}_a(s)|^2 + |\hat{d}_b(s)|^2 \right) \right\}.$$

where $\hat{\mathbb{A}}_1(s)$ is the first row of $\hat{\mathbb{A}}(s) = - \left\{ \partial \hat{\mathbb{S}}(\mathrm{d};s)/\partial \mathrm{d}' \right\}^{-1}|_{d=0}$. Using arguments similar to those in the previous section, one can show that $\hat{\mathbb{A}}_1(s)$ converges to $\mathbb{A}_1(s)$, the first row of $\mathbb{A}(s)$, uniformly in s. Furthermore, with the convergence rate of $\hat{\mathbb{S}}_1(\mathrm{d}s)$, it is not difficult to show that the remainder term is bounded by $O_p \left\{ (nh)^{-\frac{1}{2}} log(n)^2 + (nh)^{\frac{1}{2}} h^2 \right\}$ uniformly in s. It follows that

$$(nh)^{\frac{1}{2}} \hat{d}_a(s) = \frac{(nh)^{\frac{1}{2}} \hat{\mathbb{P}}_1 0;(s)}{\tau \{\psi(s)\} \dot{g}_0} \{a(s)\} + O_p \left\{ (nh)^{\frac{1}{2}} log(n)^2 + (nh)^{\frac{1}{2}} h^2 \right\}.$$

This, together with the convergence rate of $\hat{\$}(0;s)$, implies $sup|\hat{d}_a(s)|=O_p\left\{(nh)^{\frac{1}{2}}log(n)\right\}$.

It follows that $(nh)^{\frac{1}{2}}\hat{d}_a(s)$ is asymptotically equivalent to

$$\frac{(nh)^{\frac{1}{2}}\mathbb{P}_n\left(K_h\{\hat{\epsilon}(s)\}\omega\left[\hat{M}-\mathscr{G}_0\{s,\psi(\hat{p}U)\}\right]\right)}{\tau\{\psi(s)\}\dot{g}_0\{a(s)\}}$$

where $G_0(s,y)=g_0[a(s)+b(s)\{y-\psi(s)\}]$. We next show that $(nh)^{\frac{1}{2}}\hat{d}_a(s)$ is asymptotically equivalent to

$$(nh)^{\frac{1}{2}}\tilde{d}_2(s)=\frac{(nh)^{\frac{1}{2}}\mathbb{P}_n\left(K_h\{\varepsilon(s)\}w\right)\left[M-\mathscr{G}_0\left\{s,\psi\left(\hat{p}_U\right)\right\}\right]}{\tau\{\psi(s)\}\dot{g}_0\{a(s)\}}$$

i.e., $\hat{p}_U$ and $\hat{P}_{\{U,V\}}$ can be replaced by their respective $\bar{p}_U$ and $\bar{P}_{\{U,V\}}$ in $\hat{d}_a(s)$, where

$\varepsilon(s)=\psi\left(\bar{p}_U\right)-\psi(s)$. To this end, noticing the fact that $\tau\{\psi(s)\}\dot{g}_0\{a(s)\}$ is bounded away from zero uniformly in s, we have

$$(nh)^{\frac{1}{2}}|\hat{d}_a(s)-\tilde{d}_a(s)|\gtrsim h^{-\frac{1}{2}}\|\mathbb{G}_n\|_{\mathscr{F}_\delta}+h^{-\frac{1}{2}}\|\mathbb{G}_n\|_{\mathscr{H}_\delta}+O_p\left\{(nh)^{\frac{1}{2}}|\hat{\beta}-\beta_0|+|\hat{\gamma}-\gamma_0|+h^2\right\}$$

where

$\mathscr{F}_\delta=\left\{\omega yI\left\{g_2\left(\gamma'w\right)\in(v_0,v_1)\right\}I\left[\psi\left\{g_1\left(\beta'x\right)\right\}\leq c\right]-\omega yI\left\{g_2\left(\gamma_0'w\right)\in(v_0,v_1)\right\}I\left(\psi\left\{g_1\left(\beta_0'x\right)\right\}\leq c\right):|\gamma$
is the class of functions indexed by (γ,β,c). By the maximum inequality and

$|\hat{\beta}-\beta_0|+|\hat{\gamma}-\gamma_0|=O_p\left(n^{-\frac{1}{2}}\right)$, we have $h^{-\frac{1}{2}}\|\mathbb{G}_n\|_{\mathscr{F}_\delta}=O_p\left\{h^{-\frac{1}{2}}n^{-\frac{1}{4}}log(n)\right\}$. It follows that

$(nh)^{\frac{1}{2}}\hat{d}_a(s)-(nh)^{\frac{1}{2}}\tilde{d}_a(sq)$ is asymptotic $o_p(1)$ uniformly in s. This, together with the

standard arguments for local linear regression fitting, implies that $(nh)^{\frac{1}{2}}\{\hat{a}(s)\}-a(s)$

converges to a normal with mean 0 and variance $\sigma^2=\dot{g}_0\{a(s)\}^{-2}$, where $m_2=\int K(s)^2ds$ and

$\sigma^2(s)=m_2\tau\{\psi(s)\}^{-2}E\left[\left\{M-\eta_1^{(1)}(s)\right\}^2/G\left(T_0\Omega t_0\right)|\bar{p}_U=s\right]$. Then, from a delta method,

$(nh)^{\frac{1}{2}}\left\{\hat{\eta}_1^{(s)}(s)-\eta_1^{(1)}(s)\right\}$ is asymptotically normal with mean 0 and variance $\sigma^2(s)$.

To justify the resampling method, we note that from (A.0),

$$n^{-\frac{1}{2}}h^{\frac{1}{2}}\sum_{i=1}^n\hat{\mathscr{V}}_{li}^{(q)}(s)(V_i-1)+(nh)^{\frac{1}{2}}\left\{\tilde{\eta}_l^{(q)}\left(s,\hat{\Theta}^*\right)-\tilde{\eta}_1^{(q)}\left(s,\hat{\Theta}\right)\right\}=n^{-\frac{1}{2}}h^{\frac{1}{2}}\sum_{i=1}^n\hat{\mathscr{V}}_{li}^{(q)}(s)(V_i-1)+o_{p*}(1)$$

where $o_{p_*}(1)(1)$ is with respect to the product probability measure generated by the observed data and $\mathbf{V}$,

$$\hat{\mathscr{V}}_{li}^{(q)}(s) = \hat{\tau}\{\psi(\hat{p}U_i)\}^{-1} K_h\{\hat{\epsilon}_i(s)\}\hat{\omega}_i\left\{\hat{M}_{li}^{(q)} - \hat{\eta}_l^{(q)}(s)\right\}, \quad \text{(A.1)}$$

$$\tilde{\eta}_l^{(q)}(s,\Theta) = \frac{\sum_{i=1}^n \frac{O_i}{G(T_i\Lambda t_0)} K_{\tilde{h}}\left\{\psi\left(g_1\left(\beta'X_i\right)\right) - \psi(s)\right\} I\left\{g_2\left(\gamma'W_i\right) \in (v_{l-1}, v_l), Y_i = q\right\}}{\sum_{i=1}^n \frac{O_i}{G(T_i\Lambda t_0)K_{\tilde{h}}\left\{\psi(g_1(\beta'X_i)) - \psi(s)\right\}}} \quad \text{(A.2)}$$

$\hat{\tau}(s) = \sum_{i=1}^n K_h\{\hat{\varepsilon}_i(s)\}/n$ is the estimated density function of

$\psi\left(\hat{p}_U\right)$, $O_i = \Delta_i I\left(T_i < t_0\right) + I\left(T_i \geq t_0\right)$ $\hat{\beta}^*$ and $\hat{\gamma}^*$ are the respective solutions to

$$\sum_{i=1}^n \frac{O_i}{\hat{G}^*(T_i\Lambda t_0)} X_i\left\{Y_i - g_1\left(\beta'X_i\right)\right\} V_i = 0, \qquad \sum_{i=1}^n \frac{O_i}{\hat{G}^*(T_i\Lambda t_0)} W_i\left\{Y_i - g_2\left(\gamma'W_i\right)\right\} V_i = 0 \quad \text{(A.3)}$$

$\hat{G}^*(\cdot)$ is the weighted Kaplan-Meier estimates with weights $\mathbf{V}$ and $\tilde{h} = h \times n^{d_0+1/5-1/7}$. Thus, conditional on the data the random variable given in (3.4) is asymptotically equivalent

to $\hat{\tau}\{\psi(s)\}^{-1}(nh)^{\frac{1}{2}}\mathbb{P}_n\left[K_h\{\hat{\varepsilon}(s)\}\hat{\omega}\left\{\hat{M} - \hat{\eta}_1^{(1)}(s)\right\}(v-1)\right]$, which is asymptotical normally with mean zero and variance

$\hat{\sigma}^2(s) = h\hat{\tau}\{\psi(s)\}^{-2}\mathbb{P}_n\left[K_h\left\{\hat{\varepsilon}(s)^2\hat{\omega}\right\}^2\hat{\omega}^2\left\{\hat{M} - \hat{\eta}_1^{(1)}(s)\right\}^2\right]$. From the same arguments

given in Appendix A, $\hat{\sigma}^2(s) \to \sigma^2(s)$ in probability, as $n \to \infty$. Note that the second term, which accounts for the variability in $\hat{\Theta}$, involves estimating the partial derivative of

$P\left\{g_2\left(\gamma'W_i\right) \in [v_{l-1}, v_l], Y_i = q | g_1\left(\beta'X\right) = s\right\}$ with respect to β and γ. Since the optimal bandwidth for such derivative functions is of order $n^{-1/7}$ which could be substantially larger

than h used for the estimation of $\hat{\eta}_l^{(q)}$, we find that the second term in (3.4) approximates the

additional variability in $\hat{\Theta}$ better when a larger bandwidth is used to obtain $\tilde{\eta}_l^{(q)}$. As a result, the standard bootstrap method, which essentially uses the same h for b $^{-l}$ performs poorly in finite sample.

References

1. Bickel PJ, Rosenblatt M. On some global measures of the deviations of density function estimates. The Annals of Statistics. 1973; 1:1071–1095.

2. Buring J, Hennekens C. The Women's Health Study: summary of the study design. Journal of Myocardial Ischemia. 1992; 4:27–9.

3. Chi Y, Zhou X. The need for reorientation toward cost-effective prediction: Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by Pencina et al (2008). Statistics in Medicine. 2008; 27(2):182–184. [PubMed: 17712782]

4. Cook N. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007; 115(7):928. [PubMed: 17309939]

5. Cook N. Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by Pencina et al (2008). Statistics in Medicine. 2008; 27(2):191–195. [PubMed: 17671959]

6. Cook NR, Buring JE, Ridker P. The effect of including c-reactive protein in cardiovascular risk prediction models for women. Annals of Internal Medicine. 2006; 145:21–29. [PubMed: 16818925]

7. D'Agostino R. Risk prediction and finding new independent prognostic factors. Journal of Hypertension. 2006; 24(4):643. [PubMed: 16531791]

8. Davison, A.; Hinkley, D. Bootstrap Methods and Their Application. Cambridge Univ Pr.; 1997.

9. Fan, J.; Gijbels, I. Local polynomial modelling and its applications, Vol. 66 of Monographs on Statistics and Applied Probability. Chapman Hall; London: 1996.

10. Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A. The Cardiovascular Health Study: design and rationale. Annals of Epidemiology. 1991; 1(3):263–76. [PubMed: 1669507]

11. Greenland P. Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by Pencina et al (2008). Statistics in Medicine. 2008; 27(2):188–190. [PubMed: 17579927]

12. Grundy S, Cleeman J, et al. Implications of recent clinical trials for the national cholesterol education program adult treatment panel III guidelines. Journal of the American College of Cardiology. 2004; 44(3):720–732. [PubMed: 15358046]

13. Mealiffe M, Stokowski R, Rhees B, Prentice R, Pettinger M, Hinds D. Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. J Natl Cancer Inst. 2010; 102:1618–1627. [PubMed: 20956782]

14. Mosca L, Appel L, Benjamin E, Berra K, Chandra-Strobos N, Fabunmi R, Grady D, Haan C, Hayes S, Judelson D. e. a. Evidence-based guidelines for cardiovascular disease prevention in women (American Heart Association). Circulation. 2004; 109(5):672–93. [PubMed: 14761900]

15. Mosca L, Banka CL, Benjamin E. J. e. a. Evidence-based guidelines for cardiovascular disease prevention in women: 2007 update (for the expert panel/writing group). Circulation. 2007; 115:1481–501. [PubMed: 17309915]

16. Park B, Kim W, Ruppert D, Jones M, Signorini D, Kohn R. Simple transformation techniques for improved non-parametric regression. Scandinavian journal of statistics. 1997; 24(2):145–163.

17. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in Medicine. 2008; 27(2):157–72. [PubMed: 17569110]

18. Pepe M, Feng Z, Huang Y, Longton G, Prentice R, Thompson I, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. American Journal of Epidemiology. 2008; 167(3):362. [PubMed: 17982157]

19. Pepe M, 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(9):882–890. [PubMed: 15105181]

20. 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. Journal of the American Medical Association. 2007; 297:611–9. [PubMed: 17299196]

21. Tian L, Cai T, Goetghebeur E, Wei LJ. Model evaluation based on the sampling distribution of estimated absolute prediction error. Biometrika. 2007; 94(2):297.

22. Uno H, Cai T, Tian L, Wei LJ. Evaluating prediction rules for t-Year survivors with censored regression models. Journal of the American Statistical Association. 2007; 102(478):527–537.

23. van der Vaart, A.; Wellner, J. Weak Convergence and Empirical Processes. Springer; 1996.

24. Waholder S, Hartge P, Prentice R, et al. Performance of common genetic variants in breast-cancer risk models. New England Journal of Medicine. 2010; 362:986–993. [PubMed: 20237344]

25. Wand, M.; Jones, M. Kernel Smoothing. Chapman & Hall/CRC; 1995.

26. Wand MP, Marron JS, Ruppert D. Transformation in density estimation (with comments). Journal of the American Statistical Association. 1991; 86:343–361.

27. Wang T, Gona P, et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. New England Journal of Medicine. 2006; 355(25):2631. [PubMed: 17182988]

28. Wilson P, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97:1837–47. [PubMed: 9603539]
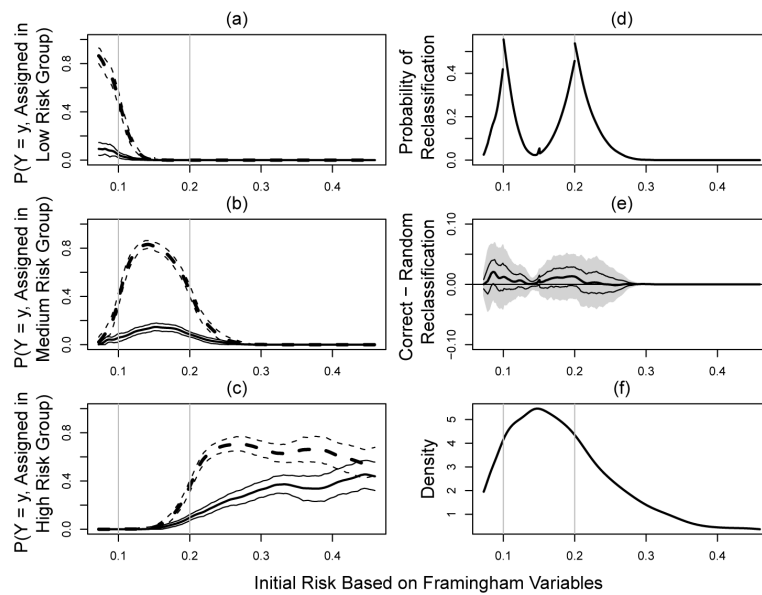
Fig. 1.
Evaluating CRP incremental values for female participants from Cardiovascular Heath

Study by treating non-CHD death as non-event; (a)-(c): $\eta_l^{(y)}(\cdot)$, $l = 1, 2, 3$, (solid lines for $y = 1$ and dashed lines for $y = 0$; thick lines for point estimators and thinner lines for 95% confidence intervals); (d) proportion of reclassification to other risk categories; (e) net gain in reclassification as measured by $D_+(s)$ (point estimator: thick solid line, 95% point wise confidence intervals: dashed lines, and simultaneous confidence intervals: shaded region); (f) density function of the initial risks.
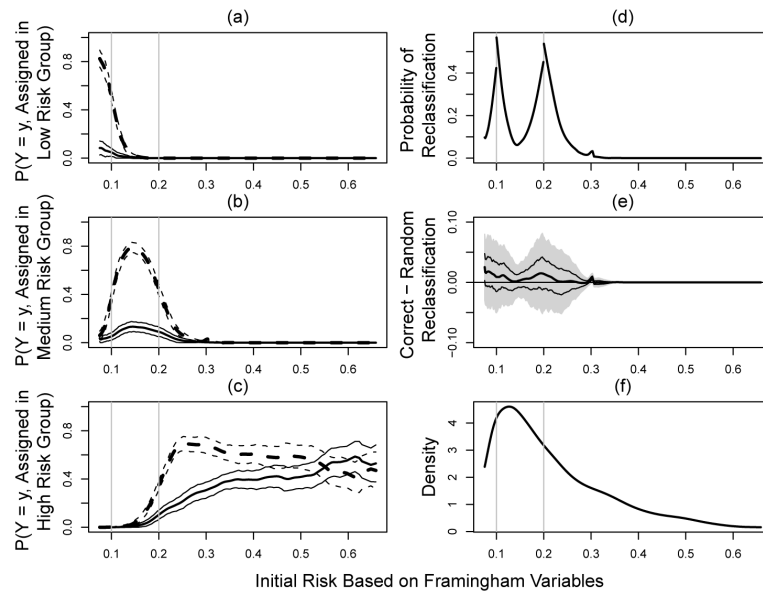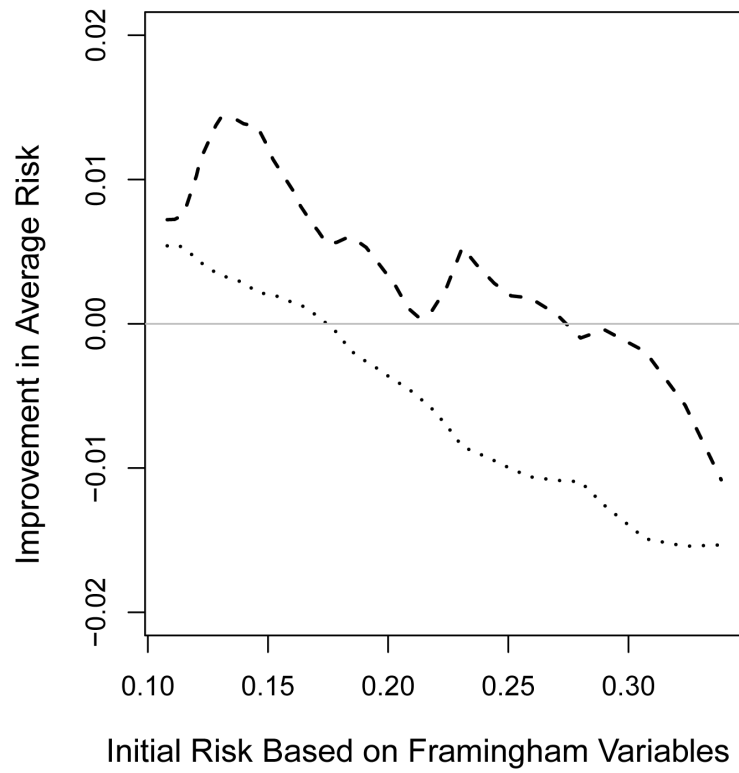
Fig. 2.
Evaluating CRP incremental values for female participants from Cardiovascular Heath Study by treating non-CHD death as censoring; (a)-(c): $\eta_l^{(y)} (\cdot)$, $l = 1, 2, 3$, (solid lines for $y = 1$ and dashed lines for $y = 0$; thick lines for point estimators and thinner lines for 95% confidence intervals); (d) proportion of reclassification to other risk categories; (e) net gain in reclassification as measured by $D_+(s)$ (point estimator: thick solid line, 95% point wise confidence intervals: dashed lines, and simultaneous confidence intervals: shaded region); (f) density function of the initial risks.
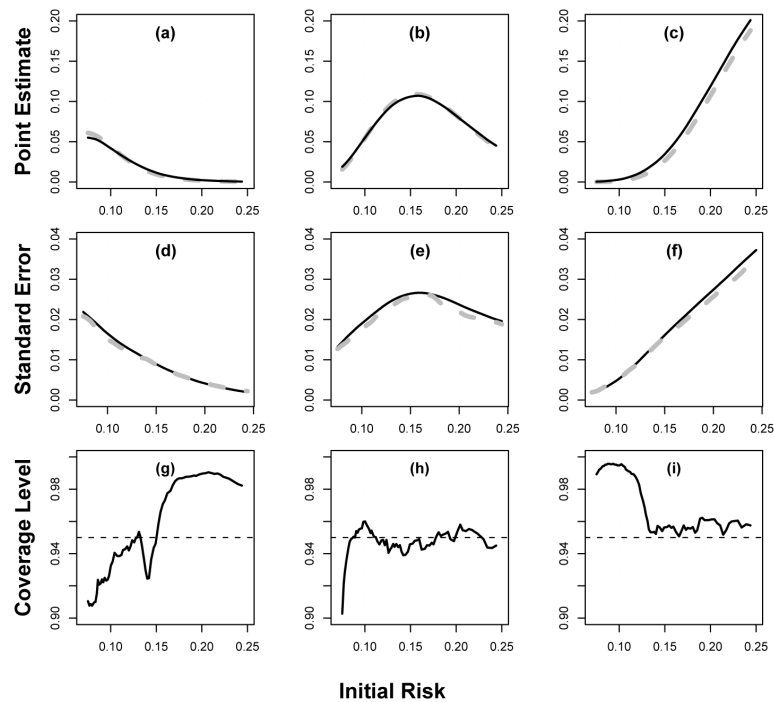
Fig. 3.
Performance of the proposed procedure under a mis-specified model with sample size 3000 where the underlying effect of CRP is twice as big as that estimated from CHS data: (a,b,c) the sample average of $\hat{\eta}_l^{(1)}(s)$ (l=1,2,3) (solid curve) compared to the truth (gray dashed curve); (d,e,f) the average of the estimated standard errors (solid curve) for $\hat{\eta}_l^{(1)}(s)$ compared to the empirical standard error estimates (gray dashed curves); and (g,h,i) the empirical coverage of the 95% confidence intervals based on the proposed resampling procedures. The three columns represents $l = 1$, 2 and 3, respectively.

Fig. 4.
Evaluating CRP incremental values with respect to improvement in mean risk among cases with $Y = 1$ (dashed curve) and among controls with $Y = 0$ (dotted curve) for female participants from Cardiovascular Heath Study by treating non-CHD death as non-event.