Biostatistics (2011), **12**, 4, *pp.* 597–609
doi:10.1093/biostatistics/kxr001
Advance Access publication on March 16, 2011

Comparing costs associated with risk stratification rules for *t*-year survival

TIANXI CAI*

Department of Biostatistics, Harvard University, Boston, MA 02115, USA
tcai@hsph.harvard.edu

LU TIAN

Department of Health Research & Policy, Stanford University, Stanford, CA 94305, USA

DONALD M. LLOYD-JONES

Department of Preventive Medicine, Northwestern University, Chicago, IL 60611, USA

SUMMARY

Accurate risk prediction is an important step in developing optimal strategies for disease prevention and treatment. Based on the predicted risks, patients can be stratified to different risk categories where each category corresponds to a particular clinical intervention. Incorrect or suboptimal interventions are likely to result in unnecessary financial and medical consequences. It is thus essential to account for the costs associated with the clinical interventions when developing and evaluating risk stratification (RS) rules for clinical use. In this article, we propose to quantify the value of an RS rule based on the total expected cost attributed to incorrect assignment of risk groups due to the rule. We have established the relationship between cost parameters and optimal threshold values used in the stratification rule that minimizes the total expected cost over the entire population of interest. Statistical inference procedures are developed for evaluating and comparing given RS rules and examined through simulation studies. The proposed procedures are illustrated with an example from the Cardiovascular Health Study.

Keywords: Disease prognosis; Optimal risk stratification; Risk prediction.

1. INTRODUCTION

Accurate risk assessment and disease prognosis are essential in health care. To improve disease prevention and management, risk stratification (RS) rules are often developed to assign subjects into different risk groups where each group corresponds a particular intervention. For example, a commonly used RS rule in cardiovascular disease prevention stratifies patients into low, intermediate, and high risk groups. Patients are typically recommended to receive antihypertensive therapy if in the intermediate risk group and receive statin if in the high risk group. In studies designed to develop RS rules, measurements of risk factors are often ascertained at baseline and patients are followed over time for the occurrence of a certain clinical

*To whom correspondence should be addressed.

event. Since the risk of experiencing such an event may change over time, one must incorporate the time domain when constructing RS rules. For example, cardiovascular RS rules are often based on the risk of experiencing a cardiovascular event within 10 years since the measurement of the risk factors. In this paper, we are interested in stratification rules for the risk of experiencing an event within t years since marker measurement. Throughout, we use the terms "cases" and "controls" to denote subjects who will and will not experience an event within t years, respectively, if the RS of interest has not been employed. The potential disease status may be changed after patients receive RS-guided intervention.

When developing and evaluating RS rules, it is crucial to understand the potential clinical and financial costs associated with assigning patients into incorrect risk groups and thus receiving suboptimal interventions. Unnecessary medication costs arise when controls are incorrectly assigned to high risk groups. Assigning cases to the low-risk category may lead to costs of life-years lost, productivity, and the subsequent medication. This signifies the importance of precise risk prediction and rigorous evaluation of RS rules prior to their wide spread use in clinical practice.

In practice, RS rules are often derived from risk prediction models with a panel of markers. Based on the predicted risk from the model, future subjects are assigned to different risk categories to receive the corresponding intervention. There are 3 important steps in developing an effective RS rule: (1) constructing a regression model predictive to the clinical response of interest (2) determining the appropriate risk category corresponding to specific intervention, and (3) evaluating the resulting RS rule in an objective and transparent way. While most of statistical methodological research focuses on the step of empirical model building, the clear answer to latter 2 steps remains elusive. When evaluating the performance of risk prediction models, measures of accuracy based on the discrimination and calibration have been considered (Gail and Pfeiffer, 2005; Cook, 2007). Discrimination measures the ability of the risk prediction model in discriminating cases from controls. Calibration measures how well the predicted risk approximates the true conditional risk given the marker measurements. However, neither of these 2 types of measures are appropriate for evaluating the performance of RS. One of the most commonly used discrimination measure is the receiver operating characteristic (ROC) curve (Pepe, 2003). Since the ROC curve is scale invariant, a monotone transformation of predicted risks does not affect the discriminatory accuracy but could lead to dramatic changes in the assignment of risk groups. Calibration measures such as the Hosmer–Lemeshow goodness of fit statistic are also inadequate because a perfectly calibrated model may have poor performance in RS if the available markers have little power in predicting the outcome. To comprehensively assess a risk model, Pepe *and others* (2008) advocated the use of a predictiveness curve in conjunction with discriminatory measures. However, such an approach could not be directly applied to evaluate the performance of RS-guided intervention. In the context of evaluating the incremental value of a new marker for risk reclassification, Pencina *and others* (2008) proposed to measure the net reclassification improvement (NRI) based on the proportion of subjects reclassified into higher- or lower-risk categories. The NRI can be used to compare RS rules but not to evaluate a single RS rule. Furthermore, the NRI does not account for the differential costs associated with different types of incorrect assignment.

The ultimate value of an RS rule can be represented as the extra total cost/benefit if the RS-guided intervention applied to the target population. Therefore, to effectively construct and evaluate an RS rule, one should have information on the financial and medical costs/benefits associated with the interventions. Therefore, an ideal data set to evaluate the RS rule would consist of patients whose intervention status is known. With such a data set along with the cost/benefit information on the interventions, one may comprehensively evaluate an RS rule based on the expected cost associated with incorrect assignment of risk groups. In this paper, we propose a unified framework to determine the optimal risk categorization and quantify the value of the corresponding RS rule based on the expected costs when the cost parameters are assumed to be given. As a simple example, patients may be stratified into low or high risk groups where low-risk patients would be managed without intervention and high-risk patients would receive a treatment. Two types of costs may arise from such a stratification: the unnecessary intervention for controls, denoted

by C_0; and the cost of not receiving treatment for cases, denoted by C_1. When evaluating an RS rule that differentiates the high- and low-risk patients, it is important to account for the trade-off between these 2 types of costs (Cantor *and others*, 1999; Obuchowski, 2003) and develop an RS rule with a cutoff value that optimizes the trade-off between these costs. In Section 2, we discuss the relationship between costs and optimal threshold values of an RS rules based on a single marker. Procedures for comparing multiple RS rules are also discussed. These procedures are generalized to the setting where multiple risk factors are available for RS in Section 3. The proposed methods are illustrated in Section 3 with a data set from the Cardiovascular Health Study (CHS) and simulation studies. Some remarks are given in Section 5.

2. Optimal RS rules with a single marker

Let T denote the time to developing a clinical event and suppose interest lies in predicting the risk of failing by time t, that is, $Y = I(T \leqslant t)$. We first consider the setting where a single continuous marker Z is used to construct RS rules. Z could be a biomarker or a composite risk score established in the literature. Without loss of generality, we assume that the goal is to assign subjects into $k = 1, \ldots, K$ increasingly ordered risk categories.

2.1 *Optimal threshold values of RS and prespecified costs*

Let $R(z)$: $(-\infty, \infty) \to \{1, \ldots, K\}$ denote the risk group assignment with marker $Z = z$ based on its predicted risk $m(z)$. A subject would be assigned to a low-risk category if $m(z)$ is close to 0 and to a high-risk category if $m(z)$ is sufficiently larger than 0. The risk threshold values for the optimal RS rule are directly related to the costs associated with incorrect assignment of risk groups. An optimal rule would assign cases to the highest risk category and controls to the lowest category. Let c_{1k} and c_{0k} denote the cost associated with assigning cases and controls to the kth risk category, respectively. Then we expect that $c_{11} > c_{12} > \cdots > c_{1K}$ and $c_{01} < c_{02} < \cdots < c_{0K}$. Without loss of generality, we assume that $c_{1K} = c_{01} = 0$. Under this assumption, c_{1k} essentially represents the additional cost incurred by assigning a case to the kth category as opposed to the highest category; and c_{0k} represents the additional cost associating with assigning a control to the kth category when compared to the lowest category.

We propose to summarize the performance of $R(z)$ for the subpopulation with marker value $Z = z$ using the expected cost associated with the stratification:

$$\mathfrak{C}_z(R) = E\{Yc_{1R(z)} + (1 - Y)c_{0R(z)} | Z = z\}.$$

We show in online Appendix A that among all possible stratification rules based on Z, the optimal stratification rule that achieves the lowest expected cost $\mathfrak{C}_z(R)$ is

$$R^{\text{opt}}(z) = \min\left\{k: \max_{0 \leqslant l \leqslant k-1} P_{kl} \leqslant \mu_0(z) \leqslant \min_{k+1 \leqslant l \leqslant K+1} P_{kl}\right\},$$

where $\mu_0(z) = \text{pr}(Y = 1 | Z = z)$, $P_{k0} = 0$, $P_{k(K+1)} = 1$, $P_{kl} = 1/(1 + r_{kl})$, and $r_{kl} = r_{lk} = (c_{1l} - c_{1k})/(c_{0k} - c_{0l})$, for $1 \leqslant l, k \leqslant K$. Here, r_{kl} is the incremental cost by moving a diseased subject from risk category k to l relative to the incremental cost by moving a disease-free subject from risk category l to k.

For a given set of cost parameters, the optimal RS rule may suggest that not all K categories are necessary. As an example, we show the optimal RS rule for $K = 3$ in Table 1. When the relative cost $r_{32} \geqslant r_{12}$, the optimal rule classifies no subject as intermediate risk suggesting that there is no gain of having the intermediate risk category under such a condition. In general for assigning K risk categories,

Table 1. *Optimal stratification rules under various configurations when $K = 3$. Here, $\emptyset$ represents an empty set suggesting that no subject would be assigned to the risk category*

$R^{\text{opt}}(z)$	$r_{32} < r_{21}$	$r_{32} \geqslant r_{21}$
1	$\mu_0(z) \leqslant P_{21}$	$\mu_0(z) \leqslant P_{31}$
2	$\mu_0(z) \in (P_{21}, P_{32}]$	$\emptyset$
3	$\mu_0(z) > P_{32}$	$\mu_0(z) > P_{31}$

the optimal RS rule will not contain empty cells or unnecessary risk strata if and only if

$$r_{21} > r_{32} > \cdots > r_{K(K-1)}. \tag{2.1}$$

Under such an assumption, the optimal RS rule is

$$R^{\text{opt}}(z) = \sum_{l=1}^{K} I\{\mu_0(z) \leqslant p_k\}, \tag{2.2}$$

$$\text{where} \quad p_k = P_{(k+1)k} = \frac{1}{1 + r_{(k+1)k}}, \tag{2.3}$$

which infers that

$$c_{1k-1} = c_{1k} + (1 - p_{k-1})(c_{0k} - c_{0k-1})/p_{k-1}. \tag{2.4}$$

Equations (2.3) and (2.4) characterize the relationship between optimal threshold values of an RS and the cost parameters.

2.2 *The expected cost of an RS rule with optimal threshold values*

Suppose the optimal threshold values, $\mathbf{p} = (p_1, \ldots, p_K)^{\mathrm{T}}$, are used to create K increasing risk categories. A subject with predicted risk $m(z)$ will be assigned to the kth risk category if $m(z) \in (p_{k-1}, p_k]$, where $0 = p_0 < p_1 < \cdots < p_K = 1$. The overall performance of such a stratification rule can be evaluated based on the expected cost, $\mathfrak{C}(m) = E\{\mathfrak{C}(m, Z)\}$, where

$$\begin{aligned}\mathfrak{C}(m, Z) &= \sum_{k=1}^{K} I\{m(Z) \in (p_{k-1}, p_k]\} E[\{Y c_{1k} + (1 - Y)c_{0k}\} \mid Z] \\ &= c_{11} E(Y \mid Z) + \sum_{k=1}^{K-1} d_{0k} I\{m(Z) \geqslant p_{k-1}\}\{1 - E(Y p_{k-1}^{-1} \mid Z)\},\end{aligned}$$

where $c_{11} = \sum_{k=2}^{K}(1 - p_{k-1})(c_{0k} - c_{0(k-1)})/p_{k-1}$ and $d_{0k} = c_{0(k+1)} - c_{0k}$. Utilizing (2.4), we represent $\mathfrak{C}(m)$ in terms of $\mathbf{p}$ and $\mathbf{c}_0 = (c_{01}, \ldots, c_{0K})'$. The advantage of this representation is that c_{0k} is relatively easy to ascertain based on the financial cost of applying the corresponding intervention to a healthy subject if one is willing to ignore the side effects of the intervention. On the contrary, it is generally difficult to determine c_{1k}, the cost of cases receiving the incorrect intervention. Obviously, the RS with the true conditional risk function $\mu_0(\cdot)$, achieves the lowest expected cost among all RS rules based on Z.

Since the commonly used risk threshold values are often derived from a series of careful adjustments based on the empirical results from long-term clinical practice, it is not unreasonable to assume that

such threshold values of a well-established RS rule are "optimal" with respect to a set of underlying cost values, which are implicitly accepted by public. Under such an optimality assumption, one may use $\mathfrak{C}(m)$ to evaluate the RS rule. Furthermore, the expected cost function $\mathfrak{C}(\mu_0)$ provides a mechanism for comparing RS rules based on different risk scores. For example, if 2 risk scores $Z^{(1)}$ and $Z^{(2)}$ are available for RS, one may prefer the risk score $Z^{(1)}$ over $Z^{(2)}$ if $\mathfrak{C}(\mu_{10}) < \mathfrak{C}(\mu_{20})$ and $Z^{(2)}$ over $Z^{(1)}$, otherwise, where $\mu_{j0}(z) = \mathrm{pr}(Y = 1 \mid Z^{(j)} = z)$, $j = 1, 2$. When the risk scores involve potentially expensive or invasive markers, one may also incorporate the cost associated with the ascertainment of the risk scores when comparing their performances. Specifically, let C_j be the average cost associated with ascertaining the risk score $Z^{(j)}$, then the expected costs associated with $Z^{(j)}$ is $\mathfrak{C}(\mu_{j0}) + C_j$. Thus, one may prefer $Z^{(1)}$ over $Z^{(2)}$ if $\mathfrak{C}(\mu_{10}) + C_1 < \mathfrak{C}(\mu_{20}) + C_2$.

2.3 *Evaluating the optimal RS rules*

Let T_i and Z_i denote the event time and marker value for the ith subject, respectively. Due to censoring, for T_i, one observes (X_i, δ_i), where $X_i = \min(T_i, T_i^{\mathrm{cen}})$, $\delta_i = I(T_i \leqslant T_i^{\mathrm{cen}})$, and T_i^{cen} are the follow-up time for the ith subject assumed to be independent of T_i and Z_i with a common $G(t) = \mathrm{pr}(T^{\mathrm{cen}} \geqslant t)$. Data for analysis consist of n i.i.d. random vectors, $\{(X_i, \delta_i, Z_i), i = 1, \ldots, n\}$.

Estimating expected cost. Without censoring, the expected cost associated with a risk score $m(z)$ based on known $\mathbf{p}$ and $\mathbf{c}_0$, may be estimated nonparametrically by

$$\widetilde{\mathfrak{C}}(m) = n^{-1} \sum_{i=1}^{n} \eta\{Y_i, m(Z_i)\},$$

where $Y_i = I(T_i \leqslant t)$ and $\eta\{Y_i, m(Z_i)\} = c_{11} Y_i + \sum_{k=1}^{K-1} d_{0k}(1 - Y_i p_k^{-1}) I\{m(Z_i) > p_{k-1}\}$. However, Y_i is not always observable due to censoring. To incorporate censoring, we propose to modify $\widetilde{\mathfrak{C}}(m)$ based on the inverse probability weighting (IPW) estimator

$$\widetilde{\mathfrak{C}}(m) = n^{-1} \sum_{i=1}^{n} \widehat{\omega}_i \eta\{Y_i, m(Z_i)\}, \tag{2.5}$$

where $\widehat{\omega}_i = \{I(X_i \leqslant t)\delta_i + I(X_i > t)\}/\widehat{G}(t \wedge X_i)$ and $\widehat{G}(\cdot)$ is the Kaplan–Meier estimator of $G(t)$.

Estimating the conditional risk. The true conditional risk function $\mu_0(z) = \mathrm{pr}(Y = 1 | Z = z)$ involved in the optimal RS is unknown in general. To estimate $\mu_0(z)$ nonparametrically, we consider the use of the local logistic likelihood estimator (Tibshirani and Hastie, 1987) with IPW to account for censoring. Specifically, we estimate $\mu_0(z)$ by $\widetilde{\mu}(z) = g_0\{\widetilde{\theta}_0(z)\}$, where $g_0(x) = e^x/(1 + e^x)$, $\{\widetilde{\theta}_0(z), \widetilde{\theta}_1(z)\}$ is the solution to the local IPW score equation

$$\widetilde{\mathcal{S}}_z(\theta_0, \theta_1) = \sum_{i=1}^{n} \widehat{\omega}_i \begin{pmatrix} 1 \\ Z_i - z \end{pmatrix} K_h(Z_i - z)[Y_i - g_0\{\theta_0 + \theta_1(Z_i - z)\}], \tag{2.6}$$

where $K_h(z) = h^{-1}K(z/h)$, $K(\cdot)$ is a smooth symmetric density function and h is the bandwidth with $h \to 0$ and $nh^2 \to \infty$ as $n \to \infty$. In practice, the bandwidth h for estimating the conditional risk function may be selected via $\mathcal{K}$-fold cross validation.

Interval estimation procedures for $\mathfrak{C}(\mu_0)$. To obtain interval estimates for $\mathfrak{C}(\mu_0)$, we show in online Appendix B that $\widetilde{\mathfrak{C}}(\widetilde{\mu})$ is consistent for $\mathfrak{C}(\mu_0)$. Furthermore, under mild regularity conditions, $n^{\frac{1}{2}}\{\widetilde{\mathfrak{C}}(\widetilde{\mu}) - \mathfrak{C}(\mu_0)\}$ converges in distribution to a zero-mean normal with variance σ^2. A 95% confidence interval for $\mathfrak{C}(\mu_0)$ may be obtained as $\widetilde{\mathfrak{C}}(\widetilde{\mu}) \pm 1.96\widehat{\sigma}$, where $\widehat{\sigma}$ is a consistent estimator of σ obtained by replacing all the theoretical quantities in σ by their empirical counterparts.

When there are multiple markers available, one may compare the performance in RS based on the difference between the expected costs. For example, when there are 2 risk scores, $Z^{(1)}$ and $Z^{(2)}$, one may compare their corresponding RS performances based on $d_{\mathfrak{C}} = \mathfrak{C}_1(\mu_0^{(1)}) - \mathfrak{C}_2(\mu_0^{(2)})$, where

$$\mathfrak{C}_j(m) = \sum_{k=1}^{K} E[I\{m(Z^{(j)}) \in (p_{k-1}, p_k]\}\{Yc_{1k} + (1-Y)c_{0k}\}].$$

$d_{\mathfrak{C}}$ may be consistently estimated by $\widetilde{\Delta}_{\mathfrak{C}} = \widetilde{\mathfrak{C}}_1(\widetilde{\mu}^{(1)}) - \widetilde{\mathfrak{C}}_2(\widetilde{\mu}^{(2)})$, where $\widetilde{\mu}^{(j)}(\cdot)$ is the estimated risk function based on $Z^{(j)}$ and $\widetilde{\mathfrak{C}}_j(m) = n^{-1}\sum_{i=1}^{n} \widehat{w}_i \eta\{Y_i, m(Z_i^{(j)})\}$, for $j = 1, 2$. Using similar arguments as given in online Appendix B, one may show that $n^{\frac{1}{2}}(\widetilde{\Delta}_{\mathfrak{C}} - d_{\mathfrak{C}})$ converges in distribution to a zero-mean normal with variance σ_{Δ}^2. The confidence intervals can be constructed accordingly.

3. APPROXIMATING OPTIMAL RS RULES BASED ON MULTIPLE MARKERS

3.1 *Developing RS rules*

When there are a panel of markers, denoted by a column vector $\mathbf{Z}$, available for risk prediction, one may derive RS rules by first ascertaining the conditional risk function, $\mu_0(\mathbf{Z}) = \mathrm{pr}(Y = 1 \mid \mathbf{Z})$. For the subpopulation with $\mathbf{Z} = \mathbf{z}$, the optimal RS rule may be constructed by replacing $\mu_0(z)$ in (2.2) with $\mu_0(\mathbf{z})$. In general, for any given risk prediction function $\mu(\mathbf{z})$ and cost parameters, one may construct an RS rule based on $R_{\mu}(\mathbf{Z}) = \sum_{k=1}^{K} I\{\mu(\mathbf{Z}) \leqslant p_k\}$, where p_k is given in (2.3). The total expected cost associated with such a rule for is $\mathbb{C}(\mu) = E\{\mathbb{C}(\mu, \mathbf{Z})\}$, where

$$\mathbb{C}(\mu, \mathbf{Z}) = \sum_{k=1}^{K} I\{\mu(\mathbf{Z}) \in (p_{k-1}, p_k]\}E[Yc_{1k} + (1-Y)c_{0k} \mid \mathbf{Z}].$$

The true conditional risk function $\mu_0(\cdot)$ minimizes the total expected cost $\mathbb{C}(\mu)$ among all functions of $\mathbf{Z}$. To approximate the optimal RS rule based on available data, one needs to estimate $\mu_0(\mathbf{Z})$. When the number of markers is not small, it is implausible to estimate $\mu_0(\mathbf{Z})$ nonparametrically. A practical approach is to approximate the conditional risk through regression modeling. For example, one may consider regression models such as the Cox proportional hazards model (Cox, 1972), semiparametric transformation models (Cheng *and others*, 1995) or time-specific generalized linear models (GLMs) (Zheng *and others*, 2006; Uno *and others*, 2007). When the assumed regression model is the true model, one may estimate $\mu_0(\mathbf{Z})$ consistently by fitting the regression model.

In practice, simple regression models may fail to hold. As such, the estimated conditional risk function may be a poor approximation to the true conditional risk and thus leads to suboptimal RS rules. Furthermore, inference procedures about the performance of the RS rule may be invalid if such procedures are derived under the assumption of correct model specification. To overcome such difficulties, we propose to employ simple statistical models as "working" models for approximating the true conditional risk and derive procedures for making inference about the performance of RS rules without requiring the fitted model to hold. A wide range of survival models including those mentioned above may be considered as the working model. A simple example is to model the conditional risk function $\mu_0(\mathbf{Z})$ via the time-specific GLM,

$$\mathrm{pr}(Y = 1 \mid \mathbf{Z}) = \mathrm{pr}(T \leqslant t \mid \mathbf{Z}) = g(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}), \tag{3.1}$$

where $g(\cdot)$ is a prespecified monotone and smooth link function and $\boldsymbol{\beta}$ is an unknown regression coefficient. Without loss of generality, we assume that the vector $\mathbf{Z}$ includes 1 as its first component. Here,

both g and $\boldsymbol{\beta}$ could may vary with time t. Through the working model (3.1), one may approximate the conditional risk for a subject with $\mathbf{Z} = \mathbf{z}$ as $g(\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{z})$, where $\widehat{\boldsymbol{\beta}}$ is the solution to

$$n^{-1}\sum_{i=1}^{n}\hat{\omega}_i\mathbf{Z}_i\{Y_i - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}_i)\} = 0. \tag{3.2}$$

Uno *and others* (2007) showed that $\widehat{\boldsymbol{\beta}}$ is always convergent to $\boldsymbol{\beta}_0$, the unique solution to $E[\mathbf{Z}\{Y - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z})\}] = 0$.

Based on the working model (3.1), one may construct RS rule using the risk prediction function $g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z})$. However, if the working model (3.1) fails to be a good approximation to the true model, it is unclear whether such an RS rule is optimal in any sense and may not perform well. To improve the RS, we propose to use $\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}$ as a scoring system and predict the risk for a subject with $\mathbf{Z}$ as $\mu_{\boldsymbol{\beta}_0}(\mathbf{Z})$, where $\mu_{\boldsymbol{\beta}}(\mathbf{Z}) = \mathrm{pr}(Y = 1 \mid \boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z})$. An optimal RS rule based on $\mu_{\boldsymbol{\beta}_0}(\mathbf{Z})$ may be constructed as in (2.2). Such a rule would be optimal, with respect to $\mathbb{C}(\mu)$, among all rules based on $\mathbf{Z}$ if the working model holds and optimal within rules based on the linear risk score $\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}$ if the working model fails to hold. Compared with $g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}^0)$, it is straightforward to show that

$$E[\{g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}^0) - \mu_0(\mathbf{Z}^0)\}^2] \geqslant E[\{\mu_{\boldsymbol{\beta}_0}(\mathbf{Z}^0) - \mu_0(\mathbf{Z}^0)\}^2]$$

and we expect that RS rules based on $\mu_{\boldsymbol{\beta}}(\mathbf{Z})$ will have lower expected cost compared to that based on $g(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z})$.

For any given $\mathbf{z}$, a consistent estimate of $\mu_{\boldsymbol{\beta}_0}(\mathbf{z})$ may be obtained as $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}(\mathbf{z})$, where $\widehat{\mu}_{\boldsymbol{\beta}}(\mathbf{z})$ is the nonparametric local likelihood estimator similar to that proposed in Section 2.3 based on the synthetic data $\{(X_i, \delta_i, \widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{Z}_i), i = 1, \ldots, n\}$.

3.2 *Evaluating RS rules based on the total expected cost*

The expected cost associated with $\mu_{\boldsymbol{\beta}_0}(\mathbf{Z})$ averaged over the population, $\mathbb{C}(\mu_{\boldsymbol{\beta}_0}) = E\{\mathbb{C}(\mu_{\boldsymbol{\beta}_0}, \mathbf{Z})\}$ can be estimated by $\widehat{\mathbb{C}}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}})$, where

$$\widehat{\mathbb{C}}(m) = n^{-1}\sum_{i=1}^{n}\widehat{\omega}_i I\{m(\mathbf{Z}_i) \in (p_{k-1}, p_k]\}\{Y_i c_{1k} + (1 - Y_i)c_{0k}\}. \tag{3.3}$$

In online Appendix C, we demonstrate that $\widehat{\mathbb{C}} = \widehat{\mathbb{C}}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}})$ is a consistent estimator of $\mathbb{C}(\mu_{\boldsymbol{\beta}_0})$. Furthermore, $n^{\frac{1}{2}}\{\widehat{\mathbb{C}} - \mathbb{C}(\mu_{\boldsymbol{\beta}_0})\}$ converges in distribution to a normal with mean 0 and variance $E(\zeta_{\mathbb{W}i}^2)$, where $\zeta_{\mathbb{W}i}$ is defined in (C.3) of online Appendix C.

As for most model evaluation measures, $\widehat{\mathbb{C}}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}})$ is likely to underestimate the total expected cost associated with the RS due to overfitting, especially when sample size is not large compared to the number of markers. An effective approach to reducing the overfitting bias is the cross validation. We consider the commonly used $\mathcal{K}$-fold cross validation, which randomly splits the data into $\mathcal{K}$ disjoint sets of about equal size and labels them as $\mathcal{I}_\kappa, \kappa = 1, \ldots, \mathcal{K}$. For each κ, based on all observations which are not in $\mathcal{I}_\kappa$, we obtain an estimate $\widehat{\boldsymbol{\beta}}_{(-\kappa)}$ for $\boldsymbol{\beta}$ via (3.2) and subsequently an estimate $\widehat{\mu}_{\boldsymbol{\beta}}^{(-\kappa)}(\mathbf{z})$ for $\mu_{\boldsymbol{\beta}}(\mathbf{z})$. Based on $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}_{(-\kappa)}}^{(-\kappa)}(\mathbf{z})$, we then compute the total expected cost estimate $\widehat{\mathbb{C}}_{(\kappa)}(m)$ via (3.3) based on observations in $\mathcal{I}_\kappa$. Then, a bias corrected estimate for $\mathbb{C}(\mu_{\boldsymbol{\beta}_0})$ is

$$\widehat{\mathbb{C}}^{(cv)} = \mathcal{K}^{-1}\sum_{\kappa=1}^{\mathcal{K}}\widehat{\mathbb{C}}_{(\kappa)}\left\{\widehat{\mu}_{\widehat{\boldsymbol{\beta}}_{(-\kappa)}}^{(-\kappa)}\right\}. \tag{3.4}$$

For any fixed $\mathcal{K}$, it is straightforward to show that $\widehat{\mathbb{C}}_{cv}$ is consistent for $\mathbb{C}(\mu_{\boldsymbol{\beta}_0})$. Using arguments given in Tian *and others* (2007), it is not difficult to show that the standardized $\widehat{\mathbb{C}}_{cv}$, $\mathcal{W} = n^{1/2}\{\widehat{\mathbb{C}}^{(cv)} - \mathbb{C}(\mu_{\boldsymbol{\beta}_0})\}$ has the same limiting distribution as that of $n^{\frac{1}{2}}\{\widehat{\mathbb{C}} - \mathbb{C}(\mu_{\boldsymbol{\beta}_0})\}$. Therefore, one may use the standard error estimate based on (C.3) of the online Appendix to construct interval estimates for $\mathbb{C}(\mu_{\boldsymbol{\beta}_0})$, which are centered around the cross-validation estimate.

4. NUMERICAL STUDIES

4.1 *Example: CHS*

We illustrate our methods by evaluating stratification rules for predicting the risk of coronary heart disease (CHD) using data from the CHS sponsored by the National Heart, Lung and Blood Institute. The CHS is a population-based observational prospective study of risk factors for cardiovascular disease in adults 65 years or older. A full description of the design of CHS is reported in Fried *and others* (1991). One of the most widely used clinical prediction score for CHD risk is the Framingham risk score (FR-score). The FR-score was originally derived from proportional hazards models by Anderson *and others* (1991) and updated by Wilson *and others* (1998) based on the Framingham heart study. Separate models were fitted for men and women with predictors including age, blood cholesterol, high-density lipoprotein (HDL) cholesterol, blood pressure, present smoking status, and diabetes mellitus. We construct the FR-score based on the coefficients given in Table 6 of Wilson *and others* (1998). Since FR-score may have different ability in RS among men and women, we evaluate its performance separately for the 2 populations and only use women for illustration. We are interested in evaluating the performance of various risk scores in stratifying patients into different risk categories for the occurrence of CHD events within 10 years. To apply the proposed procedures, we ideally need (i) RS rules with optimal threshold values and (ii) a good estimate of cost $\mathbf{c}_0$. Since some CHS patients may have already received their intervention per American Heart Association (AHA) guideline, the well-accepted risk threshold values of 10% and 20% may not be optimal to CHS population and the accurate estimation of $\mathbf{c}_0$ becomes complicated. While acknowledging these limitations, we still simply assume the optimality of the threshold values of RS rules of interest and estimate $\mathbf{c}_0$ as if no one had received the intervention for illustration purpose.

The analysis here includes 3313 females who have available information on the baseline FR-score variables and event times. Subjects in this data set were between 65 and 95 years old with a median age of 71. There was little loss to follow up in CHS and the median follow-up time was 14.47 years. There were about 26.2% of subjects who experienced a CHD event during follow up and 19.6% of subjects experienced a CHD event within 10 years. 43.0% were censored and 30.8% (1020) subjects in the sample died from other causes without a CHD. Since the RS rules were developed for the prevention of CHD, we focused our analysis on CHD events only and thus define $Y_i = 1$ if subject i experienced CHD within 10 years since and $Y_i = 0$ if she did not experience CHD or died of other causes within 10 years. Thus, the model (3.1) is assumed for the subdistribution of CHD. To estimate the censoring probability for the IPW weights, we note that censoring occurs only if a patient drops out of the study prior to the occurrence of CHD or death.

To construct RS rules based on the FR-score, we follow the current guideline from the AHA (Mosca *and others*, 2004, 2007) and consider a stratification rule which assigns a patient with FR-score $= z$ into the low risk group if $\mu_0(z) \leqslant 0.10$; the intermediate risk group if $0.10 < \mu_0(z) \leqslant 0.20$; and the high risk group if $\mu_0(z) > 0.20$, where $\mu_0(\cdot)$ is estimated based on the local IPW likelihood estimator discussed in Section 2.3. Lifestyle interventions were recommended for all women. For patients in the intermediate risk group, antihypertensive therapy such as with a thiazide diuretic was recommended. The AHA guidelines call for simultaneous lifestyle interventions and statin therapy for patients with high risk. Based on these guidelines and the yearly cost of the corresponding medications, we assume that the cost associated with assigning a patient who will not experience CHD events to the low, intermediate, and high risk groups

to be 0, \$240, and \$600, which were calculated based on the annual costs for hydrochlorothiazide at the dosage of 12.5 mg per day and simvastatin at the doseage of 40 mg per day. These parameters lead to an estimated average yearly cost of \$454 (per person) with a 95% confidence interval (\$439, \$469) for the RS rule based on the FR-score.

Cook *and others* (2006) advocated the inclusion of C-reactive protein (CRP) for predicting cardiovascular risk for women. They derived a risk prediction model based on the women's health study by including age, systolic blood pressure, antihypertensive use, present smoking status, log(HDL), log(total cholestrol), and log(CRP). We constructed the risk score based on the coefficients provided in Table 1 of Cook *and others* (2006). In Table 2, we show the proportion of subjects stratified into each of the risk categories based on the FR-score and based on the new score with CRP. Overall, the FR-score appears to assign most subjects into the intermediate risk group. The new score with CRP appears to assign more cases, that is, subjects experienced a CHD event within 10 years, to the high risk groups. For subjects who did not experience a CHD event within cases, the new score assigns 5.3% of those to the low risk group, 42.7% to the intermediate risk group, and 32.4% to the high risk group. To assess the overall effectiveness of the RS rule based on the new score, we use the same cost parameters as given above and obtained an estimated average yearly cost of \$431 with a 95% confidence interval (\$416, \$446). To compare the RS rules based on the FR-score and the score incorporating the CRP, we estimated the cost reduction due to CRP $d_{\mathfrak{C}}$ as \$23 with 95% confidence interval (\$5, \$42), suggesting that including the CRP information could potentially improve the accuracy of RS with respect to the expected cost. On the other hand, we note that the cost of the fully automated quantitative CRP test based on the quantitative immunoassay is reported to be around \$50.00. When the additional cost of the CRP is taken into account, it is unclear whether the improvement in RS due to CRP is substantial enough to recommend CRP for the general population.

Instead of using the FR-score or score given by Cook *and others* (2006), we also constructed RS rule with risk factors based on the CHS data. We first fit a logistic regression working model relating the risk factors to the binary response $Y = I(T \leqslant 10)$ and obtained a risk score $\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{Z}$. The risk of a subject experiencing CHD within 10 years is predicted based on the aforementioned nonparametric local likelihood estimator. Subsequently, the total expected cost was estimated via (3.3) and also (3.4) with 5-fold cross validation. Three models are considered: (i) age only; (ii) all variables but log(CRP); and (iii) full model. The point and interval estimates for the expected cost under these settings are given in Table 3. With the refitting, the RS rule derived under the full model results in an average cost of \$418 with standard error \$9 based on the cross-validated estimate. This cost is only slightly lower than the average cost for the RS rule obtained based on the score provided by Cook *and others* (2006).

To evaluate the incremental value of log(CRP), we compare the expected cost associated with the RS derived from the full model and the model without CRP in terms of the difference in the total expect cost. In this example, the bias-corrected estimate of the incremental value of CRP is \$3 (95% confidence interval [−9,15]). This confirms that there is a minimal gain for having the extra CRP information when we obtain risk estimates by fitting models (ii) and (iii).

Table 2. *Proportion of subjects who have (not) experienced a cardiovascular event within* 10 *years and are assigned to low, intermediate, and high risk groups based on the FR-score and the score with CRP proposed by Cook and others (2006)*

	Low	With CHD intermediate	High	Low	without CHD intermediate	High
FR-score	0.000	0.145	0.052	0.002	0.653	0.149
Score with CRP	0.005	0.081	0.111	0.053	0.425	0.325

Table 3. *Estimated expected cost* $\mathbb{C}$ *and the incremental values based on the apparent error estimate and the cross-validated estimator. Shown also are the standard error estimates (StdErr) and the lower and upper bounds of the* 95% *confidence intervals (CIs)*

		Apparent	Cross validated	StdErr	95% CI bounds	
					Lower	Upper
$\mathbb{C}$	(i) Age only	440	444	7	429	458
	(ii) w/o CRP	417	422	9	404	439
	(iii) Full model	411	418	9	401	436
$d_{\mathbb{C}}$	Full versus age only	29	25	9	8	42
	Full versus w/o CRP	6	3	6	−9	15

Table 4. *Bias, sampling standard error (SSE), average of the estimated standard error (ASE), and empirical coverage levels of the* 95% *confidence intervals (CovP) for the estimated costs under settings when there is a single marker. For each configuration, results are summarized based on* 1000 *simulated data sets*

n	Truth	Bias	SSE	ASE	CovP
200	151.53	−1.03	20.73	19.76	0.94
400	151.53	−0.43	14.03	13.98	0.96

4.2 *Simulation studies*

Simulation studies were conducted to evaluate the finite sample performance of the proposed procedures. We generated marker value Z from an exponential distribution with mean rate 10 and the survival time T from a log-normal model

$$\log T = \log(100) - (Z + \sqrt{Z} + 1) + \epsilon, \tag{4.1}$$

with $\epsilon \sim N(0, 1)$. We generated the censoring C from a log-normal distribution which resulted in about 20% of censoring. For each simulated data set, we constructed RS rules for predicting 10-year survival similar to the CHS example. That is, we assign subjects into the low, intermediate, and high-risk categories, determined by risk intervals, $(0, 0.1]$, $(0.1, 0.2]$, and $(0.2, 1.0]$, respectively. We also assume that the cost associated with assigning a patient who will not fail within 10-years to the low, intermediate, and high risk groups to be 0, \$240, and \$600. Under such configuration, the total cost associated with the optimal RS rule is $\mathfrak{C} = \$151.53$. For all the simulation studies considered here, we estimated the conditional risk function nonparametrically to ensure the consistency of the estimators. The bandwidth for estimating the conditional risk function was selected via 5-fold cross validation by minimizing the mean squared error, as described in Section 2.3. The results for sample sizes $n = 200$ and 400 are summarized in Table 4. In general, the estimated expected cost has little bias. The estimated standard errors are close to the empirical standard errors and the confidence intervals have proper coverage levels.

To evaluate the performance of the proposed procedure for comparing multiple markers, we generated markers Z_1 and Z_2 from a multivariate normal with mean 0, unit variance and correlation 0.3. The survival time T was generated from

$$\log T = \log(20) - Z_1 - 0.5Z_2 + \epsilon$$

with $\epsilon \sim N(0, 0.5^2)$. The censoring was generated from a log-normal with mean 4 and unit variance 1 which resulted in about 30% of censoring. For each data set, we obtained point and interval estimates for the expected costs associated with optimal RS rules based on each of the 2 markers. To compare the performance of the 2 stratification rules, we obtained point and interval estimates for $d_{\mathfrak{C}}$, the difference in

Table 5. *Bias, sampling standard error (SSE), average of the estimated standard error (ASE), and empirical coverage levels of the* 95% *confidence intervals (CovP) for the estimated costs under settings, when there are* 2 *markers for comparison. For each configuration, results are summarized based on* 1000 *simulated data sets*

	Truth	$n = 200$				$n = 400$			
		Bias	SSE	ASE	CovP	Bias	SSE	ASE	CovP
Marker 1	191.03	0.92	32.80	32.13	0.93	0.17	22.81	23.05	0.94
Marker 2	326.67	−8.78	31.91	32.94	0.94	−3.74	22.94	23.87	0.95
Difference	135.64	−9.70	46.95	47.51	0.94	−3.91	33.99	34.31	0.94

the expected cost. The results were summarized in Table 5 for sample sizes $n = 200$ and 400. Similar to the previous setting, the proposed inference procedures generally perform well in finite samples. At sample size of 200, the estimator for $d_{\mathfrak{C}}$ has about 7% of bias. This is partially due to the difficulty in estimating the conditional risk function nonparametrically at sample size of 200. However, the bias decreases as the sample size increases as we expected and all the interval estimators have proper coverage levels.

5. Discussion

In an ideal cost/benefit analysis, one needs to first specify the decision-makers's cost function, which amounts to determine all the values of $\bar{\mathbf{c}} = \{(c_{1k}, c_{0k}), 1 \leqslant k \leqslant K\}$ in evaluating an RS rule. In this paper, we proposed to circumvent the difficulty of assigning $\bar{\mathbf{c}}$ by using the established correspondence between $\bar{\mathbf{c}}$ and $\mathbf{p}$, if it is reasonable to assume the optimality of certain risk threshold values employed by the medical community. While this is a useful practical approach, it remains desirable to specify $\bar{\mathbf{c}}$ via careful health economic analyses "*a priori*". Once a good assessment of $\bar{\mathbf{c}}$ becomes available, one may improve an existing RS rule by adjusting the risk threshold values and properly evaluate the adjusted RS rule. In our CHS example, we used very simple and naive cost parameters based on the financial cost alone. However, to comprehensively evaluate the performance of RS rules, it is crucial to consider all the financial and medical consequences of assigning cases and controls to different risk categories such as the probability of preventing subjects from developing CHD and the subsequent life-years saved. A comprehensive analysis should also take into account the current intervention the subjects are receiving and assess the cost/benefit of changing the current intervention to the RS rule–suggested intervention. Our current analysis of the CHS data assumed that the population is naive to the intervention of interest. It would also be interesting to explore other mechanisms to account for death due to other causes since no additional medication costs would incur after death. However, we do not have sufficient information on the benefit and risk of the treatments and thus the results provided in the example section may have limited applicability in practice.

One limitation of the proposed approach is to impose a universal cost function for the entire population without allowing for heterogeneous cost–benefit profiles. It is conceivable that the appropriate cost function and hence the corresponding optimal RS rule could be individualized for each patient. With any specified cost function for an individual patient, the proposed methods can still be used in constructing an optimal RS rule and provide the corresponding recommendation of appropriate clinical interventions. On the other hand, when making public policy and regulatory decisions, it remains crucial to develop a general guideline with stratification rules that are optimal on average over the entire population.

When there is a single marker or score, the proposed RS rule is constructed based on the conditional risk $\mu_0(z)$, which can be estimated nonparametrically. If one is willing to assume that $\mu_0(z)$ is a monotone function, then one may estimate $\mu_0(z)$ using a nonparametric isotonic regression techniques (Friedman

and Tibshirani, 1984; Bloch and Silverman, 1997; Hall and Huang, 2001). Last, in order for the total expected cost estimated from the current population applicable to a new population, the 2 populations need to have the same joint distribution of $(\mathbf{Z}, Y)$. If the future population has a different marginal distribution of $\mathbf{Z}$ and shares the same conditional risk function given $\mathbf{Z}$, then one may infer about the total expected cost for the new population via appropriate reweighting.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

The authors are grateful to the editor, the associate editor and referees for their insightful and constructive suggestions. *Conflict of Interest:* None declared.

FUNDING

National Institute of Health (R01-HL089778, R01-GM079330).

REFERENCES

ANDERSON, K. M., ODELL, P. M, WILSON, P. W. AND KANNEL, W. B. (1991). Cardiovascular disease risk profiles. *American Heart Journal* **121**, 293–298.

BLOCH, D. A. AND SILVERMAN, B. W. (1997). Monotone discriminant functions and their applications in rheumatology. *Journal of the American Statistical Association* **92**, 144–153.

CANTOR, S. B., SUN, C., TORTOLERO-LUNA, G., RICHARDS-KORTUM, R. AND FOLLEN, M. (1999). A comparison of c/b ratios from studies using receiver operating characteristic curve analysis. *Journal of Clinical Epidemiology* **52**, 885–892.

CHENG, S., WEI, L. J, AND YING, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–845.

COOK, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **115**, 928.

COOK, N. R., BURING, J. E. AND RIDKER, P. M. (2006). The effect of including c-reactive protein in cardiovascular risk prediction models for women. *Annals of Internal Medicine* **145**, 21–29.

FRIED, L. P., BORHANI, N. O., ENRIGHT, P., FURBERG, C. D., GARDIN, J. M., KRONMAL, R. A., KULLER, L. H., MANOLIO, T. A., MITTELMARK, M. B., NEWMAN, A. *and others* (1991). The cardiovascular health study: design and rationale. *Annals of Epidemiology* **1**, 263–276.

FRIEDMAN, J. AND TIBSHIRANI, R. (1984). The monotone smoothing of scatterplots. *Technometrics* **26**, 243–250.

GAIL, M. AND PFEIFFER, R. (2005). On criteria for evaluating models of absolute risk. *Biostatistics* **6**, 227–239.

HALL, P. AND HUANG, L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics* **29**, 624–647.

MOSCA, L., APPEL, L. J., BENJAMIN, E. J., BERRA, K., CHANDRA-STROBOS, N., FABUNMI, R. P., GRADY, D., HAAN, C. K., HAYES, S. N., JUDELSON, D. R. *and others* (2004). Evidence-based guidelines for cardiovascular disease prevention in women. *Circulation* **109**, 672–693.

MOSCA, L., BANKA, C. L., BENJAMIN, E. J., BERRA, K., BUSHNELL, C., DOLOR, R. J., GANIATS, T. G., GOMES, A. S., GORNIK, H. L., GRACIA, C., *and others* FOR THE EXPERT PANEL/WRITING GROUP (2007). Evidence-based guidelines for cardiovascular disease prevention in women: 2007 update. *Circulation* **115**, 1481–1501.

OBUCHOWSKI, N. (2003). Receiver operating characteristic curves and their use in radiology. *Radiology* **229**, 3–8.

PENCINA, M., D'AGOSTINO, R., D'AGOSTINO, R. AND VASAN, R. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**, 157–172.

PEPE, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford University Press.

PEPE, M. S., FENG, Z., HUANG, Y., LONGTON, G., PRENTICE, R., THOMPSON, I. M. AND ZHENG, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology* **167**, 362.

TIAN, L., CAI, T., GEOTGHEBEUR, E. AND WEI, L. J. (2007). Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* **94**, 297–311.

TIBSHIRANI, R. AND HASTIE, T. (1987). Local likelihood estimation. *Journal of American Statistical Association* **82**, 559–567.

UNO, H., CAI, T., TIAN, L. AND WEI, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of American Statistical Association* **102**, 527–537.

WILSON, P. W. F., D'AGOSTINO, R. B., LEVY, D., BELANGER, A. M., SILBERSHATZ, H. AND KANNEL, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837–1847.

ZHENG, Y., CAI, T. AND FENG, Z. (2006). Application of the time-dependent ROC curves for prognostic accuracy with multiple markers. *Biometrics* **62**, 279–287.

[*Received June 16, 2009; revised October 27, 2010; accepted for publication January 8, 2011*]