



Published in final edited form as:

Lifetime Data Anal. 2013 July ; 19(3): 350–370. doi:10.1007/s10985-012-9239-z.

Evaluating incremental values from new predictors with net reclassification improvement in survival analysis

Yingye Zheng,

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

Layla Parast,

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

Tianxi Cai, and

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

Marshall Brown

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

Abstract

Developing individualized prediction rules for disease risk and prognosis has played a key role in modern medicine. When new genomic or biological markers become available to assist in risk prediction, it is essential to assess the improvement in clinical usefulness of the new markers over existing routine variables. Net reclassification improvement (NRI) has been proposed to assess improvement in risk reclassification in the context of comparing two risk models and the concept has been quickly adopted in medical journals. We propose both nonparametric and semiparametric procedures for calculating NRI as a function of a future prediction time t with a censored failure time outcome. The proposed methods accommodate covariate-dependent censoring, therefore providing more robust and sometimes more efficient procedures compared with the existing nonparametric-based estimators. Simulation results indicate that the proposed procedures perform well in finite samples. We illustrate these procedures by evaluating a new risk model for predicting the onset of cardiovascular disease.

Keywords

Inverse probability weighted (IPW) estimator; Net reclassification improvement (NRI); Risk prediction; Survival analysis

1 Introduction

Developing individualized prediction rules for disease risk and prognosis is fundamental for successful disease prevention and treatment selection. For many diseases, risk prediction models have been developed and incorporated into clinical practice guidelines. For example, the Gail model was developed for predicting individual breast cancer risk (Gail et al. 1989) and a risk calculator based on that model can be used to assist physicians making screening

recommendations. For cardiovascular disease (CVD), prediction models such as the Framingham Risk Score (FRS) are used for stratifying patients into different levels of risks. However, much refinement is needed even for the best of these models because of their limited discriminatory accuracy. For example, the Framingham model, largely based on traditional clinical risk factors, has recognized limitations in its clinical utility (Hemann et al. 2007). A considerable fraction of patients who experienced CVD events had none of the identified risk factors, indicating a need to explore avenues beyond routine clinical measures for more accurate prediction (Khot et al. 2003). This fuels much of the current search for novel biologic markers and genetic factors that, when combined with routine clinical risk factors, may provide accurate prediction at the individual level.

When new genomic or biological markers become available to assist in risk prediction, it is essential to assess the clinical usefulness of these new markers compared to existing routine markers. Careful evaluation of the incremental value is particularly crucial when markers are either expensive or invasive to measure. To quantify the added clinical value of new markers over a conventional risk scoring system for predicting disease risk, one may calculate the difference in the prediction measures for the existing conventional model and the new model, which includes information from the new markers. For example the difference in the areas under the receiver operating characteristic curves (AUC of ROC) are often used to quantify the improvement in discrimination attributable to added markers. Since a risk model is often used to stratify patients into proper risk categories, statistical summaries that depend on clinically meaningful risk thresholds may be more relevant (Cook 2007; Cui 2009; Lloyd-Jones 2010). As an alternative to measuring the difference between AUCs, net reclassification improvement (NRI) has also been proposed to assess improvement in risk reclassification in the context of comparing two risk models constructed with and without novel markers (Pencina et al. 2008). Using “up” and “down” to denote changes in one or more risk categories in the upward and downward directions, respectively, for a subject between their baseline and augmented risk values, the NRI is defined as

$$\text{NRI} = [\text{Pr}(\text{up}|\text{Diseased}) + \text{Pr}(\text{down}|\text{Healthy})] - [\text{Pr}(\text{down}|\text{Diseased}) + \text{Pr}(\text{up}|\text{Healthy})].$$

Such a measure is appealing because it acknowledges both desirable risk reclassifications (up for diseased and down for healthy subjects) and undesirable risk reclassifications (down for diseased and up for healthy subjects). Due to its simplicity, NRI has been quickly adopted in medical journals. However, compared with many other measures for incremental values, the concept has not received much attention in the statistical literature.

Since a risk model is often used for predicting an individual's future outcome, it is essential to incorporate the additional dimension of time when assessing the performance of a risk model in a cohort study. For both deriving and evaluating risk models, prospective cohort data is often used. In this setting a subject's health status at a future time t is sometimes unknown due to loss of follow-up, termination of a study or the occurrence of a competing risk event. Such censoring poses additional challenges compared with settings previously examined in the literature which focus on incremental value calculation with a dichotomous outcome. Currently there is limited development in methods to estimate the incremental value of novel markers with censored failure time outcomes. Recently Pencina and D'Agostino (2011) proposed a method for calculating time-dependent NRI, based on nonparametric Kaplan–Meier (KM) estimators in order to account for censoring in cohort data. The asymptotic properties of a similar estimator is studied in detail in Uno et al. (2009). However, the validity of these estimators relies critically on the assumption that censoring is independent of predictors used in the risk models. Furthermore, the

nonparametric procedure considered in these estimators may potentially lead to efficiency loss. A more flexible and more efficient estimating procedure is needed in practice.

In this manuscript, we propose quantitative procedures for calculating NRI as a function of a future prediction time t with a censored failure time outcome. Compared with existing nonparametric estimators, our procedures do not require the assumption that censoring is independent of predictors, therefore the methods would be widely applicable to many practical situations. We also consider procedures that aim to improve efficiency while maintaining robustness. This manuscript is organized as follows. In Sect. 2, we specify models and define NRI suitable for event time outcomes. In Sect. 3, we describe procedures for estimating time-dependent NRI using data obtained from a prospective cohort study with a failure time outcome. We comment on the theoretical properties of our proposed estimators in Sect. 4. We then describe simulation studies to evaluate the performance of the proposed estimators. The results are reported in Sect. 5. An application of our procedures for comparing two CVD risk models is presented in Sect. 6. Concluding remarks are in Sect. 7.

2 Measures of risk stratification and reclassification

Consider the situation that a vector of predictor \mathbf{Y} measured at baseline is used for predicting the time to event outcome T . Risk models can be built using sub-vectors of \mathbf{Y} . Let $\mathbf{Y}_{(1)}$, a function of \mathbf{Y} , denote a vector of conventional predictor values in the existing model. Let $\mathbf{Y}_{(2)}$, also a function of \mathbf{Y} , denote a vector of predictors used in the new model that contains $\mathbf{Y}_{(1)}$, but also new predictor values. Individual-level risk at a future time t can be derived as $P = \Pr(T \leq t | \mathbf{Y}_{(1)})$, based on the conventional model, and $Q = \Pr(T \leq t | \mathbf{Y}_{(2)})$, the corresponding risk based on the new model, respectively. Since, in practice, risk categories are often uncertain for many diseases, a more objective and flexible measure of improvement in risk prediction would be based on P or Q in their original continuous scales. Therefore, following the definition of Pencina and D'Agostino (2011), in this manuscript we focus on the time-dependent continuous NRI, which is a more general definition that does not rely on the existence of risk categories. In the time-dependent setting, we further denote an 'event' person at time t as those with $T \leq t$, and a 'nonevent' person as $T > t$. Here, $\text{NRI}(t)$ is equal to the sum of 'event NRI' and 'nonevent NRI', which are defined as:

$$\begin{aligned} \text{event NRI}_u(t) &= \Pr(Q - P > u | T \leq t) - \Pr(Q - P \leq u | T \leq t) \\ &\equiv 2\Pr(Q - P > u | T \leq t) - 1, \end{aligned}$$

and

$$\begin{aligned} \text{nonevent NRI}_v(t) &= \Pr(Q - P \leq v | T > t) - \Pr(Q - P > v | T > t) \\ &\equiv 1 - 2\Pr(Q - P > v | T > t). \end{aligned}$$

Since, $\text{NRI}_{u,v}(t) = \text{event NRI}_u(t) + \text{nonevent NRI}_v(t)$, it follows that $\text{NRI}_{u,v}(t) = 2\{\Pr(Q - P > u | T \leq t) - \Pr(Q - P > v | T > t)\}$. In practice we may choose u and v such that improvement in risk estimates is meaningful (Uno et al. 2009). Setting $u = v = 0$ gives the 'continuous NRI' considered in Pencina and D'Agostino (2011). For the ease of presentation, in the sequel, we'll omit the subscript u and v from our notations and assume $u = v = 0$, but note that our estimators can be constructed for any arbitrary u and v . In the next section, we show how each component of $\text{NRI}(t)$ can be estimated.

3 Estimation

Suppose we have a cohort of N individuals from the targeted population followed prospectively. Due to censoring, the observed data consist of N i.i.d copies of vector, $\mathcal{D} = \{\mathbf{D}_i = (X_i, \delta_i, \mathbf{Y}_i)^\top, i=1, \dots, N\}$, where $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$ for T_i and C_i denote failure time and censoring time respectively. \mathbf{Y}_i are predictors from individual i measured at time 0, including subset $\mathbf{Y}_{i(1)}$ used in the existing model (model 1) and $\mathbf{Y}_{i(2)}$ in the new model (model 2) such that $\mathbf{Y}_{i(1)} \in \mathbf{Y}_{i(2)}$. For illustration, we first assume that P and Q both follow the conventional Cox regression models. Specifically, at time t , we assume $P(\theta_1) = 1 - \exp[-\Lambda_{01}(t) \exp\{\beta_1^\top \mathbf{Y}_{(1)}\}]$ and $Q(\theta_2) = 1 - \exp[-\Lambda_{02}(t) \exp\{\beta_2^\top \mathbf{Y}_{(2)}\}]$, where Λ_{0k} is the baseline cumulative hazard function, β_k are unknown vector of parameters, for model $k = 1, 2$, and $\theta_1 = (\Lambda_{01}(t), \beta_1^\top)^\top$, $\theta_2 = (\Lambda_{02}(t), \beta_2^\top)^\top$. It is important to note that these models are most likely not correctly specified. Nevertheless under a mild regularity condition, the standard maximum partial likelihood estimator $\widehat{\beta}_k$ for β_k converges to a constant vector, as $n \rightarrow \infty$ (Hjort 1992). This provides theoretical ground for our asymptotic studies.

To estimate $\text{NRI}(t)$, Pencina and D'Agostino (2011) first expressed the two key components as

$$\Pr\{B(\theta) > 0 | T \leq t\} = \frac{\Pr\{T \leq t | B(\theta) > 0\} \Pr\{B(\theta) > 0\}}{\Pr\{T \leq t\}}$$

and

$$\Pr\{B(\theta) > 0 | T > t\} = \frac{\Pr\{T > t | B(\theta) > 0\} \Pr\{B(\theta) > 0\}}{\Pr\{T > t\}},$$

where $B(\theta) = Q(\theta_2) - P(\theta_1)$ and $\theta = (\theta_1, \theta_2)^\top$. To account for censoring, Pencina and D'Agostino (2011) proposed to use the KM estimator to estimate the survival function using data from all subjects for $\Pr\{T \leq t\}$ and using subjects with $B(\theta) > 0$ for estimation of $\Pr\{T \leq t | B(\theta) > 0\}$. We refer to the resulting estimator as the 'KM estimator' hereafter.

Uno et al. (2009) considered estimating $\text{NRI}(t)$ based on an inverse-probability-of-censoring weighted (IPW) estimator (hereafter referred to as the 'IPW estimator'), with its key components estimated as

$$\widehat{\Pr}^{\text{IPW}}\{B(\theta) > 0 | T \leq t\} = \frac{\sum_i I\{B_i(\widehat{\theta}) > 0, X_i \leq t\} \widehat{W}_i(t)}{\sum_i I(X_i \leq t) \widehat{W}_i(t)} \quad (3.1)$$

$$\widehat{\Pr}^{\text{IPW}}\{B(\theta) > 0 | T > t\} = \frac{\sum_i I\{B_i(\widehat{\theta}) > 0, X_i > t\} \widehat{W}_i(t)}{\sum_i I(X_i > t) \widehat{W}_i(t)} \quad (3.2)$$

where $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2)^\top$, $\widehat{\theta}_1 = (\widehat{\Lambda}_{01}(t), \widehat{\beta}_1^\top)^\top$, $\widehat{\theta}_2 = (\widehat{\Lambda}_{02}(t), \widehat{\beta}_2^\top)^\top$, $\widehat{W}_i(t) = I(X_i \leq t) \delta_i / \widehat{H}(X_i) + I(X_i > t) / \widehat{H}(t)$ and $\widehat{H}(\cdot)$ is the KM estimator of $H(\cdot) = P(C > \cdot)$. Due to the equivalence between the KM estimator and the IPW estimator for marginal survival functions under independent censoring (Satten and Datta 2001), the two estimators are likely to have very similar robustness and efficiency. Both estimators are consistent

under an independent censoring assumption regardless of the adequacy of the two fitted models, $P(\theta_1)$ and $Q(\theta_2)$. This is particularly appealing for model comparisons.

One potential weakness of both estimators is that they can be biased if censoring is dependent on a subset of $Y_{(2)}$. On the other hand, when model 2 is correctly specified, such covariate-dependent censoring can be incorporated based on the model since $C \perp T$ given $\beta_2^T Y_{(2)}$ or $Q(\theta_2)$. This motivates us to propose a more robust alternative to the Uno et al. (2009) estimator by estimating censoring probabilities given $Y_{(2)}$ via kernel smoothing over $Q(\theta_2)$. Let $H_q^1(t) = P(C > t \mid Q(\theta_2) = q, \Delta_i(\theta) = 1)$ and $H_q^\bullet(t) = P(C > t \mid Q(\theta_2) = q)$ where $\Delta_i(\theta) = I\{B_i(\theta) > 0\}$. To estimate $NRI(t)$, we propose to modify equations (3.1) and (3.2) by considering the following more robust IPW censoring weights

$$\tilde{W}_i^{(\iota)}(t) = \frac{I(X_i \leq t) \delta_i}{\widehat{H}_{Q_i(\hat{\theta}_2)}^{(\iota)}(X_i)} + \frac{I(X_i > t)}{\widehat{H}_{Q_i(\hat{\theta}_2)}^{(\iota)}(t)} \quad \text{for } \iota=1 \text{ and } \bullet,$$

where $\widehat{H}_q^{(\iota)}(t) = \exp\{-\widehat{\Lambda}_q^{(\iota)}(t)\} = \exp\left\{-\int_0^t \widehat{\pi}_q^{(\iota)}(s)^{-1} d\widehat{N}_{C_q}^{(\iota)}(s)\right\}$,

$$\begin{aligned} \widehat{N}_{C_q}^{(\iota)}(s) &= n^{-1} \sum_{i: \Delta_i(\hat{\theta}) \in \mathcal{U}_\bullet} K_h\{Q_i(\hat{\theta}_2) - q\} N_{C_i}(s), \\ \widehat{\pi}_q^{(\iota)}(s) &= n^{-1} \sum_{i: \Delta_i(\hat{\theta}) \in \mathcal{U}_\bullet} K_h\{Q_i(\hat{\theta}_2) - q\} I(X_i \geq s), \end{aligned}$$

$N_{C_i}(s) = I(X_i \leq s)(1 - \delta_i)$, $\mathcal{U}_1 = 1$ and $\mathcal{U}_\bullet = \{0, 1\}$, $K_h(\cdot) = \frac{1}{h} K\left\{\frac{\cdot - Q_i(\hat{\theta}_2)}{h}\right\}$, K is a symmetric kernel density function, with $h = h(n) \rightarrow 0$ as the bandwidth. Note that $\Delta_i(\hat{\theta}) \in \mathcal{U}_1$ is simply the subset of individuals with $B_i(\hat{\theta}) > 0$ and $\Delta_i(\hat{\theta}) \in \mathcal{U}_\bullet$ is the set of *all* individuals. Consequently we can then use these more robust kernel smoothing weights in the IPW estimator, to obtain the ‘Smooth-IPW (S-IPW) estimators’,

$$\widehat{\Pr}^S - \text{IPW}\{B(\theta) > 0 | T \leq t\} = \frac{\sum_i \Delta_i(\hat{\theta}) \tilde{W}_i^{(1)}(t) I(X_i \leq t)}{\sum_i \tilde{W}_i^{(\bullet)}(t) I(X_i \leq t)} \quad \text{and} \quad (3.3)$$

$$\widehat{\Pr}^S - \text{IPW}\{B(\theta) > 0 | T > t\} = \frac{\sum_i \Delta_i(\hat{\theta}) \tilde{W}_i^{(1)}(t) I(X_i > t)}{\sum_i \tilde{W}_i^{(\bullet)}(t) I(X_i > t)}. \quad (3.4)$$

This resulting estimator for $NRI(t)$ is

$$\widehat{NRI}(\hat{\theta}, t) = 2 \times \left[\widehat{\Pr}^S - \text{IPW}\{B(\hat{\theta}) > 0 | T \leq t\} - \widehat{\Pr}^S - \text{IPW}\{B(\hat{\theta}) > 0 | T > t\} \right].$$

The estimator can be shown to have the property of ‘double robustness’, i.e., it only requires that the risk model Q is correctly specified or that the independent censoring assumption holds.

Additionally, to improve upon the efficiency of the class of nonparametric estimators, we propose considering a semiparametric estimator. Note that

$$\Pr \{B(\theta) > 0 | T > t\} = \frac{E[E\{I(B(\theta) > 0, T > t) | Y_{(2)}\}]}{E[E\{I(T > t) | Y_{(2)}\}]} = \frac{E\{I(B(\theta) > 0)P(T > t | Y_{(2)})\}}{E\{P(T > t | Y_{(2)})\}}.$$

Therefore $NRI(t)$ can be estimated semiparametrically as

$$\widehat{NRI}(\widehat{\theta}, t) = 2 \times \left\{ \widehat{\Pr}^{SEM}(B(\theta) > 0 | T \leq t) - \widehat{\Pr}^{SEM}(B(\theta) \leq 0 | T > t) \right\},$$

with the ‘SEM’ estimators,

$$\widehat{\Pr}^{SEM}(B(\theta) > 0 | T \leq t) = \frac{\sum_{i=1}^n \Delta_i(\widehat{\theta}) Q_i(\widehat{\theta}_2)}{\sum_{i=1}^n Q_i(\widehat{\theta}_2)}, \quad (3.5)$$

$$\widehat{\Pr}^{SEM}(B(\theta) > 0 | T > t) = \frac{\sum_{i=1}^n \Delta_i(\widehat{\theta}) \{1 - Q_i(\widehat{\theta}_2)\}}{\sum_{i=1}^n \{1 - Q_i(\widehat{\theta}_2)\}}. \quad (3.6)$$

Under the correctly specified model $Q(\theta_2)$, the semiparametric estimator accommodates a covariate-dependent censoring situation and would be more efficient compared to the Smooth-IPW estimator. In practice, to estimate $NRI(t)$, if estimates from such a semiparametric method agree well with those of the nonparametric methods, one may choose to report results based on the semiparametric method for additional gain in efficiency. To automatize the procedure, we suggest considering a combined estimator (hereafter referred as the ‘combined estimator’), which takes the form

$$\widehat{p} \times \widehat{NRI}(\widehat{\theta}, t) + (1 - \widehat{p}) \times \widetilde{NRI}(\widehat{\theta}, t),$$

with \widehat{p} being a weight that is dependent on the aptness of the semiparametric model. For example, \widehat{p} can be taken to be the p -value from a consistent test of the proportional hazards assumption for a Cox regression model fit. Such an estimator provides a simple procedure which is robust over a wide variety of situations. In numerical studies, we show that such a combined estimator can be more efficient compared with the nonparametric estimators, while maintaining the double robustness property.

We note that the proposed estimators can be easily generalized to NRI based on risk categories. Consider a situation where individuals are classified as low, intermediate or high risk: low risk if their risks are below r_1 , and high risk if their risks are above r_2 . The reclassification accuracy of risk models in such a setting can be quantified with a 3-category NRI of the form $NRI^{category}(\widehat{\theta}, t) = P(up|T \leq t) - P(down|T \leq t) + P(down|T > t) - P(up|T > t)$. To estimate $P(up|T \leq t)$ and $P(up|T > t)$, we may simply replace $\Delta_i(\widehat{\theta})$ with $\Omega_i^{up}(\widehat{\theta}) = I(P_i(\theta_1) \leq r_1, Q_i(\theta_2) > r_1) + I(r_1 < P_i(\theta_1) \leq r_2, Q_i(\theta_2) > r_2)$ in Eqs. 3.3 and 3.4, respectively. Similarly, to estimate $P(down|T \leq t)$ and $P(down|T > t)$, one may replace $\Delta_i(\widehat{\theta})$ with $\Omega_i^{down}(\widehat{\theta}) = I(Q_i(\theta_1) \leq r_1, P_i(\theta_2) > r_1) + I(r_1 < Q_i(\theta_1) \leq r_2, P_i(\theta_2) > r_2)$ in Eqs. 3.3 and 3.4.

Similarly, one may obtain a semiparametric estimator of $\text{NRI}^{\text{category}}(\widehat{\theta}, t)$ by replacing $\Delta_i(\widehat{\theta})$ with $\Omega_i^{\text{up}}(\widehat{\theta})$, or $\Omega_i^{\text{down}}(\widehat{\theta})$ in Eqs. 3.5 and 3.6.

4 Inference

To make inference about $\widehat{\text{NRI}}(\widehat{\theta}, t)$, we study the asymptotic properties of proposed estimators. In the Appendix, we show that $\widehat{\text{NRI}}(\widehat{\theta}, t)$ is uniformly consistent for $\text{NRI}(\theta_0, t)$, where $\theta_0 = (\Lambda_{k0}(\cdot), \beta_{k0}^T)^T$ with β_{k0} being the unique maximizer of the expected value of the corresponding partial likelihood. Furthermore, we show that the process

$\widetilde{\mathcal{W}}(t) = \sqrt{n} \{ \widehat{\text{NRI}}(\widehat{\theta}, t) - \text{NRI}(\theta_0, t) \}$ is asymptotically equivalent to a sum of i.i.d terms,

$n^{-1/2} \sum_{i=1}^n \epsilon_i(t)$ where $\epsilon_i(t)$ is defined in the the Appendix. By a functional central limit theorem of Pollard (1990), the process $\widetilde{\mathcal{W}}(t)$ converges weakly to a mean zero Gaussian process in t . We also show that $\widehat{\text{NRI}}(\widehat{\theta}, t)$ is uniformly consistent for $\text{NRI}(\theta_0, t)$, and that the process

$\widetilde{\mathcal{N}}(t) = \sqrt{n} [\widehat{\text{NRI}}(\widehat{\theta}, t) - \text{NRI}(\theta_0, t)]$ is asymptotically equivalent to a sum of i.i.d terms $n^{-1/2} \sum_{i=1}^n \zeta_i(t)$ where $\zeta_i(t)$ is defined in the Appendix. Again, by a functional central limit theorem, the process $\widetilde{\mathcal{N}}(t)$ converges weakly to a mean zero Gaussian process in t .

With weak convergence of both $\widehat{\text{NRI}}(\widehat{\theta}, t)$ and $\widetilde{\mathcal{N}}(t)$, it follows that the combined estimator converges to a zero-mean process. Due to the variation in \widehat{p} , the combined estimators may not be a Gaussian process. We show in our simulation that to make inference, resampling procedures such as a bootstrap method can provide a valid approximation of the limit distribution. Specifically, at each of the b th bootstrap iterations, with $b = 1, \dots, B$, we conduct a random sampling with replacement of the original dataset, and fit our new and old risk models based on the sampled dataset, denoted as $P^b(\widehat{\theta})$ and $Q^b(\widehat{\theta})$. These estimates from the fitted models are then used to calculate $\widehat{\text{NRI}}^b(\widehat{\theta}, t)$ and $\widetilde{\text{NRI}}(\widehat{\theta}, t)$ based on the bootstrapped samples. This procedure will be repeated B times, and confidence intervals can be constructed either based on the percentile method, or a normal approximation where the standard error is calculated based on the empirical standard errors of $\{ \widehat{\text{NRI}}^b(\widehat{\theta}, t), b=1, \dots, B \}$ and $\{ \widetilde{\text{NRI}}^b(\widehat{\theta}, t), b=1, \dots, B \}$. The combined estimator can be inferred similarly by repeatedly calculate the weights based on each bootstrap sample in addition to $\widehat{\text{NRI}}^b(\widehat{\theta}, t)$ and $\widetilde{\text{NRI}}(\widehat{\theta}, t)$.

In the absence of an independent validating set, often in practice the same dataset is used for both fitting the model with several predictors and calculating a measure such as $\text{NRI}(t)$. Such an ‘apparent’ summary may potentially lead to the so-called ‘overfitting’ phenomenon, i.e. estimates of model performance will tend to be more optimistic compared with the corresponding estimates if the model were to applied to a new dataset. Several methods for correcting the bias from apparent estimates can be considered. The 0.632 Bootstrap method (Efron and Tibshirani 1997) has been shown to have better performance compared with a simple cross-validated approach. The estimator was derived in our simulation as follows: we first obtained a bootstrapped estimate $\widehat{\text{NRI}}^{bt}(t)$ by sampling the data with replacement to obtain the training set. The training set is used to estimate the model parameters $\{ \widehat{\theta}_k^{(\text{train})}, k=1, 2 \}$. The remaining subjects make up the validation set, and are used to calculate

the various estimates of NRI using parameter values $\{\widehat{\theta}_k^{(\text{train})}, k=1, 2\}$. This is repeated B times and $\widehat{\text{NRI}}^{bt}(t)$ is the mean across the repetitions. The 0.632 bootstrap estimate is,

$$\widehat{\text{NRI}}^{0.632bt}(t) = 0.632\widehat{\text{NRI}}^{bt}(t) + (1 - 0.632)\widehat{\text{NRI}}^{\text{apparent}}(t),$$

where $\widehat{\text{NRI}}^{\text{apparent}}(t)$ is the estimate without using cross-validation. To construct a confidence interval based on $\widehat{\text{NRI}}^{0.632bt}(t)$, we follow the suggestions given in Tian et al. (2007) and Uno et al. (2007) by shifting the apparent error based confidence interval in the amount of bias estimated as $\widehat{\text{bias}} = \widehat{\text{NRI}}^{\text{apparent}}(t) - \widehat{\text{NRI}}^{0.632bt}(t)$. Specifically, if $[L, R]$ is the confidence interval calculated based on the procedure described above, then the bias corrected confidence interval is $[L - \widehat{\text{bias}}, R - \widehat{\text{bias}}]$.

5 Simulation studies

To examine the performance of various $\text{NRI}(t)$ estimators, we conducted simulation studies under several different scenarios. Throughout we chose $n = 500$ and used 200 bootstrap samples to calculate standard errors. Results for each setting were produced from 1,000 simulations. We calculated $\text{NRI}(t)$, for $t = 3$ for comparing two risk models using the KM, IPW, Smooth-IPW, SEM and the combined estimators described in Sect. 3. We fitted Cox regression models to calculate risks for both the new and existing models using corresponding predictors.

For the first setting presented in Table 1, two predictors Y_1 and Y_2 were simulated from a multivariate normal distribution with mean $(0, 0.5)$, $\sigma_{y1} = \sigma_{y2} = 1$ and a correlation ρ of 0.25. The relationship between survival time T and \mathbf{Y} followed a proportional hazards model with parameters $\beta_1 = \log(3)$ and β_2 equal to $\log(1.5)$. Censoring time was generated from a $U(0, a)$ distribution where a was chosen to produce approximately 40% censoring. Note that in this setting, model Q is correctly specified and the independent censoring assumption is correct. We took the baseline model to consist of Y_1 and the new model to include both predictors. As expected, all estimators shown in Table 1 provide unbiased estimates. The bootstrap-based variance estimators perform well with coverage percentage close to the 95% nominal level. Since the risk based on the new model is correctly specified, the semiparametric method is the most efficient. Improvement in efficiency over the nonparametric procedures is observed with our combined estimators.

Under this setting we also considered a null model where $\beta_2 = 0$ i.e. there is no incremental value of the new marker and $\text{NRI}(t) = 0$. We found that in this situation all estimators tend to slightly over estimate $\text{NRI}(t)$, and variance estimators based on the bootstrap estimators tend to be conservative (see Table 2). We do not recommend calculating $\text{NRI}(t)$ in the case when the new marker does not independently predict outcome in a model with conventional predictors. Note that all theoretical results in the Appendix are derived under the assumption that $\beta_2 \neq 0$ and thus our proposed procedures are only valid under this assumption. In practice, if the null setting is a likely possibility, estimation should be treated with care.

The second setting we considered was identical to the first setting, except that censoring time was dependent on marker values. Here, censoring time,

$$C = U \cdot B + \exp(X - 3Y_2) \cdot (1 - B),$$

where U was generated from a Uniform(0, a) distribution where with a chosen to yield about 40% censoring, X was generated from a $N(0, 1)$ distribution and B was generated from a $N(2 \cdot Y_1, 1)$ distribution. Note that in this setting, model Q is correctly specified but the independent censoring assumption is not correct. As seen in the results presented in Table 3, the KM estimator yields biased estimators for both $NRI(t)$ and its two key components. The IPW estimator is biased for both $\Pr(P > Q | T \leq t)$ and $NRI(t)$, whereas the smooth-IPW estimator substantially alleviates such biases. However, we observed large variation in nonparametric estimators of $NRI(t)$ as compared with the semiparametric and combined estimators (Table 3).

The third setting we investigated considers the case where survival time depends on four markers Y_i , for $i = 1, \dots, 4$, but we only have access to the first two. In particular, \mathbf{Y} comes from a multivariate normal distribution with mean 0, and $\sigma_{ij} = 1$ for $i = j$ and 0.25 otherwise. Survival time relates to the marker data through a model where the hazard function is specified as $\lambda(t|\mathbf{Y}) = 0.1 * \{3 Y_1 + 1.5 Y_2 + 2 Y_3 + 2.5 Y_4 + \exp(3 Y_1)\}$. Note that in this setting, model Q is misspecified as depending only on Y_1 and Y_2 . Censoring time in this setting is generated the same as in the first setting, which does not depend on T or \mathbf{Y} . Since the SEM estimator misspecified the relationship between T and \mathbf{Y} as $\lambda(t|\mathbf{Y}) = \lambda_0 \exp(\beta_1 Y_1 + \beta_2 Y_2)$, it yields biased results. All other estimators are unbiased (Table 4). Throughout the three settings we considered, the combined estimator remained unbiased and more efficient than other nonparametric estimators.

To evaluate the procedures described above, we simulated 10 markers from a multivariate normal distribution with mean $\mathbf{0}$, $\sigma_{Y_i} = 1$ and pairwise correlations equal to 0.25. The number of parameters and sample size were chosen to mimic the setting of our data example described in Sect. 6. We consider a Cox model for failure time, with hazard ratio parameters for 10 markers specified as $\beta = (\log(2), \log(.77), 0, \log(1.81), 0, 0, 0, \log(0.5), 0, \log(1.2))$. The baseline model consists only of the first marker. To derive a new model based on the information on all 10 markers, for each simulation, we first fit a model with all ten markers. The expanded model consists of all markers that have non-zero β at an $\alpha = 0.05$ level. We found that in the case of estimating NRI, under our simulated scenario, the apparent summaries are quite close to the true values in many cases. Since the bias is at the rate of g/N , where g is the number of predictors under consideration for risk model building, overfitting may be of more concern when large numbers of genetic markers are involved with a relatively small sample size. In the situation there is a slight indication of overfitting, the 0.632 bootstrap procedure appears to be adequate in correcting the bias (see Table 5).

6 Example

The Framingham risk model (FRM) has been used for population-wide CVD risk assessment. The model was developed based on several common clinical risk factors, including age, gender, total cholesterol level, high-density lipoprotein (HDL) cholesterol level, smoking, systolic blood pressure and high blood pressure treatment (Wilson et al. 1998). To improve the predictive capacity of the FRM, a new risk model has been developed recently using data from the Women's Health Study (Cook et al. 2006), based on variables in the Framingham risk model and an inflammation marker, C-reactive protein (CRP). Prior to adapting the new model in routine practice, it is important to quantify its prediction performance, especially in comparison to that of FRM. We illustrate here how our proposed procedures can be used to evaluate and compare the clinical utility of the two risk models using an independent dataset from the Framingham Offspring Study (Kannel et al. 1979).

The Framingham Offspring Study was established in 1971 with 5,124 participants who were monitored prospectively for epidemiological and genetic risk factors for CVD. We consider

here 1,728 female participants who have CRP measurement and other clinical information at the second exam and are free of CVD at the time of examination. The average age of this subset was about 44 years (standard deviation = 10). The outcome we consider is the time from exam date to first major CVD event including CVD-related death. During the followup period 269 participants were observed to encounter at least one CVD event and the 10-year event rate was about 4%. For illustration we chose $t = 10$ years as in Wilson et al. (1998). For each individual, two risk scores were calculated: one based on the FRM (Model 1), combining information on age, systolic blood pressure, smoking status, high-density lipoprotein (HDL), total cholesterol, medication for hypertension; the other based on an algorithm developed in Cook et al. (2006) (Model 2), with the addition of CRP concentration. We use Cox models to specify the relation between the time-to-CVD events and model scores (linear predictors from the models).

Both models are well calibrated based on calibration plots (not shown). For comparison, we first give AUC results and use the bootstrap to obtain confidence intervals. The AUC for an ROC curve at 10-years is 0.752 (95% CI: 0.721,0.783) for Model 1 and 0.758 (95% CI: 0.729,0.787) for Model 2. The difference between the two AUCs is not statistically significant: 0.006 (95% CI: -0.033, 0.046). We now investigate whether the new models reclassify patients in terms of their risks and CVD outcome at 10 years. We consider *NRI* (10-years) for such an evaluation using the methods described in Sect. 3. Table 6 shows that estimates from the three nonparametric models are quite consistent, all indicating that the new model does not add significant improvement gauged by NRI. The semiparametric model, however, does indicate a significant incremental value with $NRI = 0.167$ ($SE = 0.067$), and the combined estimator indicates a similar magnitude of improvement, though not significant ($NRI = 0.132$, $SE = 0.137$). Note that since we considered a continuous NRI with $u = v = 0$, the observed improvement at this magnitude may not be interpreted as clinically substantial. Since different conclusions could be reached depending on which estimation method is chosen, this analysis highlights the need to consider multiple robust approaches for calculating NRI.

7 Discussion

NRI provides an alternative tool for evaluating risk prediction models (Pencina et al. 2008) beyond the traditional ROC curve framework. The concept has continued to gain popularity in the medical literature, yet its statistical properties have not been well studied to date in the statistical literature, and existing methods for calculating NRI under the failure time outcome setting are limited. In this manuscript, we provide a more thorough investigation of a variety of estimation procedures. Our proposed nonparametric and semiparametric estimators improve upon existing methods both in terms of robustness and efficiency under a variety of practical situations. Such improvement is quite important, since we observe that compared with other measures such as AUC, NRI estimates, in general, are not very stable with substantial variations in the estimators we have considered. The proposed procedures can be used for estimating both continuous NRI and NRI with pre-specified fixed categories. As illustrated in the example, the choice of estimation method can lead to different conclusions. In practice, the method chosen should depend on a number of important considerations including the likelihood that the model has been correctly specified and that the assumptions concerning censoring are correct. In addition, in situations where the new marker may be expensive or difficult to ascertain, an approach which considers both the risks and benefits of obtaining the marker should be considered in a decision-making process. We recommend such measures to be used in practice with caution. A thorough evaluation of a risk model should consider a wide spectrum of measures for assessing discrimination and calibration, and NRI may be better served as one of the summary measures to complement graphical displays of risk distributions (Gu and Pepe 2009). All

analyses were performed in R. Code for implementing the proposed procedures is available upon request.

Acknowledgments

The Framingham Heart Study and the Framingham SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. The Framingham SHARe data used for the analyses described in this manuscript were obtained through dbGaP (access number: phs000007.v3.p2). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI. The work is supported by grants U01-CA86368, P01-CA053996, R01- GM085047, R01-GM079330 awarded by the National Institutes of Health.

Appendix

Throughout, we assume that the joint density of (T, C, \mathbf{Y}) is twice continuously differentiable, \mathbf{Y} are bounded, and $1 > P(T > t) > 0, 1 > P(C > t) > 0$. The kernel function K is a symmetric probability density function with compact support and bounded second derivative. The bandwidth $h \rightarrow 0$ such that $nh^4 \rightarrow 0$. In addition, the estimator $\widehat{\theta}_k$ converges to θ_{0k} for $k = 1, 2$ as $n \rightarrow \infty$ (Hjort 1992), where β_{k0} is the unique maximizer of the expected value of the corresponding partial likelihood and Λ_{k0} is the baseline cumulative hazard for $k = 1, 2$. We denote the parameter space for θ_k by Ω_k and assume that Ω_k is a compact set containing θ_{0k} . Furthermore, we assume that $\beta_2 = 0$ and note that $Q(\theta_2) = 1 - \exp\{\Lambda_{02}(t) e^{\beta_2^T Y_{(2)}}\}$ and $P(\theta_1) = 1 - \exp\{\Lambda_{01}(t) e^{\beta_1^T Y_{(1)}}\}$ are the respective limits of $Q(\widehat{\theta}_2)$ and $P(\widehat{\theta}_1)$, for any given $\mathbf{Y}_{(2)}$ and $\mathbf{Y}_{(1)}$. The in-probability convergence of $Q(\widehat{\theta}_2) \rightarrow Q(\theta_{02})$ and $P(\widehat{\theta}_1) \rightarrow P(\theta_{01})$ are uniform in $\mathbf{Y}_{(2)}$ and $\mathbf{Y}_{(1)}$ due to the convergence of $\widehat{\theta} \rightarrow \theta_0 = (\theta_{01}^T, \theta_{02}^T)^T$.

Asymptotic Properties of $\widehat{\text{NRI}}(\widehat{\theta}, t)$

From the same arguments as given in Cai et al. (2010) and Dabrowska (1997), it follows that we have the uniform consistency of $\widetilde{H}_q^{(\iota)}(t)$ to $\widetilde{H}_q^{(\iota)}(t) = P(C \geq t \mid Q(\theta_2) = q, \Delta(\theta) \in \mathcal{U}_\bullet)$, where $\mathcal{U}_1 = 1$ and $\mathcal{U}_\bullet = \{0, 1\}$, for $\iota = 1$ and \bullet . It follows, using the law of numbers (Pollard 1990), that

$$\sup_{\theta} |\widehat{\text{NRI}}(\theta, t) - \text{NRI}(\theta, t)| \rightarrow 0.$$

This along with the convergence of $\widehat{\theta}$ to θ_0 implies that $\widehat{\text{NRI}}(\widehat{\theta}, t)$ is uniformly consistent for $\text{NRI}(\theta_0, t)$.

Throughout, we will use the fact that

$$E \left\{ \Delta_i(\theta) I(X_i \leq t) \delta_i H_{Q_i(\theta_2)}^{(1)}(X_i)^{-1} \mid Q_i(\theta_2) = q \right\} = P(\Delta_i(\theta) = 1, T_i \leq t \mid Q_i(\theta_2) = q)$$

if either $C \perp T, \mathbf{Y}_{(2)}$ (model may be misspecified) or $Q(\theta_2) = \Pr(T \leq t | Y_{(2)})$ i.e. the Cox model is correctly specified though censoring may be such that $C \perp T | \mathbf{Y}_{(2)}$ (double robustness). We

first write the i.i.d representation of $\sqrt{n} [\widehat{\text{NRI}}(\theta, t) - \text{NRI}(\theta, t)]$ for any θ . Note that

$$\sqrt{n} [\widehat{\text{NRI}}(\theta, t) - \text{NRI}(\theta, t)] = 2 \sqrt{n} \left\{ \widetilde{\Pr}(\Delta(\theta) = 1 | T \leq t) - \Pr(\Delta(\theta) = 1 | T \leq t) \right\} - 2 \sqrt{n} \left\{ \widetilde{\Pr}(\Delta(\theta) = 1 | T > t) - \Pr(\Delta(\theta) = 1 | T > t) \right\}.$$

We first examine the initial component,

$$\widehat{\Pr}(\Delta(\widehat{\theta})=1|T \leq t) = \frac{\sum_i \Delta(\widehat{\theta}) I(X_i \leq t) \delta_i / \widehat{H}_{Q_i(\widehat{\theta}_2)}^{(1)}(X_i)}{\sum_i I(X_i \leq t) \delta_i / \widehat{H}_{Q_i(\widehat{\theta}_2)}^{(\bullet)}(X_i)} \equiv \frac{\widehat{N}(t, \widehat{\theta}, \widehat{H})}{\widehat{D}(t, \widehat{\theta}, \widehat{H})}$$

where $\widehat{N}(t, \theta, H) = n^{-1} \sum_i \Delta_i(\theta) I(X_i \leq t) \delta_i / H_{Q_i(\theta_2)}^{(1)}(X_i)$ and $\widehat{D}(t, \theta, H) = n^{-1} \sum_i I(X_i \leq t) \delta_i / H_{Q_i(\theta_2)}^{(\bullet)}(X_i)$. Let $N(t, \theta) = \Pr(\Delta(\theta) = 1, T \leq t)$ and $D(t) = \Pr(T \leq t)$. Then by the uniform consistency of the IPW weights, we have

$$\begin{aligned} \sqrt{n} \{ \widehat{\Pr}(\Delta(\theta) = 1|T \leq t) - \Pr(\Delta(\theta) = 1|T \leq t) \} \\ \approx \sqrt{n} \{ \widehat{N}(t, \theta, \widehat{H}) D(t) - N(t, \theta) \widehat{D}(t, \theta, \widehat{H}) \} / D(t)^2. \end{aligned}$$

Examining the numerator, $\sqrt{n} \{ \widehat{N}(t, \theta, \widehat{H}) D(t) - N(t, \theta) \widehat{D}(t, \theta, \widehat{H}) \} = \sqrt{n} \{ (1) + (2) - (3) \}$ where (1) = $\widehat{N}(t, \theta, H) D(t) - \widehat{D}(t, \theta, H) N(t, \theta)$, (2) = $\widehat{N}(t, \theta, \widehat{H}) D(t) - \widehat{N}(t, \theta, H) D(t)$, and (3) = $[N(t, \theta) \widehat{D}(t, \theta, \widehat{H}) - \widehat{D}(t, \theta, H) N(t, \theta)]$. Note that

$$\begin{aligned} (1) &= \sqrt{n} \{ \widehat{N}(t, \theta, H) D(t) - \widehat{D}(t, \theta, H) N(t, \theta) \} = n^{-\frac{1}{2}} \sum U_{1i}(t), \text{ where} \\ U_{1i}(t) &= \frac{I(X_i \leq t) \delta_i}{H_{Q_i(\theta_2)}^{(1)}(X_i)} \Delta_i(\theta) D(t) - \frac{I(X_i \leq t) \delta_i}{H_{Q_i(\theta_2)}^{(\bullet)}(X_i)} N(t, \theta) \end{aligned}$$

Using a Taylor series expansion, Lemma A.3 of Biliias et al. (1997) and the asymptotic expansion for $\widehat{\Lambda}_q(t)$ given in Du and Akritas (2002),

$$\begin{aligned} (2) &= D(t) \sqrt{n} \{ \widehat{N}(t, \theta, \widehat{H}) - \widehat{N}(t, \theta, H) \} \\ &= D(t) n^{-1/2} \sum_i \frac{\Delta_i(\theta) I(X_i \leq t) \delta_i}{H_{Q_i(\theta_2)}^{(1)}(X_i)} \left[\frac{H_{Q_i(\theta_2)}^{(1)}(X_i)}{\widehat{H}_{Q_i(\theta_2)}^{(1)}(X_i)} - 1 \right] \\ &= D(t) n^{-1/2} \int \int_0^t \left[\frac{H_q^{(1)}(s)}{\widehat{H}_q^{(1)}(s)} - 1 \right] d \sum_i \frac{\Delta_i(\theta) \delta_i I(X_i \leq s, Q_i(\theta_2) \leq q)}{H_{Q_i(\theta_2)}^{(1)}(X_i)} \\ &\approx D(t) \int \int_0^t \sqrt{n} \left[\widehat{\Lambda}_q^{(1)}(s) - \Lambda_q^{(1)}(s) \right] d \left\{ \frac{1}{n} \sum_i \frac{\Delta_i(\theta) \delta_i I(X_i \leq s, Q_i(\theta_2) \leq q)}{H_{Q_i(\theta_2)}^{(1)}(X_i)} \right\} \\ &\approx D(t) \int \int_0^t \left[n^{-\frac{1}{2}} \sum K_h \{ q - Q_i(\theta_2) \} M_{C_q}^{(1)}(s, X_i, \delta_i) \right] dP(\Delta(\theta) = 1, T \leq t, Q(\theta_2) \leq q) \end{aligned}$$

where

$$M_{C_q}^{(1)}(t, X_i, \delta_i) = \int_0^t \frac{dN_{C_i}(s) - I(X_i \geq s) d\Lambda_q^{(1)}(s)}{\pi_s^{(1)}(q)}.$$

Now by a change of variable, $\psi = \frac{q - Q_i(\theta_2)}{h}$ and $f(t, q) \equiv \partial^2 P(\Delta(\theta) = 1, T \leq t, Q(\theta_2) \leq q) / \partial t \partial q$,

$$\begin{aligned} (2) &\approx D(t) \int \int_0^t \sqrt{n} \left[\frac{1}{n} \sum K(\psi) M_{C(\psi h + Q_i(\theta_2))}^{(1)}(s, X_i, \delta_i) \right] f(t, \psi h + Q_i) ds d\psi \\ &= D(t) n^{-1/2} \sum \int \int_0^t K(\psi) a \{ s, h\psi + Q_i(\theta_2), X_i \} ds d\psi = n^{-\frac{1}{2}} \sum U_{2i}(t), \end{aligned}$$

where $U_{2i}(t) = D(t) \int_0^t a(s, q^*, X_i) ds$ and $a(t, q, X_i) = M_{Cq^*}(t, X_i, \delta_i) f(t, q^*)$. Similar arguments can be used to obtain an asymptotic expansion for (3) as $(3) \approx n^{-\frac{1}{2}} \sum U_{3i}(t)$ and therefore, the numerator,

$$\begin{aligned} \sqrt{n} [\widehat{N}(t, \theta, \widehat{H}) D(t) - N(t, \theta) \widehat{D}(t, \theta, \widehat{H})] &\approx n^{-\frac{1}{2}} \sum \{U_{1i}(t) + U_{2i}(t) + U_{3i}(t)\}. \text{ The same arguments as given above can be used to obtain an asymptotic expansion for} \\ \sqrt{n} \{ \widehat{\Pr}(\Delta(\theta) = 1 | T > t) - \Pr(\Delta(\theta) = 1 | T > t) \} &\text{ as } n^{-\frac{1}{2}} \sum_{i=1}^n D(t)^{-2} \{U_{-1i}(t) + U_{-2i}(t) + U_{-3i}(t)\} \\ \text{where } D(t)_-, U_{-1}(t), U_{-2}(t), \text{ and } U_{-3}(t) &\text{ are defined similarly to } D(t), U_{1i}(t), U_{2i}(t), \text{ and } U_{3i}(t) \text{ with } T \leq t \text{ replaced with } T > t. \text{ Therefore,} \\ \sqrt{n} \{ \widehat{\text{NRI}}(\theta, t) - \text{NRI}(\theta, t) \} &\approx n^{-\frac{1}{2}} \sum_{i=1}^n 2 [D(t)^{-2} \{U_{1i}(t) + U_{2i}(t) + U_{3i}(t)\} - D(t)^{-2} \{U_{-1i}(t) + U_{-2i}(t) + U_{-3i}(t)\}] \\ &= n^{-\frac{1}{2}} \sum_{i=1}^n \eta_i(t) \end{aligned}$$

Note that regardless of correct model specification, $\sqrt{n}(\widehat{\theta} - \theta_0) = n^{-1/2} \sum \psi_i + o_p(1)$ where ψ_i are i.i.d mean zero random variables by Lin and Wei (1989) and Uno et al. (2009). Using a Taylor series approximation and the i.i.d representation of $\sqrt{n}[\widehat{\text{NRI}}(\theta, t) - \text{NRI}(\theta, t)]$ for any θ , we can write $\widetilde{\mathcal{W}}(t) = \sqrt{n}[\widehat{\text{NRI}}(\widehat{\theta}, t) - \text{NRI}(\theta_0, t)]$ as a sum of i.i.d terms, $n^{-1/2} \sum_{i=1}^n \epsilon_i(t)$ defined below.

$$\begin{aligned} \sqrt{n} [\widehat{\text{NRI}}(\widehat{\theta}, t) - \text{NRI}(\theta_0, t)] &= \sqrt{n} [\widehat{\text{NRI}}(\widehat{\theta}, t) - \text{NRI}(\widehat{\theta}, t) + \text{NRI}(\widehat{\theta}, t) - \text{NRI}(\theta_0, t)] \\ &\approx \sqrt{n} [\widehat{\text{NRI}}(\widehat{\theta}, t) - \text{NRI}(\widehat{\theta}, t) + \frac{\partial \text{NRI}(t)}{\partial \theta} |_{\theta_0} (\widehat{\theta} - \theta_0)] \\ &= \sqrt{n} [\widehat{\text{NRI}}(\widehat{\theta}, t) - \text{NRI}(\widehat{\theta}, t)] + \sqrt{n} (\widehat{\theta} - \theta_0) \frac{\partial \text{NRI}(t)}{\partial \theta} |_{\theta_0} \\ &\approx \sqrt{n} [\widehat{\text{NRI}}(\widehat{\theta}, t) - \text{NRI}(\widehat{\theta}, t)] + n^{-1/2} \sum \psi_i \frac{\partial \text{NRI}(t)}{\partial \theta} |_{\theta_0} \\ &\approx n^{-1/2} \sum_{i=1}^n \eta_i(t) + n^{-1/2} \sum \psi_i \frac{\partial \text{NRI}(t)}{\partial \theta} |_{\theta_0} \\ &= n^{-1/2} \sum_{i=1}^n \epsilon_i(t) \end{aligned}$$

where $\epsilon_i(u, v, t) = \eta_i(u, v, t) + \psi_i \frac{\partial \text{NRI}(t)}{\partial \theta} |_{\theta_0}$. By a functional central limit theorem of Pollard (1990), the process $\widetilde{\mathcal{W}}(t)$ converges weakly to a mean zero Gaussian process in t .

Asymptotic Properties of $\widehat{\text{NRI}}(\widehat{\theta}, t)$

Recall that we assume the Cox model is correctly specified and thus,

$$Q(\theta_2) = Q(\theta_2, t, \mathbf{Y}_{(2)}) = \Pr(T \leq t | Y_{(2)}) = 1 - \exp \{ \Lambda_{02}(t) e^{\beta_2^T Y_{(2)}} \} \text{ and}$$

$S_{Q_i(\theta_2)}(t) = \Pr(T > t | Y_{(2)}) = \exp \{ \Lambda_{02}(t) e^{\beta_2^T Y_{(2)}} \}$. To derive asymptotic properties of $\widehat{\text{NRI}}(\widehat{\theta}, t)$ we assume the same regularity conditions as in Andersen and Gill (1982). The uniform

consistency of $Q(\widehat{\theta}_2, t, \mathbf{Y}_{(2)})$ for $Q(\theta_2, t, \mathbf{Y}_{(2)})$ in t and $\mathbf{Y}_{(2)}$ follows directly from the uniform consistency of $\widehat{\Lambda}_{02}(t)$ and $\widehat{\beta}_2$. It follows from the uniform law of large numbers (Pollard

1990) that $\widehat{\text{NRI}}(\widehat{\theta}, t)$ is uniformly consistent for $\text{NRI}(\theta_0, t)$. Andersen and Gill (1982) show

that $\sqrt{n}(\widehat{\beta}_2 - \beta_{02})$ is a normal random variable and $\sqrt{n}(\widehat{\Lambda}_{02}(t) - \Lambda_{02}(t))$ converges to a Gaussian process. By the functional delta method it can be shown that

$\sqrt{n} \{ Q(\widehat{\theta}_2, t, \mathbf{Y}_{(2)}) - Q(\theta_2, t, \mathbf{Y}_{(2)}) \}$ converges to a zero mean Gaussian process in t and $\mathbf{Y}_{(2)}$

(Zheng et al. 2008). Similar to the derivation for $\overline{\text{NRI}}(\bar{\theta}, t)$, it can be shown that the process $\tilde{\mathcal{N}}(t) = \sqrt{n} [\overline{\text{NRI}}(\bar{\theta}, t) - \text{NRI}(\theta_0, t)]$ is asymptotically equivalent to $n^{-1/2} \sum_{i=1}^n \zeta_i(u, v, t)$. In particular, for a fixed θ , $\sqrt{n} \{\overline{\text{NRI}}(\theta, t) - \text{NRI}(\theta, t)\} \approx n^{-1/2} \sum_{i=1}^n \eta_i^*(t)$ where $\eta_i^*(t) = 2 \left[D(t)^{-2} \{\Delta_i(\theta) Q_i(\theta_2) - \Pr(\Delta_i(\theta) = 1 | T_i \leq t) Q_i(\theta_2)\} - D(t)^{-2} \{\Delta_i(\theta) [1 - Q_i(\theta_2)] - \Pr(\Delta_i(\theta) = 1 | T_i > t) [1 - Q_i(\theta_2)]\} \right]$. Thus, $\tilde{\mathcal{N}}(t) \approx n^{-1/2} \sum_{i=1}^n \zeta_i(t)$ where $\zeta_i(u, v, t) = \eta_i^*(t) + \psi_i \frac{\partial \text{NRI}(t)}{\partial \theta} |_{\theta_0}$. Once again, using a functional central limit theorem, this implies that $\tilde{\mathcal{N}}(t)$ converges to a Gaussian process with mean zero.

References

- Andersen P, Gill R. Cox's regression model for counting processes: a large sample study. *Ann Stat*. 1982; 10:1100–1120.
- Biliyas Y, Gu M, Ying Z. Towards a general asymptotic theory for Cox model with staggered entry. *Ann Stat*. 1997; 25:662–682.
- Cai T, Tian L, Uno H, Solomon S, Wei L. Calibrating parametric subject-specific risk estimation. *Biometrika*. 2010; 97:389–404. [PubMed: 23049123]
- Cook N. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007; 115:928. [PubMed: 17309939]
- Cook N, Buring J, Ridker P. The effect of including c-reactive protein in cardiovascular risk prediction models for women. *Annals of Internal Medicine*. 2006; 145:21. [PubMed: 16818925]
- Cui J. Overview of risk prediction models in cardiovascular disease research. *Ann Epidemiol*. 2009; 19:711–717. [PubMed: 19628409]
- Dabrowska D. Smoothed cox regression. *Ann Stat*. 1997; 25(4):1510–1540.
- Du Y, Akritas M. Uniform strong representation of the conditional Kaplan–Meier process. *Math Methods Stat*. 2002; 11:152–182.
- Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. *J Am Stat Assoc*. 1997; 92(438):548–560.
- Gail M, Brinton L, Byar D, Corle D, Green S, Schairer C, Mulvihill J. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst (JNCI)*. 1989; 81:1879.
- Gu W, Pepe M. Measures to summarize and compare the predictive capacity of markers. *Int J Biostat*. 2009; 5:27.
- Hemann B, Bimson W, Taylor A. The framingham risk score: an appraisal of its benefits and limitations. *Am Heart Hosp J*. 2007; 5:91–96. [PubMed: 17478974]
- Hjort N. On inference in parametric survival data models. *Int Stat Rev*. 1992; 60(3):355–387.
- Kannel W, Feinleib M, McNamara P, Garrison R, Castelli W. An investigation of coronary heart disease in families. *Am J Epidemiol*. 1979; 110:281. [PubMed: 474565]
- Khot U, Khot M, Bajzer C, Sapp S, Ohman E, Brener S, Ellis S, Lincoff A, Topol E. Prevalence of conventional risk factors in patients with coronary heart disease. *JAMA*. 2003; 290:898–904. [PubMed: 12928466]
- Lin D, Wei L. The robust inference for the Cox proportional hazards model. *J Am Stat Assoc*. 1989; 84:1074–1078.
- Lloyd-Jones D. Cardiovascular risk prediction. *Circulation*. 2010; 121:1768–1777. [PubMed: 20404268]
- Pencina M, D'Agostino R Sr. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011; 30:11–21. [PubMed: 21204120]
- Pencina M, D'Agostino R Sr, D'Agostino R Jr. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008; 27:157–172. [PubMed: 17569110]

- Pollard, D. Empirical processes: theory and applications. Institute of Mathematical Statistics; Hayward: 1990.
- Satten G, Datta S. The kaplan-meier estimator as an inverse-probability-of-censoring weighted average. *Am Stat.* 2001; 55:207–210.
- Tian L, Cai T, Goetghebeur E, Wei L. Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika.* 2007; 94:297–311.
- Uno H, Cai T, Tian L, Wei L. Evaluating prediction rules for t-year survivors with censored regression models. *J Am Stat Assoc.* 2007; 102:527–537.
- Uno, H.; Tian, L.; Cai, T.; Kohane, I.; Wei, L. Comparing risk scoring systems beyond the roc paradigm in survival analysis. *Harvard University Biostatistics Working Paper Series.* 2009. p. 107
- Wilson P, D'Agostino R, Levy D, Belanger A, Silbershatz H, Kannel W. Prediction of coronary heart disease using risk factor categories. *Circulation.* 1998; 97:1837. [PubMed: 9603539]
- Zheng Y, Cai T, Pepe M, Levy W. Time-dependent predictive values of prognostic biomarkers with failure time outcome. *J Am Stat Assoc.* 2008; 103:362–368. [PubMed: 19655041]

Table 1

Simulation results under noninformative censoring and correctly specified new risk model with mean (mean of bias (mean(bias)) and standard deviation (Std. Dev.) of the estimated parameters across simulations, the mean of the standard error estimates calculated for each simulation using bootstrapping (mean(std error)), and coverage of the 95 % bootstrap confidence interval based on the normal approximation

| Method | $\Pr(P_i - Q_i > 0 \mid T_i = t)$ | $\Pr(P_i - Q_i > 0 \mid T_i > t)$ | NRI (t) |
|------------------------|-----------------------------------|-----------------------------------|-------------|
| True values | 0.592 | 0.358 | 0.468 |
| KM | | | |
| Mean(Bias) | 0.003 | 0.001 | 0.002 |
| Std. Dev. | 0.034 | 0.030 | 0.104 |
| Mean(std error) | 0.034 | 0.030 | 0.103 |
| 95 % bootstrap CI cov. | 0.946 | 0.946 | 0.946 |
| IPW | | | |
| Mean(Bias) | 0.002 | 0.002 | -0.001 |
| Std. Dev. | 0.034 | 0.030 | 0.105 |
| Mean(std error) | 0.034 | 0.031 | 0.104 |
| 95 % bootstrap CI cov. | 0.943 | 0.95 | 0.951 |
| Smooth IPW | | | |
| Mean(Bias) | 0.001 | 0.003 | -0.003 |
| Std. Dev. | 0.034 | 0.030 | 0.104 |
| Mean(std error) | 0.034 | 0.030 | 0.103 |
| 95 % bootstrap CI cov. | 0.946 | 0.942 | 0.949 |
| SEM | | | |
| Mean(Bias) | 0.001 | 0.003 | -0.003 |
| Std. Dev. | 0.024 | 0.029 | 0.082 |
| Mean(std error) | 0.025 | 0.028 | 0.080 |
| 95 % bootstrap CI cov. | 0.952 | 0.942 | 0.937 |
| Combined | | | |
| Mean(Bias) | 0.002 | 0.003 | -0.002 |
| Std. Dev. | 0.029 | 0.028 | 0.089 |
| Mean(std error) | 0.031 | 0.029 | 0.095 |
| 95 % bootstrap CI cov. | 0.968 | 0.949 | 0.969 |

KM Kaplan–Meier estimator, *IPW* inverse probability weighted estimator, *Smooth IPW* smooth inverse probability weighted estimator, *SEM* semiparametric estimator, *Combined* combined estimator, as defined in the text

Table 2

Simulation results under noninformative censoring and correctly specified new risk model with mean of bias (mean(Bias)) and standard deviation (Std. Dev.) of the estimated parameters across simulations, the mean of the standard error estimates calculated for each simulation using bootstrapping (mean(std error)), and coverage of the 95 % bootstrap confidence interval based on the normal approximation. Data is generated under the null model that $\beta_2 = 0$

| Method | | $\Pr(P_i - Q_i > 0 \mid T_i = t)$ | $\Pr(P_i - Q_i > 0 \mid T_i > t)$ | $\text{NRI}(t)$ |
|---------------------------|------------------------|-----------------------------------|-----------------------------------|-----------------|
| Null model: $\beta_2 = 0$ | | | | |
| | True values | 0.5 | 0.5 | 0 |
| KM | | | | |
| | Mean(Bias) | 0.01 | -0.02 | 0.061 |
| | Std. Dev. | 0.034 | 0.026 | 0.091 |
| | Mean(std error) | 0.043 | 0.033 | 0.118 |
| | 95 % bootstrap CI cov. | 0.996 | 0.971 | 0.98 |
| IPW | | | | |
| | Mean(Bias) | 0.01 | -0.019 | 0.058 |
| | Std. Dev. | 0.034 | 0.026 | 0.092 |
| | Mean(std error) | 0.044 | 0.033 | 0.119 |
| | 95 % bootstrap CI cov. | 0.996 | 0.972 | 0.981 |
| Smooth IPW | | | | |
| | Mean(Bias) | 0.009 | -0.019 | 0.055 |
| | Std. Dev. | 0.034 | 0.026 | 0.092 |
| | Mean(std error) | 0.044 | 0.033 | 0.118 |
| | 95 % bootstrap CI cov. | 0.996 | 0.972 | 0.981 |
| SEM | | | | |
| | Mean(Bias) | 0.009 | -0.019 | 0.057 |
| | Std. Dev. | 0.023 | 0.025 | 0.067 |
| | Mean(std error) | 0.029 | 0.031 | 0.081 |
| | 95 % bootstrap CI cov. | 0.99 | 0.967 | 0.957 |
| Combined | | | | |
| | Mean(Bias) | 0.008 | -0.019 | 0.055 |
| | Std. Dev. | 0.029 | 0.025 | 0.077 |
| | Mean(std error) | 0.039 | 0.032 | 0.104 |
| | 95 % bootstrap CI cov. | 0.997 | 0.971 | 0.977 |

KM Kaplan–Meier estimator, *IPW* inverse probability weighted estimator, *Smooth IPW* smooth inverse probability weighted estimator, *SEM* semiparametric estimator, *Combined* combined estimator, as defined in the text

Table 3

Simulation results under covariate-dependent censoring and correctly specified new risk model with mean of bias (mean(Bias)) and standard deviation (Std. Dev.) of the estimated parameters across simulations, the mean of the standard error estimates calculated for each simulation using bootstrapping (mean(std error)), and coverage of the 95 % bootstrap confidence interval based on the normal approximation

| Method | $\Pr(P_i - Q_i > 0 T_i = t)$ | $\Pr(P_i - Q_i > 0 T_i > t)$ | NRI(t) |
|------------------------|--------------------------------|--------------------------------|--------|
| True values | 0.611 | 0.45 | 0.322 |
| KM | | | |
| Mean(Bias) | 0.067 | -0.062 | 0.259 |
| Std. Dev. | 0.040 | 0.040 | 0.126 |
| Mean(std error) | 0.041 | 0.040 | 0.129 |
| 95 % bootstrap CI cov. | 0.615 | 0.659 | 0.483 |
| IPW | | | |
| Mean(Bias) | -0.024 | 0.005 | -0.057 |
| Std. Dev. | 0.034 | 0.045 | 0.131 |
| Mean(std error) | 0.035 | 0.044 | 0.130 |
| 95 % bootstrap CI cov. | 0.897 | 0.944 | 0.918 |
| Smooth IPW | | | |
| Mean(Bias) | -0.013 | 0.007 | -0.038 |
| Std. Dev. | 0.041 | 0.041 | 0.133 |
| Mean(std error) | 0.040 | 0.040 | 0.132 |
| 95 % bootstrap CI cov. | 0.937 | 0.939 | 0.941 |
| SEM | | | |
| Mean(Bias) | 0 | -0.001 | 0.002 |
| Std. Dev. | 0.025 | 0.039 | 0.098 |
| Mean(std error) | 0.026 | 0.037 | 0.095 |
| 95 % bootstrap CI cov. | 0.951 | 0.932 | 0.938 |
| Combined | | | |
| Mean(Bias) | -0.006 | 0.002 | -0.016 |
| Std. Dev. | 0.031 | 0.039 | 0.109 |
| Mean(std error) | 0.035 | 0.039 | 0.117 |
| 95 % bootstrap CI cov. | 0.975 | 0.951 | 0.971 |

KM Kaplan–Meier estimator, *IPW* inverse probability weighted estimator, *Smooth IPW* smooth inverse probability weighted estimator, *SEM* semiparametric estimator, *Combined* combined estimator, as defined in the text

Table 4

Simulation results under noninformative censoring and misspecified new risk model with mean of bias (mean(Bias)) and standard deviation (Std. Dev.) of the estimated parameters across simulations, the mean of the standard error estimates calculated for each simulation using bootstrapping (mean(std error)), and coverage of the 95 % bootstrap confidence interval based on the normal approximation

| Method | $\Pr(P_i - Q_i > 0 T_i = t)$ | $\Pr(P_i - Q_i > 0 T_i > t)$ | NRI (t) |
|------------------------|--------------------------------|--------------------------------|---------|
| True values | 0.646 | 0.395 | 0.504 |
| KM | | | |
| Mean(Bias) | 0.007 | -0.002 | 0.016 |
| Std. Dev. | 0.072 | 0.023 | 0.160 |
| Mean(std error) | 0.074 | 0.024 | 0.164 |
| 95 % bootstrap CI cov. | 0.94 | 0.945 | 0.947 |
| IPW | | | |
| Mean(Bias) | 0.004 | -0.001 | 0.008 |
| Std. Dev. | 0.072 | 0.023 | 0.160 |
| Mean(std error) | 0.074 | 0.024 | 0.165 |
| 95 % bootstrap CI cov. | 0.945 | 0.942 | 0.95 |
| Smooth IPW | | | |
| Mean(Bias) | 0.003 | -0.001 | 0.007 |
| Std. Dev. | 0.072 | 0.023 | 0.160 |
| Mean(std error) | 0.074 | 0.024 | 0.164 |
| 95 % bootstrap CI cov. | 0.943 | 0.946 | 0.95 |
| SEM | | | |
| Mean(Bias) | -0.046 | 0.003 | -0.099 |
| Std. Dev. | 0.022 | 0.022 | 0.068 |
| Mean(std error) | 0.022 | 0.023 | 0.068 |
| 95 % bootstrap CI cov. | 0.448 | 0.943 | 0.682 |
| Combined | | | |
| Mean(Bias) | -0.009 | 0.000 | -0.020 |
| Std. Dev. | 0.057 | 0.022 | 0.128 |
| Mean(std error) | 0.062 | 0.023 | 0.139 |
| 95 % bootstrap CI cov. | 0.970 | 0.947 | 0.976 |

KM Kaplan–Meier estimator, *IPW* inverse probability weighted estimator, *Smooth IPW* smooth inverse probability weighted estimator, *SEM* semiparametric estimator, *Combined* combined estimator, as defined in the text

Table 5

Simulation results comparing apparent estimates and the 0.632 bootstrap for correcting overfitting

| Estimator | $\Pr(P_i - Q_i > 0 T_i = t)$ | $\Pr(P_i - Q_i > 0 T_i > t)$ | $\text{NRI}(t)$ |
|-----------------|--------------------------------|--------------------------------|-----------------|
| True values | 0.684 | 0.275 | 0.817 |
| Smooth IPW | | | |
| Apparent | | | |
| Mean(Bias) | 0.000 | 0.004 | -0.007 |
| Std. Dev. | 0.036 | 0.028 | 0.108 |
| CI coverage | 0.962 | 0.963 | 0.964 |
| 0.632 Bootstrap | | | |
| Mean(Bias) | -0.008 | 0.008 | -0.032 |
| Std. Dev. | 0.034 | 0.027 | 0.102 |
| CI coverage | 0.971 | 0.969 | 0.968 |
| Bootstrapped SE | | | |
| Mean(std error) | 0.039 | 0.030 | 0.114 |
| SEM | | | |
| Apparent | | | |
| Mean(Bias) | 0.003 | -0.001 | 0.009 |
| Std. Dev. | 0.023 | 0.025 | 0.072 |
| CI coverage | 0.955 | 0.954 | 0.945 |
| 0.632 Bootstrap | | | |
| Mean(Bias) | 0.005 | -0.003 | 0.015 |
| Std. Dev. | 0.022 | 0.024 | 0.072 |
| CI coverage | 0.953 | 0.962 | 0.937 |
| Bootstrapped SE | | | |
| Mean(std error) | 0.024 | 0.025 | 0.071 |
| Combined | | | |
| Apparent | | | |
| Mean(Bias) | 0.001 | 0.001 | 0.001 |
| Std. Dev. | 0.028 | 0.026 | 0.087 |
| CI coverage | 0.982 | 0.969 | 0.975 |
| 0.632 Bootstrap | | | |
| Mean(Bias) | -0.002 | 0.003 | -0.008 |
| Std. Dev. | 0.027 | 0.025 | 0.085 |
| CI coverage | 0.989 | 0.975 | 0.983 |
| Bootstrapped SE | | | |
| Mean(std error) | 0.035 | 0.028 | 0.102 |

Smooth IPW smooth inverse probability weighted estimator, *SEM* semiparametric estimator, *Combined* combined estimator, as defined in the text

Table 6

NRI estimates for two risk models for predicting 10-year CVD risk among women in the Framingham offspring cohort

| Method | $\Pr(P_i - Q_i > 0 \mid T_i = t)$ | $\Pr(P_i - Q_i > 0 \mid T_i > t)$ | $\$NRI(t)$ |
|------------|-----------------------------------|-----------------------------------|------------|
| KM | | | |
| Est | 0.483 | 0.508 | -0.049 |
| SE | 0.069 | 0.028 | 0.176 |
| IPW | | | |
| Est | 0.478 | 0.508 | -0.059 |
| SE | 0.070 | 0.028 | 0.178 |
| Smooth IPW | | | |
| Est | 0.480 | 0.508 | -0.057 |
| SE | 0.070 | 0.028 | 0.178 |
| SEM | | | |
| Est | 0.587 | 0.503 | 0.167 |
| SE | 0.015 | 0.026 | 0.067 |
| Combined | | | |
| Est | 0.570 | 0.504 | 0.132 |
| SE | 0.054 | 0.027 | 0.137 |

KM Kaplan–Meier estimator, *IPW* inverse probability weighted estimator, *Smooth IPW* smooth inverse probability weighted estimator, *SEM* semiparametric estimator, *Combined* combined estimator, as defined in the text