Covariate-adjusted measures of discrimination for survival data

Ian R. White*[1] and **Eleni Rapsomaniki**[2] for the **Emerging Risk Factors Collaboration**[3]

[1] MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK
[2] Farr Institute for Health Informatics Research, Department of Epidemiology and Public Health, University College London Medical School, 222 Euston Road, London WC1E 6BT, UK
[3] Members of the Emerging Risk Factors Collaboration are listed in Appendix A

Received 18 March 2014; revised 15 July 2014; accepted 11 August 2014

Motivation: Discrimination statistics describe the ability of a survival model to assign higher risks to individuals who experience earlier events: examples are Harrell's C-index and Royston and Sauerbrei's D, which we call the D-index. Prognostic covariates whose distributions are controlled by the study design (e.g. age and sex) influence discrimination and can make it difficult to compare model discrimination between studies. Although covariate adjustment is a standard procedure for quantifying disease-risk factor associations, there are no covariate adjustment methods for discrimination statistics in censored survival data. **Objective:** To develop extensions of the C-index and D-index that describe the prognostic ability of a model adjusted for one or more covariate(s). **Method:** We define a covariate-adjusted C-index and D-index for censored survival data, propose several estimators, and investigate their performance in simulation studies and in data from a large individual participant data meta-analysis, the Emerging Risk Factors Collaboration. **Results:** The proposed methods perform well in simulations. In the Emerging Risk Factors Collaboration data, the age-adjusted C-index and D-index were substantially smaller than unadjusted values. The study-specific standard deviation of baseline age was strongly associated with the unadjusted C-index and D-index but not significantly associated with the age-adjusted indices. **Conclusions:** The proposed estimators improve meta-analysis comparisons, are easy to implement and give a more meaningful clinical interpretation.

Keywords: C-index; D-index; Discrimination.

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

A fundamental property of a prognostic marker is its ability to discriminate high from low risk patients (Hlatky et al., 2009). Markers that do not improve discrimination are also unlikely to improve other measures of clinical performance (Mihaescu et al., 2010). The discrimination performance of a marker and its incremental value can vary significantly across different studies. Variation beyond chance can be attributed to differences between studies either in the strength of association between the marker and the outcome or in the marker's distribution (Pepe et al., 2004), or to a range of other possible biases that relate to the conduct and recording in a study (Lijmer et al., 2002). In practice, markers are often used in combination, so interest lies in evaluating the discrimination of a prognostic model including

*Corresponding author: e-mail: ian.white@mrc-bsu.cam.ac.uk, Phone: +44-1223-330399, Fax: +44-1223-330365

one or more markers together with demographic variables. Other important aspects of prognostic ability include calibration.

For a binary outcome, the standard description of discrimination is the receiver operating characteristic (ROC) curve, which displays the trade-off between specificity (the probability that a control marker value is below the cut-off) and sensitivity (the probability that a case marker value is above the cut-off) for different marker cut-offs. The ROC curve is often summarized by the area under the curve (AUC), also called the C-statistic, which can be interpreted as the probability that the marker will correctly classify a randomly chosen pair of patients as case and noncase (Hanley and McNeil, 1982). The AUC ranges from 0 (when all predictions are wrong) and 1 (perfect predictions) with 0.5 representing the average discriminative ability of random predictions. Values below 0.5 are rarely seen other than due to small sample variation.

For a survival outcome, many measures of prognostic ability have been proposed (Choodari-Oskooei et al., 2012a, 2012b). The C-statistic can be used to measure discrimination in this setting, taking the binary outcome to be survival to a fixed follow-up time (Chambless and Diao, 2006). The C-index (Harrell et al., 1982) extends the C-statistic and avoids specifying a fixed follow-up time: it estimates the probability that given two randomly drawn patients, the patient who has an event first is predicted a higher risk. Royston and Sauerbrei (2004) proposed an alternative measure, D, which we call the D-index: it is based on a proportional hazards model and has the interpretation of an average log hazard ratio between an individual in the upper half of the risk distribution and an individual in the lower half (Pennells et al., 2014). We use the C-index because it is the most widely used measure in practice (Mallett et al., 2010), and the D-index because it adapts well to the purposes of this paper; the two measures give similar conclusions when used to evaluate the discrimination added by a new marker (Fibrinogen Studies Collaboration, 2009).

Covariate adjustment is necessary for correct assessment of disease-risk factor associations in observational studies, but its importance is rarely acknowledged in assessing discrimination. A particular issue is that covariates that form part of the study design such as age and sex can impact substantially on the prognostic ability of a model (Janes and Pepe, 2008; Kerr and Pepe, 2011): for example, the prognostic ability of a cardiovascular risk model (which includes age, sex, clinical covariates, and biomarkers) is likely to be substantially larger in a study that recruited men and women aged 40–80 than in a similar study that only recruited men aged 55–65. Janes and Pepe (2008), writing in the context of ROC curves, identify three cases in which covariates Z influence the discrimination performance of a risk score R:

(1) The covariate Z is associated both with R and with disease risk. A common example is when Z is age. Janes and Pepe (2008) show that stratifying by categorical Z reduces the discrimination of R. This reduction is greater if R and Z are highly correlated.
(2) A different issue arises if the covariates Z are associated with R but not with disease risk. Ignoring this type of covariate effect may underestimate discrimination (Janes and Pepe, 2008). In cardiovascular disease, R might be a function of C-reactive protein, a cardiovascular risk factor, and Z might be acute infection, which strongly raises C-reactive protein. Allowing for Z in the analysis could remove a source of noise in R and hence improve discrimination.
(3) The discrimination of R may vary across levels of Z, which is analogous to effect modification. This situation arises if the hazards associated with various levels of Z vary: for example, associations with blood pressure measurements are attenuated with increasing age (Prospective Studies Collaboration, 2002). It also arises when associations remain constant but the spread of the distribution of X varies with Z.

Covariate adjustment for measures of discrimination has been tackled in the context of diagnostic tests using ROC curves based on binary outcomes (Janes et al., 2009). However, there are currently no methods to adjust the discrimination performance of prognostic markers for covariates for censored survival data. In this paper, we propose definitions of the C-index and D-index for censored survival

data allowing adjustment for one or more binary or continuous covariates, and ways to estimate them. Although we primarily aim to adjust for age and sex, the methods are presented for a general covariate adjustment.

The paper is arranged as follows. In Section 2 we describe the data which motivated our methods, and we define the model used to generate risk predictions. In Section 3, we review unadjusted measures of discrimination. In Section 4, we propose definitions of adjusted measures of discrimination, and our new estimators. In Section 5, we examine the performance of the proposed estimators in a simulation study. In Section 6 we apply the proposed methods to data from a large set of epidemiological cohort studies. We conclude in Section 7 with a discussion of our results, recommendations for when our proposed estimators might be appropriate, desirable extensions and limitations.

2 Data

The Emerging Risk Factors Collaboration (ERFC) has collated and harmonized individual participant data from population-based prospective studies of cardiovascular disease (CVD) (Emerging Risk Factors Collaboration, 2007). In May 2011 the data set comprised 1.9 million individuals in 108 studies with an average of 15.5 years follow-up. We used these data to model time to first fatal/nonfatal CVD event, which includes coronary heart disease and stroke. Our main dataset was restricted to prospective cohorts and clinical trials that provide information on Framingham risk factors, that is age, smoking status (current/ex vs. never), systolic blood pressure (SBP), total cholesterol (TC), high-density lipoprotein (HDL) cholesterol and history of diabetes at the baseline survey. Individual participant data were further restricted to subjects aged at least 40 years at baseline with the above risk factors recorded, no known history of CVD at the baseline survey, no recorded history of diabetes, and not known to be under statin treatment. Thus, our analysis data comprised 349,137 individuals from 82 studies (of which eight were clinical trials), of whom 24,369 experienced a CVD event.

The model fitted to these data was the Cox proportional hazards model stratified by study and sex. Studies that were randomized trials were additionally stratified by trial arm. Thus, for individual i in stratum s, the hazard at time t is

$$h_{si}(t) = h_{0s}(t) \exp(\boldsymbol{\beta}\mathbf{x}_i), \tag{1}$$

where $\mathbf{x}_i$ is the vector of covariates (age, smoking status, SBP, TC, and HDL cholesterol) for individual i, $\boldsymbol{\beta}$ is a vector of corresponding regression coefficients (assumed constant across strata), and $h_{0s}(t)$ is the baseline hazard at time t for individuals of stratum s. Table 1 summarizes the data from these 82 studies and the fitted Cox model.

The within-study distribution of baseline age and sex is determined by a study design and hence differs between studies. This may affect measures of discrimination. Model (1) is stratified by sex, so standard calculations automatically stratify measures of discrimination for sex. We therefore focus on the effect of baseline age. The left-hand panel in Fig. 1 plots the C-index for each study (computed as described in Section 6) against the within-study standard deviation (SD) of baseline age. Studies with more variation in baseline age tend to have substantially larger C-indices. The age-adjusted C-index, introduced in Section 4 below, is plotted in the right-hand panel, and shows no association with variation in baseline age.

3 Measures of discrimination in the absence of covariate adjustment

3.1 Notation

We initially work in a single dataset of n individuals. For each individual $i = 1, \ldots, n$, we assume that the model covariates are $\mathbf{x}_i$ (scalar or vector), the true event time (in the absence of censoring) is t_i^*, and

Table 1 ERFC data: variable summaries and selected model.

Variable	Mean	Within-studies SD	Between-studies SD	Fitted model Coef.	Fitted model Std. Err.
Baseline variables					
Age (years)	55.92	7.57	6.78	0.0771	0.0009
Smoking (0 = no, 1 = yes)	0.30	0.43	0.15	0.563	0.014
SBP (mm Hg)	133.30	18.51	7.70	0.0146	0.0003
Total cholesterol (mmol/L)	5.87	1.07	0.46	0.168	0.006
HDL cholesterol (mmol/L)	1.35	0.37	0.15	−0.512	0.020
Sex (0 = male, 1 = female)	0.42	0.40	0.29	(stratifier)	
Outcome variables					
Follow-up (years)	10.64	3.65	4.87		
CVD event (0 = no, 1 = yes)	0.07	0.24	0.07		

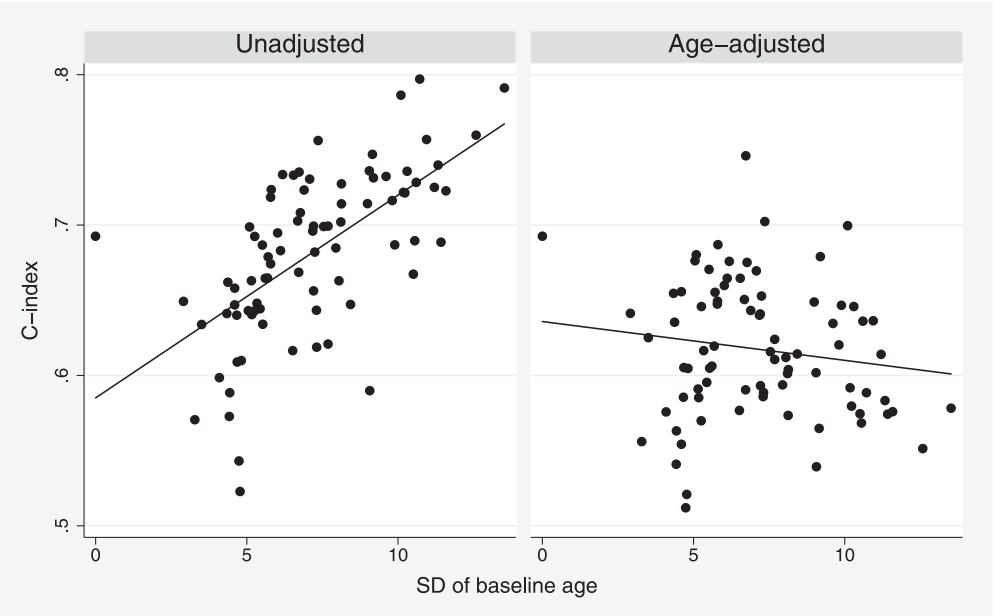

Figure 1 ERFC data: unadjusted and age-adjusted C-index for a model including baseline age, smoking, SBP, TC, and HDL, plotted for each study against the SD of baseline age in that study. Analyses are stratified by sex and trial arm. Each point represents one study.

censoring time is c_i, so the event indicator is $d_i = 1(t_i^* \leq c_i)$ and the observed time is $t_i = \min(t_i^*, c_i)$. The observed data are (x_i, d_i, t_i). Censoring is assumed to be noninformative.

We also assume that the risk score is (scalar) $r(\mathbf{x}_i)$, which is typically (but not necessarily) the linear predictor $\hat{\boldsymbol{\beta}}_x \mathbf{x}$ from fitting a survival model such as $h(t) = h_0(t) \exp(\boldsymbol{\beta}_x \mathbf{x})$ where $\boldsymbol{\beta}_x$ are coefficients and $h_0(t)$ is the baseline hazard. Our aim is to evaluate the discrimination of $r(\mathbf{x}_i)$.

3.2 C-index

Harrell et al. (1982) defined the C-index as a statistic measuring the degree to which sample pairs are concordant, where concordance occurs if the individual of higher predicted risk has the first event in the pair. This statistic is affected by censoring (see Section 3.2.1). The underlying estimand was defined by Heagerty and Zheng (2005) and Uno et al. (2011) as $C = P(r(\mathbf{x}_i) > r(\mathbf{x}_j)|t_i^* < t_j^*)$. Gonen and Heller (2005) instead stated the estimand $K = P(t_i^* < t_j^*|r(\mathbf{x}_i) \geq r(\mathbf{x}_j))$.

These estimands are equivalent in the absence of ties in $r(\mathbf{x}_i)$ (i.e., if all individuals have different values of $r(\mathbf{x}_i)$). We believe that ties in $r(\mathbf{x}_i)$ are important, since poorly discriminating models may have many ties, so it is important to account for them. Heagerty and Zheng's C counts pairs tied on $r(\mathbf{x}_i)$ as discordant, while K double-counts them (because they satisfy both $r(\mathbf{x}_i) \geq r(\mathbf{x}_j)$ and $r(\mathbf{x}_j) \geq r(\mathbf{x}_i)$). Instead, we make the natural definition of the C-index as

$$C = \mathrm{E}\left[C_{ij}\right] \text{ where } C_{ij} = \begin{cases} 1(t_i^* < t_j^*) & \text{if } r(\mathbf{x}_i) > r(\mathbf{x}_j) \\ 0.5 & \text{if } r(\mathbf{x}_i) = r(\mathbf{x}_j) \\ 1(t_i^* > t_j^*) & \text{if } r(\mathbf{x}_i) < r(\mathbf{x}_j) \end{cases} \tag{2}$$

for a random pair (i, j), where $1(a) = 1$ if a is true and 0 if a is false.

Pairs with tied event times are excluded from all calculations based on C_{ij}, so that the estimand becomes $\mathrm{E}[C_{ij}|t_i^* \neq t_j^*]$. For simplicity, however, we ignore tied event times in the notation throughout this article.

We now consider various estimators of C in Eq. (2).

3.2.1 Harrell's estimator

Estimation of C is complicated by the presence of censoring, because we do not know whether $t_i^* < t_j^*$ for pairs where the first event time is censored. Harrell et al. (1996) proposed estimating C as the mean of C_{ij} over informative pairs, where pair (i, j) is informative if $t_i^* < t_j^*$ and $d_i = 1$ or $t_i^* > t_j^*$ and $d_j = 1$: that is, if the first event in the pair is observed. Harrell's estimator is often written as

$$\hat{C}_{Har} = \frac{\#\text{concordant} + \frac{1}{2}\#\text{tied}}{\#\text{concordant} + \#\text{discordant} + \#\text{tied}}$$

where #concordant counts pairs with $t_i^* < t_j^*$ and $r(\mathbf{x}_i) > r(\mathbf{x}_j)$, or $t_i^* > t_j^*$ and $r(\mathbf{x}_i) < r(\mathbf{x}_j)$; #tied counts pairs with $r(\mathbf{x}_i) = r(\mathbf{x}_j)$; and #discordant counts pairs with $t_i^* < t_j^*$ and $r(\mathbf{x}_i) < r(\mathbf{x}_j)$, or $t_i^* > t_j^*$ and $r(\mathbf{x}_i) > r(\mathbf{x}_j)$. However, the informative pairs are not representative of all pairs—for example, a pair of low-risk individuals is likely to have no event and hence be noninformative—and this can cause bias in $\hat{C}_{Har}$ (Gonen and Heller, 2005).

3.2.2 Gonen and Heller's estimator

Gonen and Heller (2005) proposed an alternative estimator to avoid bias due to censoring. To present the idea in greater generality, suppose $r^*(\mathbf{x}_i)$ is a linear predictor from a correctly specified proportional hazards model. Then $r^*(\mathbf{x}_i) - r^*(\mathbf{x}_j)$ represents the log hazard ratio between individuals i and j, and the probability that individuals i and j are concordant is $\mathrm{expit}\{r^*(\mathbf{x}_i) - r^*(\mathbf{x}_j)\}$ if $r(\mathbf{x}_i) > r(\mathbf{x}_j)$ where $\mathrm{expit}(\eta) = 1/(1 + \exp(-\eta))$. (Similarly it is $\mathrm{expit}\{r^*(\mathbf{x}_j) - r^*(\mathbf{x}_i)\}$ if $r(\mathbf{x}_i) < r(\mathbf{x}_j)$, and 0.5 if $r(\mathbf{x}_i) = r(\mathbf{x}_j)$). Then the estimator is the average of this concordance probability, which can be written as

$$\hat{C}_{ind} = \frac{1}{n(n-1)} \sum_{i,j} \mathrm{expit}\left\{[r^*(\mathbf{x}_i) - r^*(\mathbf{x}_j)]\mathrm{sign}\left[r(\mathbf{x}_i) - r(\mathbf{x}_j)\right]\right\}. \tag{3}$$

Gonen and Heller (2005) considered the special case $r^*(\mathbf{x}_i) = r(\mathbf{x}_i)$ (that is, they assumed that $r(\mathbf{x}_i)$ is a linear predictor from a correctly specified proportional hazards model) giving the simpler expression

$$\hat{C}_{ind} = \frac{1}{n(n-1)} \sum_{i,j} \text{expit}\left(|r(\mathbf{x}_i) - r(\mathbf{x}_j)|\right). \tag{4}$$

$\hat{C}_{ind}$ is an indirect measure, since it does not use the event times and relies on correct model specification.

3.2.3 Restricted C-index

Let $\tau = \max_i t_i$ be the longest follow-up time observed. The study only gives information about discrimination at time $t \leq \tau$, and C can only be estimated by (implicitly) extrapolating to times $t > \tau$: for example, $\hat{C}_{ind}$ assumes that the proportional hazards model continues to hold at times beyond τ. To avoid extrapolation, Heagerty and Zheng (2005) proposed the restricted C-index

$$C^\tau = P(r(\mathbf{x}_i) > r(\mathbf{x}_j) | t_i^* < t_j^*, t_i^* < \tau)$$

which is estimable without extrapolation in a study with follow-up at least up to time τ. They and Uno et al. (2011) proposed estimators of C^τ to account for censoring before time τ: that of Uno et al. (2011) involves a weighted mean of C_{ij} over informative pairs, where the weight for pair (i, j) is $\hat{G}(\min(t_i, t_j))^{-2}$ and $G(t) = P(c_i \geq t)$.

3.3 D-index

Royston and Sauerbrei (2004) proposed a measure, D, which we also call the D-index, with the interpretation of the log hazard ratio between two equal-sized prognostic groups. It is estimated in a two-stage procedure. In stage 1, the values of $r(\mathbf{x}_i)$ are ranked, converted to normal scores, and multiplied by $\sqrt{\pi/8}$. In stage 2, a proportional hazards regression is performed on the scaled normal scores, and D is the regression coefficient.

As in the work of Harrell et al. (1982), the estimand is not immediately clear. A possible estimand is based on pairs: still assuming that $r^*(\mathbf{x}_i) - r^*(\mathbf{x}_j)$ is the true log hazard ratio between individuals i and j, the estimand D can be defined as the average of this log hazard ratio when i is drawn randomly from the upper half of the risk distribution and j is drawn randomly from the lower half (Pennells et al., 2014): that is,

$$D = \text{E}\left[r^*(\mathbf{x}_i) - r^*(\mathbf{x}_j) | r(\mathbf{x}_i) > \bar{r} > r(\mathbf{x}_j)\right] \tag{5}$$

where $\bar{r}$ is the mean of the $r(\mathbf{x}_i)$. The algorithm above clearly estimates this estimand consistently when the proportional hazards model is correctly specified and $r(\mathbf{x}_i)$ is normally distributed, but it may be biased when $r(\mathbf{x}_i)$ is skewed (Choodari-Oskooei et al., 2012a).

4 Measures of discrimination with covariate adjustment

Let $\mathbf{z}_i$ be covariates, which may or may not form part of $\mathbf{x}_i$. We aim to evaluate the risk score $r(\mathbf{x}_i)$ while adjusting for the covariates $\mathbf{z}_i$. Conceptually, we want to estimate C and D if we had a sample with a common value of z, or by restricting attention to pairs with equal values of z. In the ERFC data of Section 2, $r(\mathbf{x}_i)$ is a cardiovascular risk prediction while z_i is age.

4.1 Adjusted C

4.1.1 Estimand

Covariate adjustment can be defined by considering pairs that match exactly on Z, so that

$$C(z) = \mathrm{E}\left[C_{ij}|z_i = z_j = z\right]$$

is a z-specific C. For situations where $C(z)$ is roughly constant over z, or where a summary measure of discrimination is required, the z-adjusted C

$$C^{adj} = \mathrm{E}\left[C_{ij}|z_i = z_j\right]$$

is the natural measure when a risk model is stratified by z, and can be applied more widely.

Note that for continuous z with density $f(z)$, pairs matching on z have density proportional to $f(z)^2$, so that

$$C^{adj} = \int C(z)f(z)^2 dz \int f(z)^2 dz. \tag{6}$$

It is natural to consider weighting by $f(z)$ rather than $f(z)^2$ in (6), so we also define

$$C^{adj,w} = \int C(z)f(z)dz \int f(z)dz$$

although other choices of weights are also possible. Of course, $C^{adj,w} = C^{adj}$ if $C(z)$ is constant.

4.1.2 Direct estimation for categorical Z

We describe an estimator as direct (like $\hat{C}_{Har}$) if it uses actual event times, and indirect (like $\hat{C}_{ind}$) if instead it uses risks predicted under a model. Direct estimation is tricky with continuous z, as there may be few or no matching pairs (Section 4.1.3). We therefore first consider the case with categorical z. Simple estimators are $\hat{C}(z) = \left\{\sum_{(i,j):z_i=z_j=z} C_{ij}\right\}/\left\{\sum_{(i,j):z_i=z_j=z} 1\right\}$, $\hat{C}^{adj} = \left\{\sum_{(i,j):z_i=z_j} C_{ij}\right\}/\left\{\sum_{(i,j):z_i=z_j} 1\right\}$ and $\hat{C}^{adj,w} = \left\{\sum_z \hat{f}(z)\hat{C}(z)\right\}/\left\{\sum_z \hat{f}(z)\right\}$, where $\hat{f}(z) = \sum_{i:z_i=z} 1$. In the presence of censoring, these sums are restricted to informative pairs, and the weighting scheme of Uno et al. (2011) may be used to handle random censoring.

4.1.3 Direct estimation for continuous or multivariate Z

For some methods, it is helpful to decompose

$$r(\mathbf{x}_i) = m(\mathbf{x}_i, z_i) + \hat{r}(z_i) \tag{7}$$

where $m(\mathbf{x}_i, z_i)$ is uncorrelated with z_i. This is easily done by fitting a suitable regression for $r(\mathbf{x}_i)$ on z_i, and defining $\hat{r}(z_i) = \mathrm{E}[r(\mathbf{x}_i)|z_i]$ as the fitted value and $m(\mathbf{x}_i, z_i)$ as the residual. Conceptually, we want to estimate the discrimination that is due to $m(\mathbf{x}_i, z_i)$. The methods proposed in this section do not assume that $m(\mathbf{x}_i, z_i)$ is independent of z_i, unlike the methods proposed in Section 4.1.4.

We propose direct estimation by plotting C_{ij} against $\hat{r}(z_j) - \hat{r}(z_i)$ or $|\hat{r}(z_j) - \hat{r}(z_i)|$, fitting a suitable model (parametric or nonparametric), and taking $\hat{C}^{adj}_{smooth1}$ as the fitted value at $\hat{r}(z_j) - \hat{r}(z_i) = 0$. In order to automate the procedure, we use a logistic regression of C_{ij} on $(\hat{r}(z_j) - \hat{r}(z_i))^2$ with weights

$$w_1(z_i, z_j) = \exp\left(-\lambda[\hat{r}(z_j) - \hat{r}(z_i)]^2\right) \tag{8}$$

where λ controls the amount of smoothing. $\hat{C}^{adj}_{smooth1}$ is then the inverse logit of the estimated intercept. The procedure is illustrated in Supporting Information Fig. S1.

Again, in the presence of censoring, the sums are restricted to informative pairs, and the weighting scheme of Uno et al. (2011) may be used to handle random censoring.

We estimate $\hat{C}^{adj,w}$ using the same logistic regression but with weights $w_1(z_i, z_j)w_2(z_i, z_j)$ where

$$w_2(z_i, z_j) = \{\hat{f}(z_i)\hat{f}(z_j)\}^{-1/2} \tag{9}$$

since this weight approximates $1/f(z)$ when $z_i \approx z_j$. Here, $\hat{f}(z)$ might be a kernel estimate of the density of z.

The above method is based on C_{ij} which represents whether two events occur in the order predicted by $r(\mathbf{x})$. An alternative is to explore whether events occur in the order predicted by $m(\mathbf{x}, z)$. We define "m-concordance" C^m_{ij} by replacing conditions $r(\mathbf{x}_i) > r(\mathbf{x}_j)$ etc. in Eq. (2) with $m(\mathbf{x}_i, z_i) > m(\mathbf{x}_j, z_j)$ etc. Again, fitted values of the mean of C^m_{ij} at $\hat{r}(z_j) - \hat{r}(z_i) = 0$ give an estimator of $\hat{C}^{adj}$, which we denote by $\hat{C}^{adj}_{smooth2}$.

Comparing the two estimators $\hat{C}^{adj}_{smooth1}$ and $\hat{C}^{adj}_{smooth2}$ may help to detect an unsuitable value of λ. Too small a value causes bias by giving too much weight to mismatched pairs, while too large a value causes large variance by reducing the effective number of pairs used. We used the ERFC data to compare $\hat{C}^{adj}_{smooth1}$ with $\hat{C}^{adj}_{smooth2}$ (Supporting Information Fig. S2) and to compare their standard errors (Supporting Information Fig. S3), for $0 \leq \lambda \leq 10$. Values $\lambda < 1$ tended to give large differences between the two estimators, but values in the range 1–10 seemed broadly reasonable: later work uses $\lambda = 3$.

4.1.4 Indirect estimation of an approximate estimand

Now we use the correctly specified linear predictor $r^*(\mathbf{x})$, which we decompose as $r^*(\mathbf{x}_i) = m^*(\mathbf{x}_i, z_i) + \hat{r}^*(z_i)$ as in (7). Recall that $P(C_{ij} = 1|\mathbf{x}_i, \mathbf{x}_j) = \text{expit}\{r^*(\mathbf{x}_i) - r^*(\mathbf{x}_j)\}$ when $r(\mathbf{x}_i) > r(\mathbf{x}_j)$, etc. Hence for pairs that match on z, $P(C_{ij} = 1|\mathbf{x}_i, \mathbf{x}_j) = \text{expit}\,(m^*(\mathbf{x}_i, z_i) - m^*(\mathbf{x}_j, z_j))$ when $m(\mathbf{x}_i, z_i) > m(\mathbf{x}_j, z_j)$, etc. This suggests defining a new estimand

$$C^{adj*} = \mathrm{E}\left[\text{expit}\left\{\left[m^*(\mathbf{x}_i, z_i) - m^*(\mathbf{x}_j, z_j)\right]\text{sign}\left[m(\mathbf{x}_i, z_i) - m(\mathbf{x}_j, z_j)\right]\right\}\right]. \tag{10}$$

$C^{adj*} = C^{adj}$ if $m(\mathbf{x}_i, z_i)$ is independent of z_i. Appendix B and the simulation study demonstrate that $C^{adj*} \neq C^{adj}$ in general, but differences are not large.

Analogous to (3), we propose the indirect estimator in the correctly specified case $m^*(\mathbf{x}, z) = m(\mathbf{x}, z)$:

$$\hat{C}^{adj*}_{ind} = \frac{1}{n(n-1)}\sum_{i,j}\text{expit}\left(|m(\mathbf{x}_i, z_i) - m(\mathbf{x}_j, z_j)|\right). \tag{11}$$

Like $\hat{C}_{ind}$, this estimator is unaffected by censoring, but requires correct model specification.

4.1.5 Recalibrating

To be useful in practice, a risk score must be well calibrated. Ideally, this is ensured by recalibrating the model in an external validation set. However, sometimes miscalibrated risk scores are evaluated, and in this case we want to be sure that the miscalibration does not distort the C-index.

The advantage of a direct method is that it should give correct results if the risk score is miscalibrated. The indirect methods above are very susceptible to miscalibration. However, even the direct methods of Section 4.1.3 are slightly affected by miscalibration, because the weights in (8) are affected. We therefore propose preceding all the above methods, except for Harrell's method (which is unaffected by miscalibration), by a recalibration step.

For the unadjusted indirect method, we assume $r^*(\mathbf{x}) = \gamma_r r(\mathbf{x})$ and estimate γ_r by fitting the Cox model

$$h_i(t) = h_0(t) \exp\left\{\gamma_r r(\mathbf{x}_i)\right\}.$$

If $r(\mathbf{x})$ is well calibrated, then $\hat{\gamma}_r \approx 1$. The recalibrated estimate is

$$\hat{C}_{ind,recal} = \frac{1}{n(n-1)} \sum_{i,j} \text{expit}\left(\hat{\gamma}_r |r(\mathbf{x}_i) - r(\mathbf{x}_j)|\right). \tag{12}$$

For the adjusted methods, we assume $m^*(\mathbf{x}, z) = \gamma_m m(\mathbf{x}, z)$ and $\hat{r}^*(\mathbf{x}) = \gamma_r \hat{r}(\mathbf{x})$ and estimate γ_m and γ_z by fitting the Cox model

$$h_i(t) = h_0(t) \exp\left\{\gamma_m m(\mathbf{x}_i, z_i) + \gamma_z \hat{r}(z_i)\right\}$$

with fixed $m(\mathbf{x}_i, z_i)$ and $\hat{r}(z_i)$. The recalibrated estimate is

$$\hat{C}^{adj*}_{ind,recal} = \frac{1}{n(n-1)} \sum_{i,j} \text{expit}\left(\hat{\gamma}_m |m(\mathbf{x}_i, z_i) - m(\mathbf{x}_j, z_j)|\right). \tag{13}$$

Definitions (12) and (13) allow for negative values of $\hat{\gamma}_r$ and $\hat{\gamma}_m$, which could arise with a very poorly calibrated model, and would correctly give estimates less than 0.5.

If $r(\mathbf{x})$ is the linear predictor from fitting a Cox model to the data, then recalibration as proposed above is pointless: if done, it yields $\hat{\gamma}_m = \hat{\gamma}_z = 1$. However, the values of γ_r and γ_m in (12) and (13) could instead be estimated by shrinkage methods (Copas, 1983; van Houwelingen and Le Cessie, 1990).

4.2 Adjusted *D*

We define covariate-adjusted D by extending estimand (5) proposed in Section 3. First, z-specific D is

$$D(z) = \mathrm{E}\left[r^*(\mathbf{x}_i) - r^*(\mathbf{x}_j) \mid r(\mathbf{x}_i) > \hat{r}(z) > r(\mathbf{x}_j), z_i = z_j = z\right]$$

recalling that $\hat{r}(z)$ is the z-specific mean of the $r(\mathbf{x}_i)$. We can also write $D(z)$ as

$$\mathrm{E}\left[m^*(\mathbf{x}_i, z_i) - m^*(\mathbf{x}_j, z_j) \mid m(\mathbf{x}_i, z_i) > 0 > m(\mathbf{x}_j, z_j), z_i = z_j = z\right]$$

and so it is natural to define adjusted D as

$$D^{adj} = \mathrm{E}\left[m^*(\mathbf{x}_i, z_i) - m^*(\mathbf{x}_j, z_j) \mid m(\mathbf{x}_i, z_i) > 0 > m(\mathbf{x}_j, z_j), z_i = z_j\right].$$

That is, D^{adj} is the average log hazard ratio between individuals matched on z who are above-average and below-average for their value of z.

We propose the following modification to the estimation algorithm for D^{adj} given in Section 3.3. In stage 1, instead of ranking the $r(\mathbf{x}_i)$, we rank the $m(\mathbf{x}_i, z_i)$ across the whole sample, form normal scores, and scale by $\sqrt{\pi/8}$. In stage 2, the proportional hazards regression on the scaled normal scores is adjusted for z to avoid bias from omitting a prognostic covariate (Ford et al., 1995). $\hat{D}^{adj}$ is the coefficient of the scaled normal scores in the stage 2 model.

4.3 Stratification

A stratified version of C^{adj} may be computed by restricting attention to pairs within strata. For stratified versions of C^{adj*} and D^{adj} we replace $m(\mathbf{x}_i, z_i)$ and $\hat{r}(z_i)$ above with $m(\mathbf{x}_i, z_i, s_i)$ and $\hat{r}(z_i, s_i)$ where s_i is the stratum of individual i, $\hat{r}(z_i, s_i) = \mathrm{E}[r(\mathbf{x}_i)|z_i, s_i]$ and $r(\mathbf{x}_i, s_i) = m(\mathbf{x}_i, z_i, s_i) + \hat{r}(z_i, s_i)$. This decomposition is performed by regressing $r(\mathbf{x}_i, s_i)$ on z_i within strata. In estimating D^{adj}, the stage 2 proportional hazards regression is stratified by s.

5 Simulation study

We next explore the performance (bias and precision) of the proposed estimators as we vary the strengths of association of the outcome with $\mathbf{x}$ and z. We first consider an ideal setting where $r(\mathbf{x})$ is the linear predictor from a correctly specified Cox model, and $m(\mathbf{x}_i, z_i)$ is independent of z_i so that $C^{adj} = C^{adj*}$. We then consider a nonideal setting where $\mathrm{var}(m(\mathbf{x}_i, z_i))$ depends on z_i, so that estimands C^{adj}, $C^{adj,w}$, and C^{adj*} potentially differ.

5.1 Data generating model

Data sets of size $n = 1000$ were generated with covariates $\mathbf{x} = (v, z)$. This relatively large sample size was designed to make optimism negligible without excessively increasing computing time for C (which is roughly proportional to n^2).

Covariate z represents age at baseline. A fraction $1 - \phi$ of individuals belong to age group 1 and have $z \sim U(40, 50)$, the uniform distribution from 40 to 50. The remaining fraction ϕ of individuals belong to age group 2 and have $z \sim U(50, 60)$. Covariate v represents the biomarker of interest and was drawn as $v = \alpha(z - 50) + u$ with $u \sim N(0, \sigma_g^2)$ for an individual whose value of z places them in age group g. Settings for σ_g are given below. We chose α to make $\mathrm{corr}(v, z) = 0.25$: changing $\mathrm{corr}(v, z)$ to 0 or 0.5 affected the unadjusted results for both C and D, but had negligible effect on the adjusted results (results not shown).

Survival times were drawn from the Gompertz distribution

$$h(t) = h_0 \exp\left\{\beta_v v + \beta_z(z - 50 + t)\right\}$$

where t is time in years from baseline. We took $\beta_v = 0, 0.5, 1$, and $\beta_z = 0, 0.1, 0.2$. Follow-up was for 15 years, and h_0 was chosen to give 50–70% censoring. With this data generating model, $r(\mathbf{x}) = \beta_v v + \beta_z(z - 50)$, $\hat{r}(z) = (\beta_v\alpha + \beta_z)(z - 50)$ and $m(\mathbf{x}, z) = \beta_v[v - \alpha(z - 50)]$.

In simulation 1, we take $\phi = 0.5$, so that $z \sim U(40, 60)$ and weighting does not affect the estimands. We also take $\sigma_1 = \sigma_2 = 1$, so that $\hat{r}(z)$ and $m(\mathbf{x}, z)$ are independent, and the estimands C^{adj} and C^{adj*} are equal. In simulation 2, we take $\phi = 2/3$, in order to explore weighting, and $\sigma_1 = 1$ and $\sigma_2 = 2$, so that $\mathrm{var}(m(\mathbf{x}, z)|z)$ depends on z and the estimands differ.

5.2 Methods considered

For C, the unadjusted methods considered were Harrell's $\hat{C}_{Har}$ ("Harrell") and Gonen and Heller's $\hat{C}_{ind}$ ("indirect"). The adjusted methods considered were $\hat{C}^{adj}_{smooth1}$ ("smooth 1 unweighted") and $\hat{C}^{adj}_{smooth2}$ ("smooth 2 unweighted") with weights $w_1(z_i, z_j)$; $\hat{C}^{adj,w}_{smooth1}$ ("smooth 1 weighted") and $\hat{C}^{adj,w}_{smooth2}$ ("smooth 2 weighted") with weights $w_1(z_i, z_j)w_2(z_i, z_j)$; and $\hat{C}^{adj*}_{ind}$ ("indirect"). Weight $w_1(z_i, z_j)$ was computed using (8) with $\lambda = 3$, and $w_2(z_i, z_j)$ was computed using (9) and estimating $\hat{f}(z)$ in one-unit bins for z. Estimation was restricted to informative pairs without the weighting of Uno et al. (2011).

For D, we used unadjusted $\hat{D}$ and adjusted $\hat{D}^{adj}$. The methods are summarized in Table 2 and the top half of Table 3.

Table 2 Summary of methods for unadjusted measures of discrimination.

	C-index		D-index
	Harrell's C	*Indirect*	
Notation	$\hat{C}_{Har}$	$\hat{C}_{ind}$	$\hat{D}$
Description	Mean concordance among informative pairs	Mean expected concordance[a)]	Cox model on scaled rankit of risk score
Quantity estimated	C	C	D
Recalibration	No impact	Needed	Implicit in method

[a)] Expected concordance is computed assuming that $r(\mathbf{x})$ is the linear predictor in a correctly specified Cox model.

5.3 Simulation scheme

A total of 1000 datasets were drawn for each combination of parameters. For each combination of parameters and each method, we computed $\bar{C}$ and s_C, the mean and standard deviation of $\hat{C}$, and we present a forest-type plot showing $\bar{C}$ with an interval constructed as $\bar{C} \pm 1.96 s_C$. We computed the true values of C and D using the exact methods described in appendices C and D. Source code to reproduce the results is available as Supporting Information on the journal's web page (http://onlinelibrary.wiley.com/doi/10.1002/bimj.201400061/suppinfo).

5.4 Results for simulation 1

Figure 2 shows results with $\text{corr}(v, z) = 0.25$: the five panels show different combinations of β_v and β_z.

Considering the unadjusted results, $\hat{C}_{Har}$ has a slightly larger mean than $\hat{C}_{ind}$ in all panels, indicating small bias due to censoring.

Comparing the unadjusted and adjusted estimates, we see that they are similar in the second panel where $\beta_z = 0$ (i.e., where there is no covariate effect to adjust for), but markedly different in the other panels where $\beta_z > 0$.

We now compare the different adjustment methods in the first panel where $\beta_v = 0$ so that the true value is $C^{adj} = 0.5$. The "smooth 1" methods (unweighted and weighted) show substantial bias. This is likely to have arisen because concordance C_{ij} is strongly related to $\hat{r}(z_j) - \hat{r}(z_i)$ and the smoothing method inadequately allows for this association. The other adjusted methods show small positive bias. This is attributable to optimism, since the models are fitted and evaluated in the same data; however, optimism is small because of our large sample size. We therefore suggest that "smooth 2" may be preferable to "smooth 1".

In the other panels where $C^{adj} > 0.5$, the indirect estimator appeared unbiased (suggesting the bias from optimism is negligible), while the smoothing estimators had small positive bias, presumably due to censoring.

Results for adjusted D similarly suggest small optimism when $\beta_v = 0$ and little or no bias elsewhere (Fig. 3).

5.5 Results for simulation 2

Results for C are shown in Fig. 4. When $\beta_v = 0$ (top panel), the results are very similar to simulation 1. In the other panels, the estimands C^{adj} (estimated by the unweighted methods), $C^{adj,w}$ (estimated by the weighted methods), and C^{adj*} (estimated by the indirect method) are unequal and are shown by three vertical lines. These estimands differ by up to 0.014, with $C^{adj,w} < C^{adj*} < C^{adj}$.

Table 3 Summary of methods for adjusted measures of discrimination.

	C-index			Adjusted
	Smooth 1	Smooth 2	Indirect	D-index
Notation	$\hat{C}^{adj}_{smooth1}$, $\hat{C}^{adj,w}_{smooth1}$	$\hat{C}^{adj}_{smooth2}$, $\hat{C}^{adj,w}_{smooth2}$	$\hat{C}^{adj}_{ind}$	$\hat{D}^{adj}$
Description	Mean concordance (Smooth 1) or *m*-concordance (Smooth 2) among informative pairs with similar values of adjustment variable	Mean expected concordance[a)] after removing difference in adjustment variable	Covariate-adjusted Cox model on scaled rankit of risk score	
Options	Choice of weights; choice of smoothing parameter λ	–	–	
Quantity estimated	C^{adj} or C^{adjw}	C^{adj*}	D^{adj}	
Recalibration	Little impact	Little impact	Needed	Implicit in method
Properties				
Must lie between 0 and 1	✓	✓	✓	N/A
Has value 1 if all pairs are concordant	✓	✓	×	N/A
Has value 0 if all pairs are discordant	✓	✓	×	N/A
Reduces to unadjusted estimate if there is no Z	✓	✓	✓	✓
Direct – unaffected by miscalibration	✓	✓	×	(×)
Free of tuning parameter λ	×	×	✓	✓
Fast to compute	×	×	×	✓
Unaffected by optimism	×	×	×	×
Unaffected by censoring	×	×	✓	✓
Unbiased in simulation after accounting for optimism and censoring	×	✓	✓	✓

[a)] Expected concordance is computed assuming that $r(\mathbf{x})$ is the linear predictor in a correctly specified Cox model. (×) means only slightly affected.

$\hat{C}^{adj}_{ind}$ remains unbiased for C^{adj*}. The smoothing estimators are all positively biased, with weighted estimators on average slightly smaller than unweighted estimators. In the second panel, where unadjusted and adjusted C-indices are equal, the bias in the smoothing estimators is slightly smaller than that in Harrell's estimator: this suggests that all the bias observed is attributable to censoring.

Corresponding results for D are shown in Fig. 5. Small positive bias is found for all parameter values. Because no bias was seen in simulation 1, this is likely to arise from model mis-specification.

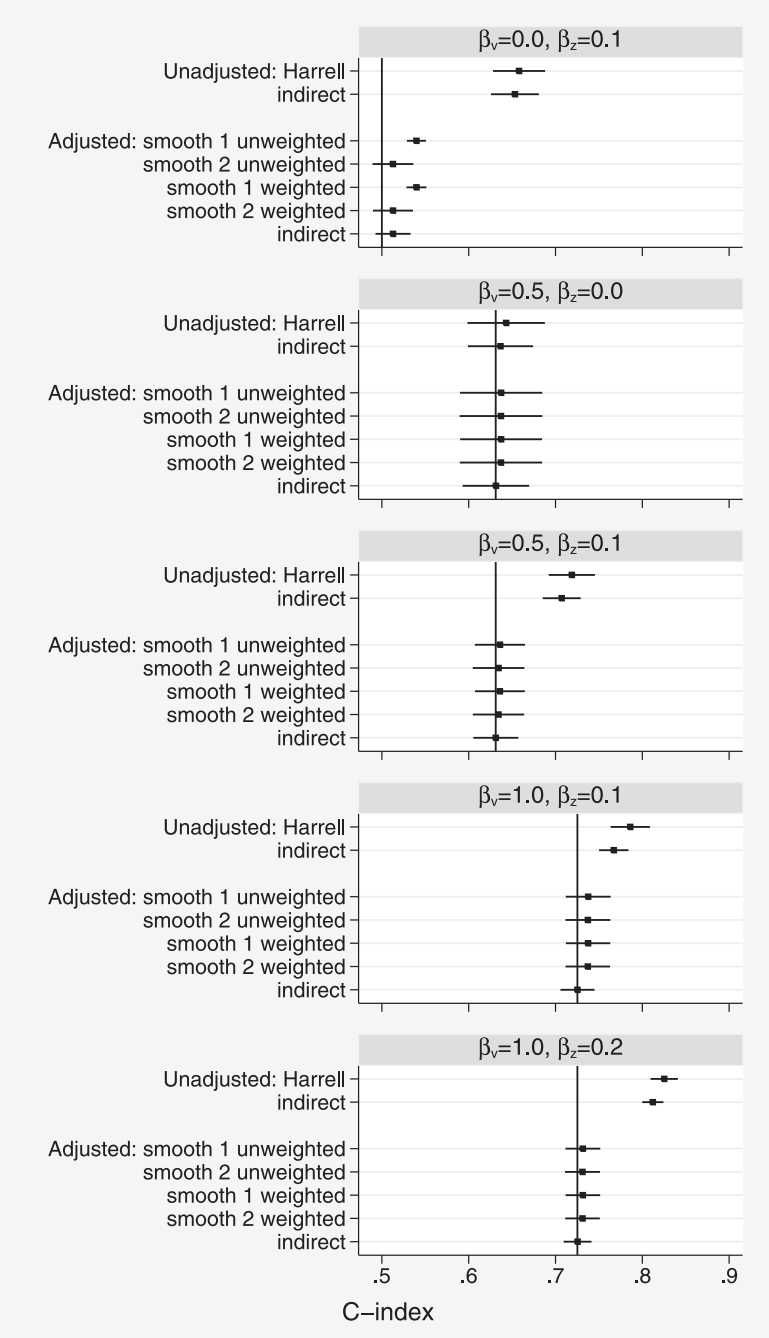

Figure 2 Simulation study 1: comparison of various unadjusted and adjusted estimates of C. Intervals show $\bar{C} \pm 1.96 s_C$ where $\bar{C}$ and s_C are the mean and standard deviation of $\hat{C}$. Vertical lines indicate the true value of adjusted C. Panels show simulated data with different values of β_v and β_z, but all have $\mathrm{corr}(v, z) = 0.25$ (see text).

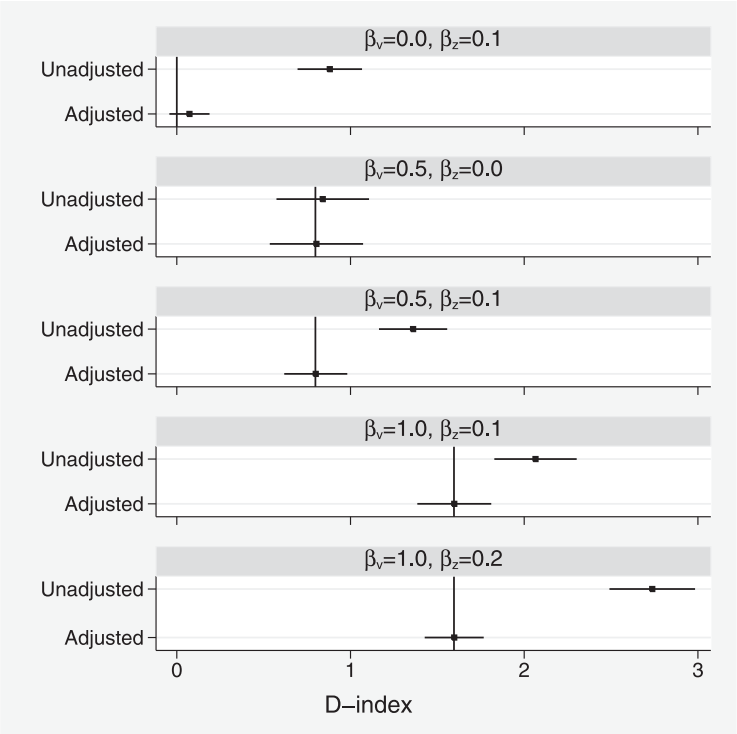

Figure 3 Simulation study 1: comparison of various unadjusted and adjusted estimates of D. Intervals show $\bar{D} \pm 1.96 s_D$. Vertical lines indicate the true value of adjusted D.

6 ERFC results

To illustrate the differences between methods and the effects of recalibration, we used the single Cox proportional hazards model fitted to all the ERFC studies, stratifying by study, sex, and trial arm, as displayed in Table 1. The linear predictor from the resulting model was evaluated using unweighted methods in each study separately, stratifying by sex and trial arm.

We first explore the effect of recalibration. The model is guaranteed to be well calibrated in the whole ERFC data, but it is likely to be miscalibrated in individual studies. Figure 6 plots, for each method considered, the difference between the C-indices after recalibration and before recalibration against their mean, as proposed by Bland and Altman (1986). Harrell's method is unaffected by recalibration and so is not shown in Fig. 6. We see that recalibration has a large impact for the indirect methods and very little impact for the smoothing methods.

We next compare the different methods after recalibration. Figure 7 shows Bland-Altman plots comparing the two unadjusted methods and the three adjusted methods. The top panel shows that the indirect method tends to give lower results than Harrell's method, probably due to censoring (Gonen and Heller, 2005). The four panels in the lower left-hand corner compare unadjusted and adjusted methods and show large differences. The three panels in the lower right-hand corner compare the adjusted methods. Again, the indirect method tends to give lower estimates than the other methods, while the two smoothing methods give very similar results.

Finally, we revisit Fig. 1, which used the indirect method with recalibration. The strong association of the unadjusted C-index with SD of baseline age (left-hand panel) is removed when we use the age-adjusted C-index (right-hand panel). Covariate adjustment reduces C by up to 0.21 in 77 of the 82

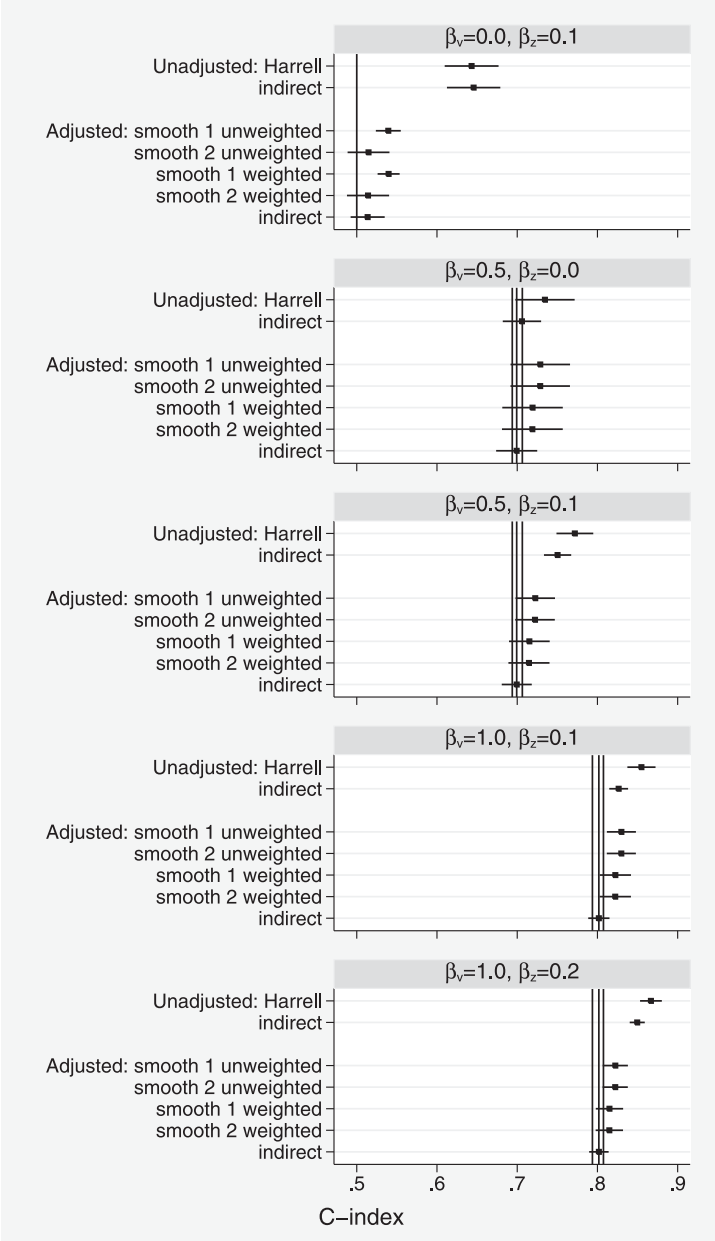

Figure 4 Simulation study 2: comparison of various unadjusted and adjusted estimates of C. Vertical lines indicate (from L to R) true values of $C^{adj,w}$, C^{adj*}, and C^{adj}.

studies; increases C by up to 0.03 in four studies (all of which are small); and leaves C unchanged for one study where all participants have the same baseline age. Figure 8 shows the corresponding results for the D-index.

Source code to analyze simulated data (like one ERFC cohort) is available as Supporting Information on the journal's web page (http://onlinelibrary.wiley.com/doi/10.1002/bimj.201400061/suppinfo).

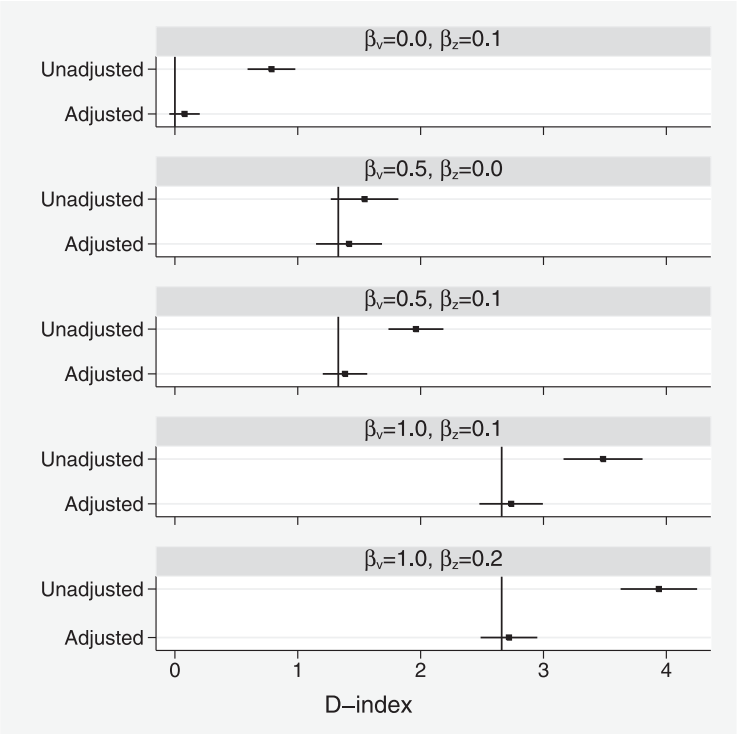

Figure 5 Simulation study 2: comparison of various unadjusted and adjusted estimates of *D*. Vertical lines indicate the true value of adjusted *D*.

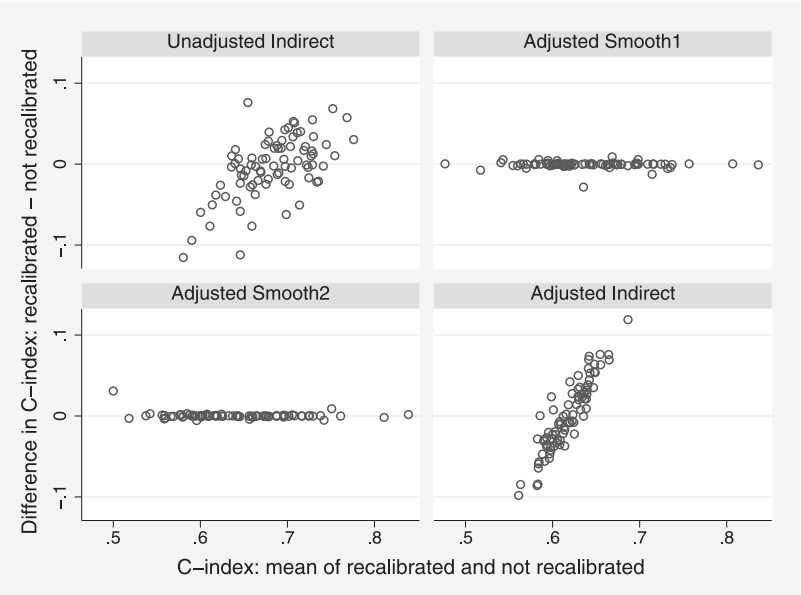

Figure 6 ERFC data: Bland-Altman plots exploring the effects of recalibration on various methods for computing the adjusted C-index. Each point represents one study.

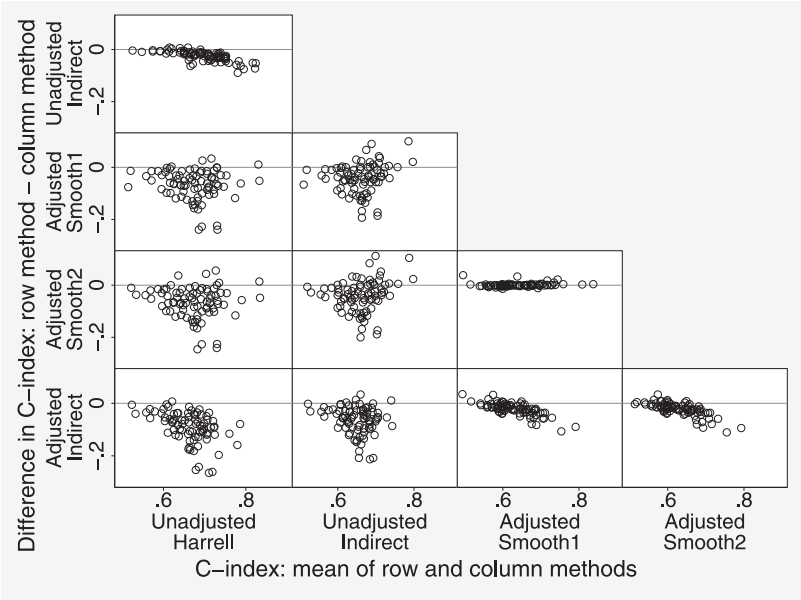

Figure 7 ERFC data: Bland-Altman plots comparing different methods for computing the adjusted C-index, after recalibration. Each point represents one study.

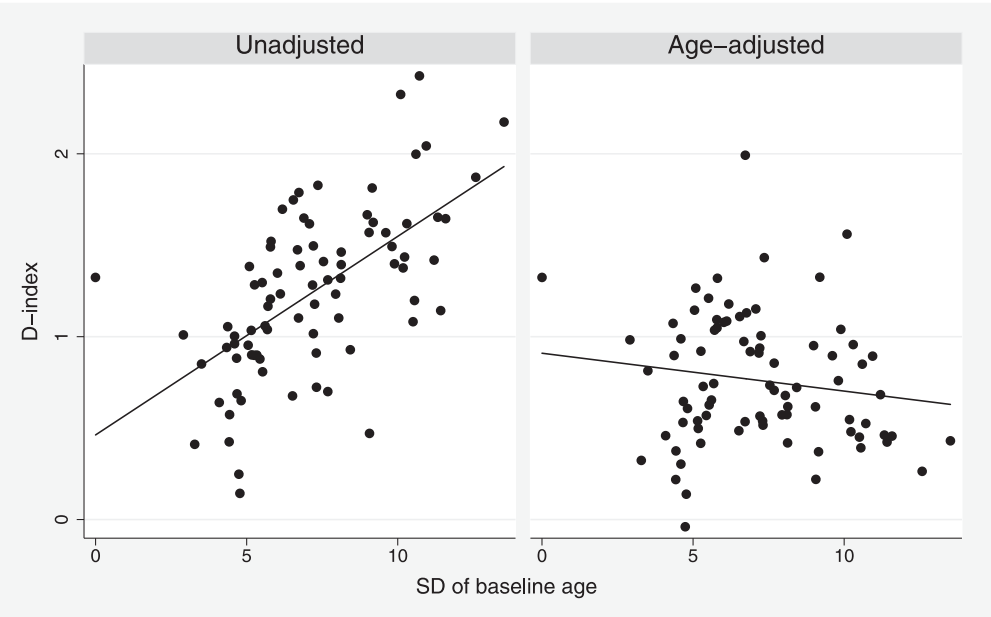

Figure 8 ERFC data: unadjusted and age-adjusted D-index for a model including baseline age, smoking, SBP, TC, and HDL, plotted for each study against the SD of baseline age in that study. Analyses are stratified by sex and trial arm. Each point represents one study.

7 Discussion

We have proposed a number of methods for estimating an adjusted C-index. The lower part of Table 3 lists a number of desirable properties of an adjusted measure of discrimination, and evaluates the proposed adjusted C-indices and the adjusted D-index against these measures. Overall, the best adjusted C-index appears to be the indirect estimator, although we caution that it is sensitive to model mis-specification. The adjusted D-index is an excellent alternative which is easy to compute.

We have not discussed computation of standard errors in this paper: for the C-index, bootstrapping could be used, while a standard error arises naturally in the calculation of the D-index.

Optimism (overfitting) is an issue whenever models are estimated and evaluated on the same dataset (Harrell et al., 1996). It is not the focus of this paper, because the ERFC data set is large and optimism is likely to be negligible. However, there were signs of optimism in the simulation studies with $\beta_v = 0$. In general, our methods should be applied after recalibrating the risk score $r(\mathbf{x})$ to allow for optimism, ideally in an external validation set, or otherwise by using internal corrections such as the bootstrap (Harrell et al., 1996).

Censoring is another potential problem for our methods. It causes bias in direct methods if noninformative pairs are simply excluded. Our simulation results show that moderate biases due to censoring do occur, especially in larger C-indices (e.g., in the bottom two panels of Fig. 4). Typically, studies have both a limit τ to the length of follow-up and random censoring before that time. The method of Uno et al. (2011) can be used to correct for the random censoring, but it estimates C^τ not C. Currently the only way to estimate C with data censored by end of follow-up is the indirect method.

Model mis-specification is a further potential problem, especially with the indirect methods which assume a correctly specified proportional hazards model. We have proposed a recalibration step which should remove bias due to miscalibration, but not necessarily other forms of model mis-specification. The direct methods such as Smooth 1 and Smooth 2 should be much less sensitive to model mis-specification, since they depend on observed concordance, not model-predicted concordance.

To reduce sensitivity to model mis-specification, we tried using the difference between the direct and indirect estimators of C (which may reflect the impact of model mis-specification) to correct the indirect estimator of C^{adj*}, defining $\hat{C}^{adj*}_{corr1} = \hat{C}_{Har} - \hat{C}_{ind} + \hat{C}^{adj*}_{ind}$ (or the equivalent on the logit scale). However, because the impact of censoring is greater in unadjusted than adjusted estimators (Fig. 2), the corrected estimator did not perform well in the simulation study and was not included in the results.

Covariate adjustment could be considered for other measures of discrimination. The net reclassification index (NRI) is a popular measure of the difference in discrimination between two models (Pencina et al., 2008). Because the NRI is based on within-individual comparisons, it is neither necessary nor possible to adjust it for covariates, although a covariate-specific NRI could be a useful quantity. However, correct calibration is required to avoid misleading NRI results (Hilden and Gerds, 2014). Another way to evaluate the value of adding a new biomarker to a risk prediction model could be to evaluate the discrimination of the new model, adjusting for the covariates in the original model.

Further extensions include ways to account for competing risks and to obtain time-dependent measures of discrimination (Wolbers et al., 2009). A common feature of these approaches is that the C-index needs to be computed for different time points which limit the comparability of different studies that have different length of follow-up. An open question is which metric is more appropriate for which data and whether these different approaches can produce different conclusions in some scenarios. The methodology presented here could be extended to incorporate these extensions.

Our approach should not be confused with ROC regression (Tosteson and Begg, 1988). ROC regression methods model the accuracy of a diagnostic test as a function of covariates, not how the disease is associated with covariates. A possible use of such an approach might be to find subgroups where the marker should not be used or to find optimal cutoffs. Here we assume predictions that have been optimized with respect to disease risk and our aim is not (primarily) to explain which covariates

affect accuracy but to adjust discrimination statistics for the confounding effect of the covariates that do.

In summary, we have proposed covariate-adjusted measures of concordance. There are many benefits to such measures (Pepe et al., 2008). In the meta-analysis setting, they facilitate comparisons between studies with different covariate distributions (Figs. 1 and 8). They also enable matched case-control studies nested within cohort studies to be compared with standard cohort studies, since the former can only yield measures of discrimination adjusted for the matching variables. We advocate adjustment, at least for study design variables such as age, sex, and study centre, whenever measures of discrimination are to be compared between studies with different distributions of the design variables.

Acknowledgments This work was supported by the Medical Research Council Grant G0700463 and Unit Programme U105260558. Funding sources for the ERFC studies, and full study names, are listed at http://www.phpc.cam.ac.uk/ceu/research/erfc/studies/. The authors thank the staff and participants of all the participating studies for their important contributions. We also thank Patrick Royston and Shaun Seaman for helpful discussions.

Conflict of interest
The authors have declared no conflict of interest.

Appendix

A Members of the Emerging Risk Factors Collaboration

Emerging Risk Factors Collaboration investigators/contributors: **AFTCAPS:** R W Tipping; **ALLHAT:** B R Davis, L M Simpson; **ARIC:** C M Ballantyne, A R Folsom, J Coresh; **AUSDIAB:** J E Shaw, R Atkins, P Z Zimmet, E L M Barr; **BHS:** M W Knuiman; **BRHS:** S G Wannamethee, R W Morris; **BRUN:** J Willeit, P Willeit, P Santer, S Kiechl; **BUPA:** N Wald; **BWHHS:** S Ebrahim, D A Lawlor; **CAPS:** J Gallacher, J W G Yarnell, Y Ben-Shlomo; **CASTEL:** E Casiglia, V Tikhonoff; CHARL; S E Sutherland, P J Nietert, J E Keil, D L Bachman; **CHS:** B M Psaty, M Cushman, see http://www.chs-nhlbi.org for acknowledgments; **COPEN:** B G Nordestgaard, A Tybjærg-Hansen, R Frikke-Schmidt; **CUORE:** S Giampaoli, L Palmieri, S Panico, L Pilotto, D Vanuzzo; **DUBBO:** L A Simons, Y Friedlander, J McCallum; **EAS:** J F Price, S McLachlan; **EPESEBOS:** J O Taylor, J M Guralnik; **EPESEIOW:** R B Wallace, F J Kohout, J C Cornoni-Huntley, J M Guralnik; **EPESENCA:** D G Blazer, J M Guralnik, C L Phillips; **EPESENHA:** C L Phillips, J M Guralnik; **EPICNOR:** N J Wareham, K-T Khaw; **ESTHER:** H Brenner, B Schöttker, H T Müller, D Rothenbacher; **FINE_FIN:** A Nissinen; **FINE_IT:** C Donfrancesco, S Giampaoli; **FINRISK92, FINRISK97:** K Harald, P R Jousilahti, E Vartiainen, V Salomaa; **FRAMOFF:** R B D'Agostino Sr., P A Wolf, R S Vasan; **FUNAGATA:** M Daimon, T Oizumi, T Kayama, T Kato; **GOH:** A Chetrit, R Dankner, F Lubin; **GOTO43:** L Welin, K Svärdsudd, H Eriksson, G Lappas; **GOTOW:** L Lissner, K Mehlig, C Björkelund; **GRIPS:** D Nagel; **HISAYAMA:** Y Kiyohara, H Arima, T Ninomiya, J Hata; **HONOL:** B Rodriguez; **HOORN:** J M Dekker, G Nijpels, C D A Stehouwer; **IKNS:** H Iso, A Kitamura, K Yamagishi, H Noda; **ISRAEL:** U Goldbourt; **KIHD:** J Kauhanen, J T Salonen, T-P Tuomainen; **LEADER:** T W Meade, B L DeStavola; **MCVDRFP:** A Blokstra, W M M Verschuren; **MESA:** M Cushman, I H de Boer, A R Folsom, B M Psaty, see http://www.mesa-nhlbi.org for acknowledgements; **MOGERAUG1, MOGERAUG2:** W Koenig, C Meisinger, A Peters; **MORGEN:** W M M Verschuren, H B Bueno-de-Mesquita, A Blokstra; **MOSWEGOT:** A Rosengren, L Wilhelmsen, G Lappas; **MRFIT:** L H Kuller, G Grandits; **NHANESIII**; **NPHSII:** J A Cooper, K A Bauer; **NSHS:** K W Davidson, S Kirkland, J A Shaffer, D Shimbo; **OSAKA:** A Kitamura, H Iso, S Sato; **PREVEND:** R P F Dullaart, S J L Bakker, R T Gansevoort; **PRIME:** P Ducimetiere, P Amouyel, D Arveiler, A Evans, J Ferrières; **PROCAM:** H

Schulte, G Assmann; PROSPER; J W Jukema, R G J Westendorp, N Sattar; **QUEBEC:** B Cantin, B Lamarche, J-P Després; **RANCHO:** E Barrett-Connor, D L Wingard, L B Daniels; **REYK:** V Gudnason, T Aspelund; **RIFLE:** M Trevisan; **ROTT:** A Hofman, O H Franco; **SHHEC:** H Tunstall-Pedoe, R Tavendale, G D O Lowe, M Woodward; **SHS:** W J Howard, B V Howard, Y Zhang, L G Best, J Umans; **SPEED:** Y Ben-Shlomo, G Davey-Smith; **TARFS:** A Onat; **TOYAMA:** H Nakagawa, M Sakurai, K Nakamura, Y Morikawa; **TROMSO:** I Njølstad, E B Mathiesen, T Wilsgaard; **ULSAM:** J Sundström; **USPHS2:** J M Gaziano, P M Ridker; **WHITE1:** M Marmot, R Clarke, R Collins, A Fletcher; **WHITE2:** E Brunner, M Shipley, M Kivimaki; **WHS:** P M Ridker, J Buring, N Rifai, N Cook; **WOSCOPS:** I Ford, M Robertson; **ZARAGOZA:** A Marín Ibañez; **ZUTE:** E J M Feskens, J M Geleijnse.

Coordinating Centre: T Bolton, S Burgess, A S Butterworth, E di Angelantonio, P Gao, E Harshfield, S Kaptoge, L Pennells, S Peters, S Spackman, S Thompson, M Walker, I White, P Willeit, A Wood, J Danesh (Principal Investigator).

B Comparison of estimands

We give an example where $\mathrm{var}(V|Z=z)$ does not depend on z and yet the estimands C^{adj*} and C^{adj} are unequal.

Suppose $\mathbf{X}=(V,Z)$ where V and Z are binary with $p(Z=1)=0.5$, $p(V=1|Z=0)=\pi_0=0.2$ and $p(V=1|Z=1)=\pi_1=0.8$. Suppose $h_i(t)=h_0(t)\exp(V+Z)$ so $r(\mathbf{X})=V+Z$. We want to adjust for Z so $m(\mathbf{X},Z)=V-\mathrm{E}[V|Z]$. When $Z_i=Z_j=z$, $|m(\mathbf{X}_i,Z_i)-m(\mathbf{X}_j,Z_j)|$ is 1 with probability $2\pi_z(1-\pi_z)$ and 0 otherwise, and this distribution is independent of z since $\pi_1=1-\pi_0$. So $C(z=0)=C(z=1)=C^{adj}=0.574$. But for pairs with $Z_i\neq Z_j$, $|m(\mathbf{X}_i,Z_i)-m(\mathbf{X}_j,Z_j)|$ has a different distribution (but the same SD). As a result, $C^{adj*}=0.598\neq C^{adj}$.

C True values of $C(z)$, C^{adj}, and C^{adj*} in the simulation study

For simplicity we assume $\beta_v\geq 0$. We also assume that the model is correctly specified. We first compute $C(z)$, for which the distribution of z is not needed. Define $I(\sigma^2)=\mathrm{E}[\mathrm{expit}\{|A|\}]$ when $A\sim N(0,\sigma^2)$. Then

$$I(\sigma^2)=2\int_0^\infty \mathrm{expit}(\sigma u)\,\phi(u)du$$

where $\phi(u)$ is the standard normal density function.

For two individuals $r=1,2$ with $z_1=z_2=z$, we can write $v_r=\alpha(z-50)+u_r$, so the probability that they are concordant is $\mathrm{expit}(|\beta_v(v_2-v_1)|)=\mathrm{expit}(|\beta_v(u_2-u_1)|)$. If z is in age group g then $u_r\sim N(0,\sigma_g^2)$ so $\mathrm{var}(\beta_v(u_2-u_1))=2\beta_v^2\sigma_g^2$ and

$$C(z)=I(2\beta_v^2\sigma_g^2). \tag{A.1}$$

For simulation 1, we evaluate (A.1) with $\sigma_g=1$. We then have $C^{adj}=C^{adj*}=C(z)$ for all z.

For simulation 2, $\sigma_1=1$ and $\sigma_2=2$. Denote the corresponding values of $C(z)$ from (A.1) as $C_1=I(2\beta_v^2\sigma_1^2)$ and $C_2=I(2\beta_v^2\sigma_2^2)$. We can then derive $C^{adj}=\{(1-\phi)^2C_1+\phi^2C_2\}/\{(1-\phi)^2+\phi^2\}$ and $C^{adj,w}=(1-\phi)C_1+\phi C_2$. Finally, we can write C^{adj*} as a weighted sum of $\mathrm{E}[\mathrm{expit}\{\beta_v|u_1-u_2|\}]$ terms over the possible groups for individuals 1 and 2:

$$C^{adj*}=(1-\phi)^2I(2\beta_v^2\sigma_1^2)+2\phi(1-\phi)I(\beta_v^2(\sigma_1^2+\sigma_2^2))+(1-\phi)^2I(2\beta_v^2\sigma_2^2).$$

D True value of D^{adj}

We have $D = \mathrm{E}[r(\mathbf{x}_i) - r(\mathbf{x}_j)|z_i = z_j, m(\mathbf{x}_i, z_i) < 0 < m(\mathbf{x}_j, z_j)]$. We again assume that the model is correctly specified. Since $m(\mathbf{x}_i, z_i) = \beta_v u_i$, $D = \beta_v\{\mathrm{E}[u_i|u_i > 0] - \mathrm{E}[u_i|u_i < 0]\}$. In simulation 1, $\mathrm{E}[u_i|u_i > 0] = \sqrt{2/\pi}$. In simulation 2, $\mathrm{E}[u_i|u_i > 0] = \sqrt{2/\pi}\{(1-\phi)\sigma_1 + \phi\sigma_2\} = \frac{5}{3}\sqrt{\frac{2}{\pi}}$. In both cases, $\mathrm{E}[u_i|u_i < 0] = -\mathrm{E}[u_i|u_i > 0]$. Hence $D^{adj} = \beta_v\sqrt{8/\pi}$ in simulation 1 and $D^{adj} = \beta_v\frac{5}{3}\sqrt{8/\pi}$ in simulation 2.

References

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between 2 methods of clinical measurement. *Lancet* **327**, 307–310.

Chambless, L. E. and Diao, G. (2006). Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine* **25**, 3474–3486.

Choodari-Oskooei, B., Royston, P. and Parmar, M. K. B. (2012a). A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Statistics in Medicine* **31**, 2627–2643.

Choodari-Oskooei, B., Royston, P. and Parmar, M. K. B. (2012b). A simulation study of predictive ability measures in a survival model II: explained randomness and predictive accuracy. *Statistics in Medicine* **31**, 2644–2659.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)* **45**, 311–354.

Emerging Risk Factors Collaboration (2007). The Emerging Risk Factors Collaboration: analysis of individual data on lipid, inflammatory and other markers in over 1.1 million participants in 104 prospective studies of cardiovascular diseases. *European Journal of Epidemiology* **22**, 839–869.

Fibrinogen Studies Collaboration (2009). Measures to assess the prognostic ability of the stratified Cox proportional hazards model. *Statistics in Medicine* **28**, 389–411. Writing committee: L Pennells, IR White, AM Wood, S Kaptoge, N Sarwar.

Ford, I., Norrie, J. and Ahmadi, S. (1995). Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine* **14**, 735–746.

Gonen, M. and Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* **92**(4), 965–970.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A. (1982). Evaluating the Yield of Medical Tests. *Journal of the American Medical Association* **247**, 2543–2546.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361–387.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105.

Hilden, J. and Gerds, T. A. (2014). A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in Medicine* **33**, 3405–3414.

Hlatky, M. A., Greenland, P., Arnett, D. K., Ballantyne, C. M., Criqui, M. H., Elkind, M. S. V., Go, A. S., Harrell, F. E., Hong, Y., Howard, B. V., Howard, V. J., Hsue, P. Y., Kramer, C. M., McConnell, J. P., Normand, S.-L. T., O'Donnell, C. J., Smith, S. C. J. and Wilson, P. W. F. on behalf of the American Heart Association Expert Panel on Subclinical Atherosclerotic Diseases and Emerging Risk Factors and the Stroke Council (2009). Criteria for Evaluation of Novel Markers of Cardiovascular Risk: A Scientific Statement From the American Heart Association. *Circulation* **119**, 2408–2416.

van Houwelingen, J. C. and Le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine* **9**, 1303–1325.

Janes, H. and Pepe, M. S. (2008). Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. *American Journal of Epidemiology* **168**, 89–97.

Janes, H., Longton, G. and Pepe, M. S. (2009). Accommodating covariates in receiver operating characteristic analysis. *Stata Journal* **9**, 17–39.

Kerr, K. and Pepe, M. (2011). Joint modeling, covariate adjustment, and interaction: contrasting notions in risk prediction models and risk prediction performance. *Epidemiology* **22**, 805–812.

Lijmer, J., Bossuyt, P. and Heisterkamp, S. (2002). Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Statistics in Medicine* **21**, 1525–1537.

Mallett, S., Royston, P., Waters, R., Dutton, S. and Altman, D. (2010). Reporting performance of prognostic models in cancer: a review. *BMC Medicine* **8**, 21.

Mihaescu, R., van, Z., van, H., Sijbrands, E., Uitterlinden, A., Witteman, J., Hofman, A., Hunink, M., van Duijn, C. and Janssens, A. (2010). Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. *American Journal of Epidemiology* **172**, 353–361.

Pencina, M. J., D'Agostino, Sr R. B., D'Agostino, Jr R. B. and Vasan, R. S. (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.

Pennells, L., Kaptoge, S., White, I. R., Thompson, S. G., Wood, A. M. and The Emerging Risk Factors Collaboration., (2014). Assessing risk prediction models using individual participant data from multiple studies. *American Journal of Epidemiology* **179**, 621–632.

Pepe, M., Janes, H., Longton, G., Leisenring, W. and Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* **159**, 882–890.

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**, 362–368.

Prospective Studies Collaboration, (2002). Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* **360**, 1903–1913.

Royston, P. and Sauerbrei, W. (2004). A new measure of prognostic separation in survival data. *Statistics in Medicine* **23**, 723–748.

Tosteson, A. and Begg, C. (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making* **8**, 204–215.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30**, 1105–1117.

Wolbers, M., Koller, M. T., Witteman, J. C. M. and Steyerberg, E. W. (2009). Prognostic models with competing risks: Methods and application to coronary risk prediction. *Epidemiology* **20**, 555–561.

