Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2017, Vol. 26(3) 1053–1077

Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280214567141
journals.sagepub.com/home/smm
SAGE

Predictive accuracy of novel risk factors and markers: A simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model

Peter C Austin,[1,2,3] Michael J Pencinca[4,5] and Ewout W Steyerberg[6]

Abstract

Predicting outcomes that occur over time is important in clinical, population health, and health services research. We compared changes in different measures of performance when a novel risk factor or marker was added to an existing Cox proportional hazards regression model. We performed Monte Carlo simulations for common measures of performance: concordance indices (*c*, including various extensions to survival outcomes), Royston's D index, R^2-type measures, and Chambless' adaptation of the integrated discrimination improvement to survival outcomes. We found that the increase in performance due to the inclusion of a risk factor tended to decrease as the performance of the reference model increased. Moreover, the increase in performance increased as the hazard ratio or the prevalence of a binary risk factor increased. Finally, for the concordance indices and R^2-type measures, the absolute increase in predictive accuracy due to the inclusion of a risk factor was greater when the observed event rate was higher (low censoring). Amongst the different concordance indices, Chambless and Diao's c-statistic exhibited the greatest increase in predictive accuracy when a novel risk factor was added to an existing model. Amongst the different R^2-type measures, O'Quigley et al.'s modification of Nagelkerke's R^2 index and Kent and O'Quigley's $\rho^2_{w,a}$ displayed the greatest sensitivity to the addition of a novel risk factor or marker. These methods were then applied to a cohort of 8635 patients hospitalized with heart failure to examine the added benefit of a point-based scoring system for predicting mortality after initial adjustment with patient age alone.

[1] Institute for Clinical Evaluative Sciences, Toronto, Canada
[2] Institute of Health Management, Policy and Evaluation, University of Toronto
[3] Schulich Heart Research Program, Sunnybrook Research Institute, Toronto, Canada
[4] Duke Clinical Research Institute, Duke University, Durham
[5] Department of Biostatistics and Bioinformatics, Duke University School of Medicine
[6] Department of Public Health, Erasmus MC – University Medical Center Rotterdam, Rotterdam, The Netherlands

Corresponding author:
Peter C Austin, Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario M4N 3M5, Canada.
Email: peter.austin@ices.on.ca

Keywords
Survival analysis, Cox proportional hazards model, discrimination, predictive accuracy, Monte Carlo simulations, predictive models, risk factors, model performance

1 Introduction

Predicting the occurrence of an adverse event or outcome over time is an important issue in clinical, population health, and health services research. Time-to-event outcomes occur frequently in the biomedical literature.[1] The Cox proportional hazards regression model is the method most frequently used to assess the effect of patient characteristics on the risk of the occurrence of a time-to-event outcome. Clinical prediction models allow clinicians to accurately assess patient prognosis and permit effective risk stratification of patients.

There is a growing interest in identifying novel risk factors or markers (e.g. genetic factors, biomarkers, lifestyle characteristics or patient characteristics) that add important prognostic information above and beyond that contained in conventional clinical prediction models. The importance of novel risk factors or markers is often quantified by the change in model performance when a novel risk factor is added to an existing risk prediction model. Several measures have been proposed to quantify the predictive performance of a Cox proportional hazards regression model. These include extensions of concordance or c-statistics, R^2-type coefficients, and Royston's D index. Furthermore, extensions of the integrated discrimination improvement (IDI) to survival models have been developed for quantifying the incremental increase in prediction accuracy due to the inclusion of novel risk factors in an existing model.[2]

In a recent paper, we examined the sensitivity of different measures of model performance to the inclusion of a novel risk factor or marker to an existing clinical prediction model when outcomes are binary.[3] Several articles have reported extensive comparisons of the performance and properties of different measures of model performance for predicting survival outcomes.[2,4–9] While the statistical properties of these performance measures have been extensively studied, there is a paucity of research comparing the sensitivity of these different performance measures for survival models to the inclusion of novel risk factors or markers. The objective of the current paper was to explore changes in popularly used measures of model performance for survival outcomes when a novel risk factor or marker is added to an existing Cox proportional hazards regression model. To address this objective, we performed an extensive set of simulations reflecting scenarios that reflect current epidemiological research on new risk factors. Of note, our focus was solely on the numerical impact of adding new predictors on the selected measures of model performance and not on selecting the most appropriate measures of model performance or investigating their limitations.

The paper is structured as follows: in Section 2 we describe different measures of predictive accuracy for use with survival models. In Section 3 we describe the design of an extensive set of Monte Carlo simulations to examine the effect of adding a novel risk factor to an existing reference model on different measures of model performance for Cox proportional hazards models. In Section 4 we report the results of these simulations. In Section 5, we present a case study illustrating the application of these different methods for assessing the added utility of a heart failure mortality point-scoring system in predicting the hazard of death in patients hospitalized with heart failure. Finally, in Section 6 we summarize our findings and place them in the context of the existing literature. In particular, we compare and contrast our findings on the effect of adding novel risk factors on different measures performance for use with survival outcomes with previously published work on the effect of adding novel risk factors on different performance measures for use with binary outcomes.

2 Measures of predictive accuracy for time-to-event analyses

In this section we briefly review different measures of model performance that have been proposed for use with time-to-event models: concordance (c) statistics, Royston's D-index, R^2-type measures, and Chambless' adaptation of the IDI for survival outcomes.

2.1 Notation and terminology

Let T^S denote the outcome variable denoting time to the occurrence of the event of interest, and Z denote a vector of p explanatory variables. Observed data for the ith subject consist of the following triple: (T_i, δ_i, Z_i) ($i = 1, \ldots, N$), where T is a variable denoting the observed event time: $T_i = \min(T_i^S, T_i^C)$, where T^C denotes the censoring time, which is independent of T^S conditional on Z, while δ_i denotes an indicator variable denoting whether the event was observed to occur for the ith subject (i.e. $\delta_i = I(T_i^S \leq T_i^C)$). Where necessary, let τ denote a specific event time.

The Cox proportional hazard regression model can be written as $\lambda(t|Z) = \lambda_0(t)\exp\{\beta' Z\}$, where $\lambda_0(t)$ denotes the baseline hazard function and β is a $p \times 1$ vector of regression parameters that are estimated from the fitted model. This model can also be written in terms of the density function: $f(t|Z;\beta) = h_0(t)\exp\left\{\beta Z - e^{\beta Z}\int_0^t h_0(u)du\right\}$ [4], which will be used for defining one of the R^2-type measures of predictive accuracy.

2.2 Concordance statistics for time-to-event outcomes

When outcomes are binary, the c-statistic is the probability that a randomly selected subject who experienced the outcome has a higher predicted probability of experiencing the outcome than a randomly selected subject who did not experience the outcome. It can be calculated by taking all possible pairs of subjects consisting of one subject who experienced the outcome of interest and one subject who did not experience the outcome. The c-statistic is the proportion of such pairs in which the subject who experienced the outcome had a higher predicted probability of experiencing the event than the subject who did not experience the outcome (i.e. out of all possible pairs in which one subject experiences the outcome and one subject does not experience the outcome, it is the proportion of pairs that are concordant).[10,11] Different adaptations of the c-statistic have been proposed for use with time-to-event outcomes in which censoring may occur.[8,9,12] Let P_i denote the model-based predicted probability of the occurrence of an event prior to time τ and let D_i be an event indicator at time τ (i.e. $D_i = 1$ if the ith subject experienced the event prior to time τ and $D_i = 0$ otherwise).

Chambless and Diao proposed a time-varying extension of the conventional c-statistic to survival outcomes that most closely reflects the definition in the setting with binary outcomes.[13] At a given time τ, one considers all possible pairs of subjects consisting of one subject who experienced the event prior to time τ, and one subject who had not experienced the event by time τ. Chambless and Diao's time-varying c-statistic is defined to be the proportion of all such pairs in which the subject who experienced the event prior to time τ had a greater predicted probability of experiencing the event prior to time τ compared to the subject who had not experienced the event prior to time τ.

We denote this approach by AUC(CD). Formally, $\mathrm{AUC(CD)}(\tau) = P(P_i > P_j\,|D_i(\tau) = 1, D_j(\tau) = 0)$. An estimator of this concordance index is $\mathrm{AUC(CD)}(\tau) = \frac{\mathrm{E}[(1-S(\tau|P_j))\cdot S(\tau|P_i)\cdot I(P_i < P_j)]}{\mathrm{E}[(1-S(\tau|P_i)]\cdot \mathrm{E}[S(\tau|P_i)]}$.[8]

The most commonly used extension of the c-statistic to survival analysis was proposed by Harrell et al.[14] and was extensively studied by Pencina and D'Agostino.[12] They formally defined $\mathrm{AUC(H)}(\tau) = \Pr(P_i > P_j\,|T_i^S < T_j^S, T_i^S < \tau)$, where T_i^S denotes the true survival time for the *i*th subject (and thus, due to censoring, may not be observable for all subjects). The estimator is based on concordant and discordant pairs. A pair of subjects is said to be concordant if the subject with the longer survival time also had the greater predicted survival time (alternatively, one can use the predicted probability of surviving until any fixed time point). A pair of subjects is said to be discordant if the subject with the longer survival time had the shorter predicted survival time or the lower probability of surviving to some fixed time point. A pair of subjects is said to be usable if at least one of the subjects experienced the outcome (i.e. both subjects were not censored). Harrell's estimate of the c-statistic is defined to be the proportion of all usable pairs that are concordant. The formal estimator is $\mathrm{AUC(H)}(\tau) = \frac{\sum_{i\neq j}\{I(P_i > P_j)\cdot I(T_i < T_j, T_i < \tau)\cdot I(D_i{=}1)\}}{\sum_{i\neq j}\{I(T_i < T_j, T_i < \tau)\cdot I(D_i{=}1)\}}$.[8]

In Harrell's definition of the c-index, pairs are excluded if the subject with the shorter follow-up time was censored. Thus, the AUC(H) depends on the censoring mechanism, which is not a desirable property. To address this limitation of AUC(H), Uno et al. proposed an alternative estimator of Harrell's concordance index that uses an inverse probability of censoring weights.[8,9] $\mathrm{AUC(U)}(\tau) = \frac{\sum_{i,j}\{I(P_i > P_j)\cdot I(T_i < T_j, T_i < \tau)\cdot I(D_i{=}1)\cdot G(T_i)^{-2}\}}{\sum_{i,j}\{I(T_i < T_j, T_i < \tau)\cdot I(D_i{=}1)\cdot G(T_i)^{-2}\}}$, where $\mathrm{G(T_i)}$ is the Kaplan–Meier estimator of the censoring time distribution.

Finally, Gönen and Heller proposed a concordance index (which we denote as GHCI) that is a reversal of the above definitions of the c-statistic.[15] The theoretical definition of their concordance index is $\Pr(T_j > T_i|P_i \geq P_j)$, where T_i denotes the observed survival time for the *i*th subject. An estimator of the GHCI is $\frac{2}{N(N-1)}\sum_{i<j}\left\{\frac{I(\beta^{tr}X_i > \beta^{tr}X_j)}{1+\exp(\beta^{tr}X_j-\beta^{tr}X_i)} + \frac{I(\beta^{tr}X_i < \beta^{tr}X_j)}{1+\exp(\beta^{tr}X_i-\beta^{tr}X_j)}\right\}$ (where the superscript *tr* denotes the transpose of a vector), assuming that subjects are ordered according to increasing linear predictors $\beta^{tr}X_i$. The GHCI assumes that the Cox model is correctly specified.

Strictly speaking, the GHCI is not a c-statistic. However, we include it in this sub-section since it is a concordance-type index. We will use the term c-statistic when we are referring to one of AUC(H), AUC(CD), or AUC(U). We will use the term concordance or concordance statistics when we are referring to one of the three c-statistics or to the GHCI.

2.3 Royston's D-index

Royston and Sauerbrei proposed a measure of prognostic separation in survival data that has been referred to as D.[16] *D* measures prognostic separation of survival curves and is closely related to the standard deviation of the prognostic index. The prognostic index for each subject is defined to be the linear predictor from the fitted Cox proportional hazards model. Ranking the prognostic index across the sample, one assigns to the *i*th subject the *i*th expected standard normal order statistic in a sample of the same size. Using a Cox proportional hazards model, one then regresses the original time-to-event outcome on the expected standard normal order statistics. Royston's D-index is defined to be the estimated regression coefficient multiplied by $\sqrt{8/\pi}$. D can be interpreted as an estimate of the log hazard ratio comparing two prognostic groups of equal size (i.e. if one had dichotomized the linear predictor at the sample median).[16]

2.4 R^2-type measures

A large number of different R^2-type measures for use with survival models have been proposed.[16–30] Comprehensive reviews of these different measures are provided by Choodari-Oskooei et al.,[4,5] Hielscher et al.,[6] and by Schemper and Stare.[7] Most of these R^2-type measures attempt to mimic the definition of R^2 for linear models: $\mathrm{R}^2 = \frac{\mathrm{Var}(\beta' Z)}{\mathrm{Var}(\beta' Z)+\sigma^2}$, where σ^2 denotes the variance of the error term in the linear model. Several authors have proposed lists of desirable properties for measures of explained variation.[5,6] Choodari-Oskooei et al. divided these measures into four different classes: measures of explained variation, measures of explained randomness, measures of predictive accuracy, and a fourth class of miscellaneous measures that do not fall into any of the above three classes.[4,5] Measures in the first class have an interpretation that most closely resembles the popular interpretation of R^2 for linear models.

Choodari-Oskooei et al. suggest that a measure of explained variation should satisfy four properties: (i) independence from censoring (i.e. the measure is not affected by the degree of censoring in the data), (ii) monotonicity (i.e. that the measure of predictive accuracy takes on higher values as the magnitude of the effect of a covariate on the outcome increases), (iii) interpretability, and (iv) robustness against influential observations.[5] Based on these criteria, Choodari-Oskooei et al. suggest that two measures of explained variation should be used: Kent and O'Quigley's $\mathrm{R}^2_{\mathrm{PM}} = \frac{\mathrm{Var}(\beta' Z)}{\mathrm{Var}(\beta' Z)+\pi^2/6}$ and Royston and Sauerbrei's $R^2_D = \frac{D^2/(8/\pi)}{D^2/(8/\pi)+\pi^2/6}$ where D denotes Royston's D index described in Section 2.3. Amongst the measures of explained randomness, they suggested that Kent and O'Quigley's ρ^2_w be used, which can be approximated by $\rho^2_{w,a} = \frac{\mathrm{Var}(\beta' Z)}{\mathrm{Var}(\beta' Z)+1}$. We note that both Kent and O'Quigley measures assume that the fitted model is correct, since the estimated β coefficients are used in the calculations of R^2 and ρ^2.

In another review, Hielscher et al. suggested that two different measures of explained variation be used: An R^2 based on the Integrated Brier Score (IBS), R^2_{IBS}, proposed by Graf et al.[27] and R^2_{SH} of Schemper and Henderson[28] (Hielscher et al. use the notation R^2_D to refer to this latter measure, however we are already using this term for Royston and Sauerbrei's measure. As a further aside, Hielscher et al. classified these two measures as measures of explained variation, whereas Choodari-Oskooei described them as measures of predictive accuracy).

For binary prediction models, the Brier score is the mean squared prediction error. The IBS, an extension of this concept to survival outcomes, is defined as $\mathrm{IBS}(\tau) = \int E[(I(T > \tau) - S(\tau|Z))^2] dF_Z(Z)$, where $I(T > \tau) \in \{0, 1\}$ is the individual survival status at time τ and $S(\tau|Z)$ is the predicted survival probabilities from the model with covariate vector Z. The R^2 measure based on the IBS is defined as $R^2_{IBS}(\tau) = 1 - \frac{\mathrm{IBS}(\tau)}{\mathrm{IBS}_0(\tau)}$, where $\mathrm{IBS}_0(\tau)$ is the IBS for the marginal estimate of survival probabilities obtained using the Kaplan–Meier estimates.[27] Whereas the IBS will decrease with improvements in model fit, the R^2_{IBS} will increase with improvements in model fit. The latter quantity thus performs as one would expect for R^2-type measures.

Schemper and Henderson's integrated measure of predictive accuracy is defined as $R^2_{SH} = 2\int_0^\tau E[S(t|Z)\{1 - S(t|Z)\}] f(t) dt W(\tau)$, where $(0,\tau)$ is the follow-up period and $W(\tau) = \left\{\int_0^\tau f(t)dt\right\}^{-1}$.

Two different R^2-type measures, which fall into Choodari-Oskooei et al.'s miscellaneous class, are based on the likelihood ratio statistic for comparing the full model with the null model. Allison suggested $R^2 = 1 - \exp(-LR/N)$,[31] while Nagelkerke defined a generalized R^2 index as

$R_N^2 = \frac{1-\exp(-LR/N)}{1-\exp(-L^0/N)}$, where LR is the global log-likelihood ratio statistic for testing the importance of all p predictors in the regression model, L^0 is the –2 log likelihood for the null model, and N denotes the sample size or number of subjects in the sample.[10,23,32] In the latter measure, Allison's suggested measure is scaled by its maximum possible value, so that the resultant measure will lie between 0 and 1. O'Quigley et al. have criticized these measures as being inconsistent in the presence of censoring, as they converge to zero as the percentage of censoring increases[25] (p. 481). They proposed a modification to these measures, in which N, the number of subjects, is replaced by K, the number of uncensored subjects (i.e. the number of observed events). We will examine the performance of the modified Nagelkerke index as a comparator to the measures described earlier.

2.5 IDI

Pencina et al. suggested that the improvement in predicting the probability of the occurrence of a binary outcome due to the addition of a novel risk factor to an existing risk-prediction model can be summarized using the IDI.[33] Given a new regression model that includes a novel risk factor and an older regression model in which this risk factor is omitted, the IDI is estimated as

$$\text{IDI} = (\bar{\hat{p}}_{\text{new,events}} - \bar{\hat{p}}_{\text{new,nonevents}}) - (\bar{\hat{p}}_{\text{old,events}} - \bar{\hat{p}}_{\text{old,nonevents}})$$

where $\bar{\hat{p}}_{\text{new,events}}$ is the mean of the new model-based predicted probabilities of an event for those who develop events, while $\bar{\hat{p}}_{\text{new,nonevents}}$ is the mean of the new model-based predicted probabilities of an event for those who do not develop an event. $\bar{\hat{p}}_{\text{old,events}}$ and $\bar{\hat{p}}_{\text{old,nonevents}}$ are defined similarly for the old regression model. The IDI is different from the measures described in the previous sub-sections because it is not a measure of model performance. Rather, it is a measure for quantifying the improvements in predictive accuracy due to the inclusion of a novel risk factor.

Chambless et al. proposed a time-varying extension of the IDI to survival outcomes, based on the observation that the original IDI is related to the change in the proportion of variance explained between the model with and without the novel risk factor ($R^2(t)_{\text{new}}\text{and}R^2(t)_{\text{old}}$, respectively).[2] Chambless et al. defined this time-varying extension as $\text{IDI}(t) = R^2(t)_{\text{new}} - R^2(t)_{\text{old}}$, with $\hat{R}^2(t) = \frac{\hat{\text{V}}\text{ar}(S(t|X))}{\hat{S}(t)\times(1-\hat{S}(t))}$, where $\hat{S}(t)$ denotes the mean survival function at time t across all subjects in the sample, while $\hat{\text{V}}\text{ar}(S(t|X))$ denotes the variance of the survival function at time t across all subjects. $\hat{S}(t|Z_i)$ is the estimated survival function derived from the estimated proportional hazards regression model, while $S(t) = E(S(t|Z))$ can be estimated by averaging the estimated survival functions across all subjects in the sample. We note that this definition of R^2 differs from those discussed in the previous sub-section.

2.6 Software

We used R software (version 2.15.2, R Foundation for Statistical Computing, Vienna, Austria) to simulate the random datasets. AUC(CD), AUC(U), and GHCI were estimated using the AUC.cd, UnoC, and GHCI functions in the survAUC package. The IBS of Graf et al., R_{IBS}^2, was estimated using the pec and ibs functions in the pec package. Schemper and Henderson's R_{SH}^2 was estimated using R code provided by Lara Lusa (http://cemsiis.meduniwien.ac.at/en/kb/science-research/software/statistical-software/surevsurev/ – site accessed 26 September 2013)[34] We used SAS® macros provided by Chambless and colleagues to compute the IDI (www.aricnews.net – site accessed 4 October 2013).[2] The other measures of model performance were estimated using components extracted from the fitted Cox proportional regression models.

3 Methods – Monte Carlo simulations

An extensive series of Monte Carlo simulations was performed, similar in design to those in a recent study examining this issue in the context of binary outcomes and logistic regression models.[3] Our simulations incorporated the following design elements: (i) an existing prediction model that related a continuous risk factor to the hazard of the occurrence of a time-to-event outcome with a given predictive accuracy (i.e. the 'reference model'), (ii) the addition of either a novel binary or continuous risk factor to the existing model, (iii) variations in the magnitude of the correlation between the existing continuous risk factor and the novel risk factor, (iv) variations in the prevalence of the novel binary risk factor, (v) variation in the magnitude of the hazard ratio relating the novel risk factor to the hazard of the occurrence of the outcome, (vi) scenarios with a high degree of censoring (low observed event rate) versus scenarios with a low degree of censoring (high observed event rate), (vii) censoring occurring due to administrative censoring versus censoring occurring due to study dropout or loss to follow-up.

We describe one scenario in detail and then describe briefly how this scenario was modified in subsequent scenarios. For each of 1000 subjects, we randomly generated a continuous predictor variable denoting an established risk factor from a standard normal distribution: $x_{1i} \sim N(0,1)$. We then simulated a novel binary risk factor from a Bernoulli distribution with parameter $\mathrm{P}_{\text{risk-factor}}$:$x_{2i} \sim Be(\mathrm{P}_{\text{risk}-\text{factor}})$. Thus, the prevalence of binary risk factor in the population is $\mathrm{P}_{\text{risk-factor}}$. In this first scenario, the established, continuous, risk factor was assumed to be independent of the novel binary risk factor (this will be modified in subsequent scenarios). A time-to-event outcome was generated for each of the 1000 subjects using a Cox–Weibull model for generating time-to-event outcomes.[35] For each subject, the linear predictor was defined as $\mathrm{LP} = \alpha_1 x_{1i} + \alpha_2 x_{2i}$. For each subject, we generated a random number from a standard uniform distribution: $\mathrm{u} \sim \mathrm{U}(0,1)$. A survival or event time was generated for each subjects as follows: $\left(-\frac{\log(u)}{\lambda \exp(\mathrm{LP})}\right)^{1/\eta}$. We set λ and η to be equal to 0.25 and 0.5, respectively. We set two different degrees of administrative censoring (25% and 90% right censoring when the study observation period ended – thus the observed event rate was 75% and 10% in these two scenarios, respectively). For the first setting, we defined t_0 to be the 75th percentile of survival times in one simulated dataset of size 1,000,000. Subjects whose survival times exceeded t_0 were then subjected to administrative censoring (t_0 was defined analogously in the setting with a high degree of administrative censoring).

Two Cox proportional hazards models were fit in each simulated dataset: a regression model consisting of only the continuous risk factor x_1 and a regression model consisting of the continuous risk factor x_1 and the novel binary risk factor x_2. The predictive accuracy of each of the two models was determined using the different methods described in the preceding section. The change in predictive accuracy was then determined across 1000 simulated datasets (each consisting of 1000 subjects). For those measures of predictive accuracy that are time varying (AUC(CD), AUC(u), and Chambless' IDI), we estimated model performance at $\tau = \mathrm{t}_0$, where t_0 is as defined earlier. Similarly, the IBS was determined over the maximum duration of follow-up.

In the Monte Carlo simulations, the following factors were varied: (1) the proportion of subjects for whom an event was observed to occur: 0.10 versus 0.75 (low versus high rate of observed events); (2) $\exp(\alpha_1)$ (the hazard ratio for the continuous risk factor): from 1 to 3 in increments of 0.2; (3) the prevalence of the novel binary risk factor: 0.10, 0.25, and 0.50; (4) $\exp(\alpha_2)$ (the hazard ratio for the novel binary risk factor): 1.10, 1.25, 1.50, and 2.0. We thus examined 264 ($2 \times 11 \times 3 \times 4$) different scenarios in which the novel risk factor was binary and was independent of the existing continuous risk factor. Each statistical method was applied to the same simulated datasets, so that the simulated datasets did not vary between the statistical methods.

We modified the above scenario by inducing a correlation between the existing continuous risk factor and the novel binary risk factor. We examined two different scenarios characterized by different degrees of correlation between the existing continuous risk factor and the novel binary risk factor. To do so, we simulated two continuous risk factors from a standard bivariate normal distribution with correlation ρ between the two components. The first component was used as the existing continuous risk factor. We categorized the second component at a given threshold, and defined the novel binary risk factor to be present if the second component lay above the threshold, and to be absent if the second component lay below the threshold. The threshold was selected so that the prevalence of the novel binary risk factor was as described earlier (0.1, 0.25, and 0.50). The simulations then proceeded as described earlier. In the first modification, ρ was set to 0.5, while in the second modification, ρ was set to 0.8.

The above three sets of simulations examined changes in predictive accuracy due to the inclusion of a novel binary risk factor to a Cox proportional hazards model that consisted of an existing continuous risk factor. We modified the original scenario to examine changes in model performance due to the inclusion of a novel continuous risk factor. In this fourth set of simulations, the novel risk factor was simulated from a standard normal distribution. Furthermore, it was generated so as to be independent of the continuous existing risk factor. The hazard ratio for the continuous novel risk factor took on the following values: 1.1, 1.25, 1.5, and 2. We thus examined 88 different scenarios (two degrees of censoring × 11 hazard ratios for the existing continuous risk factor × four hazard ratios for the novel continuous risk factor). The simulations then proceeded as described earlier.

The above sets of simulations used data-generating processes that induced administrative censoring, with either 25 or 90% of subjects being censored due to study termination (all subjects whose survival time exceeded t_0 had their event time censored at time t_0). We then repeated each of the above scenarios in which the novel risk factor was independent of the existing continuous risk factor (due to space and time constraints, we did not examine the scenarios in which the two risk factors were correlated), with censoring induced by loss to follow-up or study dropout. We induced study dropout so that some subjects dropped out prior to the true event time. We determined each subject's dropout status from a Bernoulli distribution with parameter $P_{dropout}$. Subjects who were selected as dropping out had a dropout time selected from a uniform $(0,T^S)$ distribution, where T^S denotes the subject's true event time. We examined two different values of $P_{dropout}$: 0.25 and 0.90, so that 25% and 90% of subjects were subject to censoring due to study dropout, with the event being observed for the remaining subjects. The simulations then proceeded as described earlier.

4 Results – Monte Carlo simulations

We report our results separately for the four scenarios defined by the nature of the novel risk factor (binary versus continuous) and its correlation with the existing continuous risk factor. Results are reported in detail for the first setting and more briefly for the remaining settings.

4.1 Binary risk factor independent of existing continuous risk factor

Results for the scenarios with an independent binary risk factor that is uncorrelated with the existing continuous risk factor, in which there was a low event rate (i.e. high rate of censoring), and in which censoring was due to administrative censoring are reported in Figures 1 to 3. In Figure 1, we report

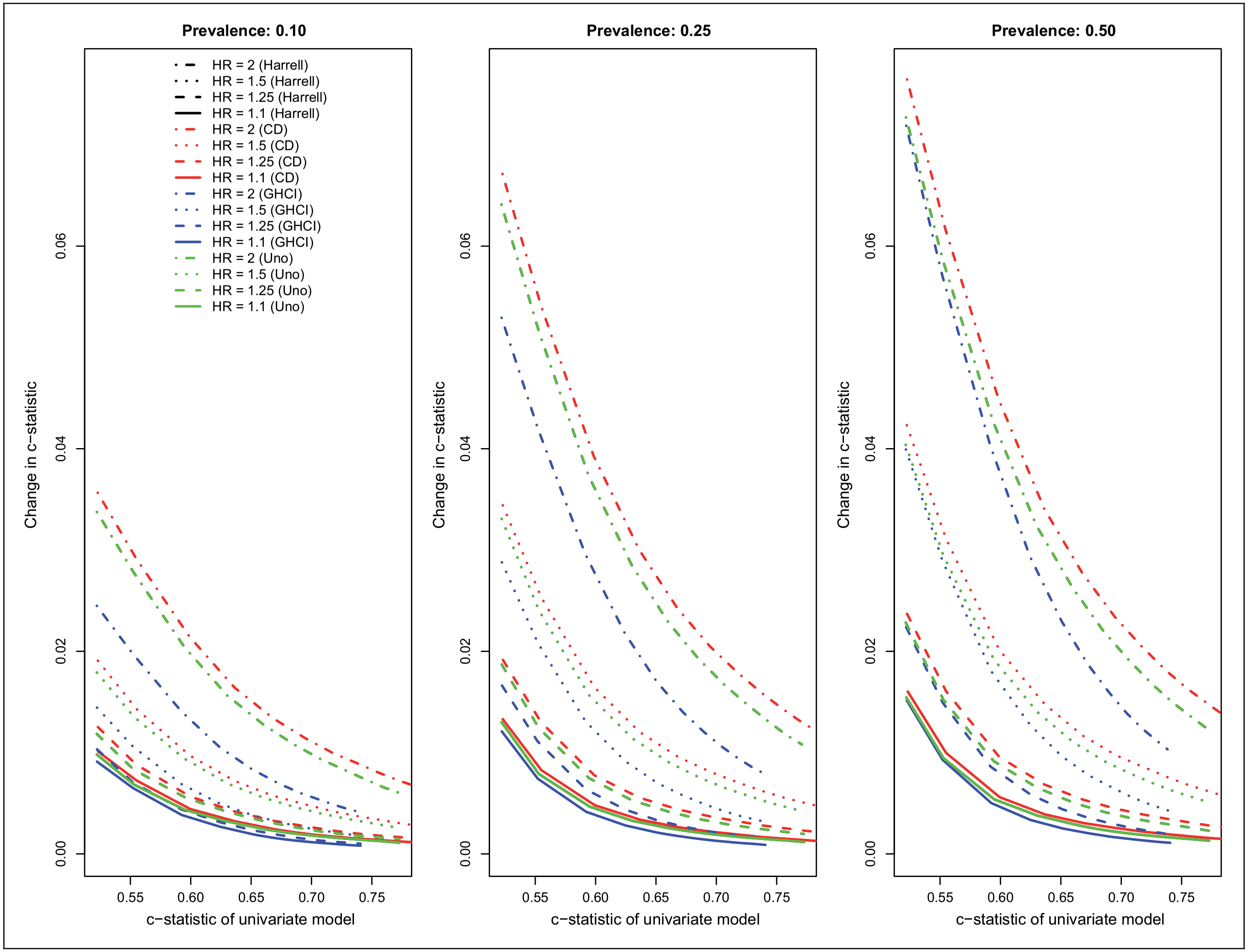

Figure 1. Relationship between change in c-statistic and c-statistic of univariate model (low event rate – uncorrelated binary).

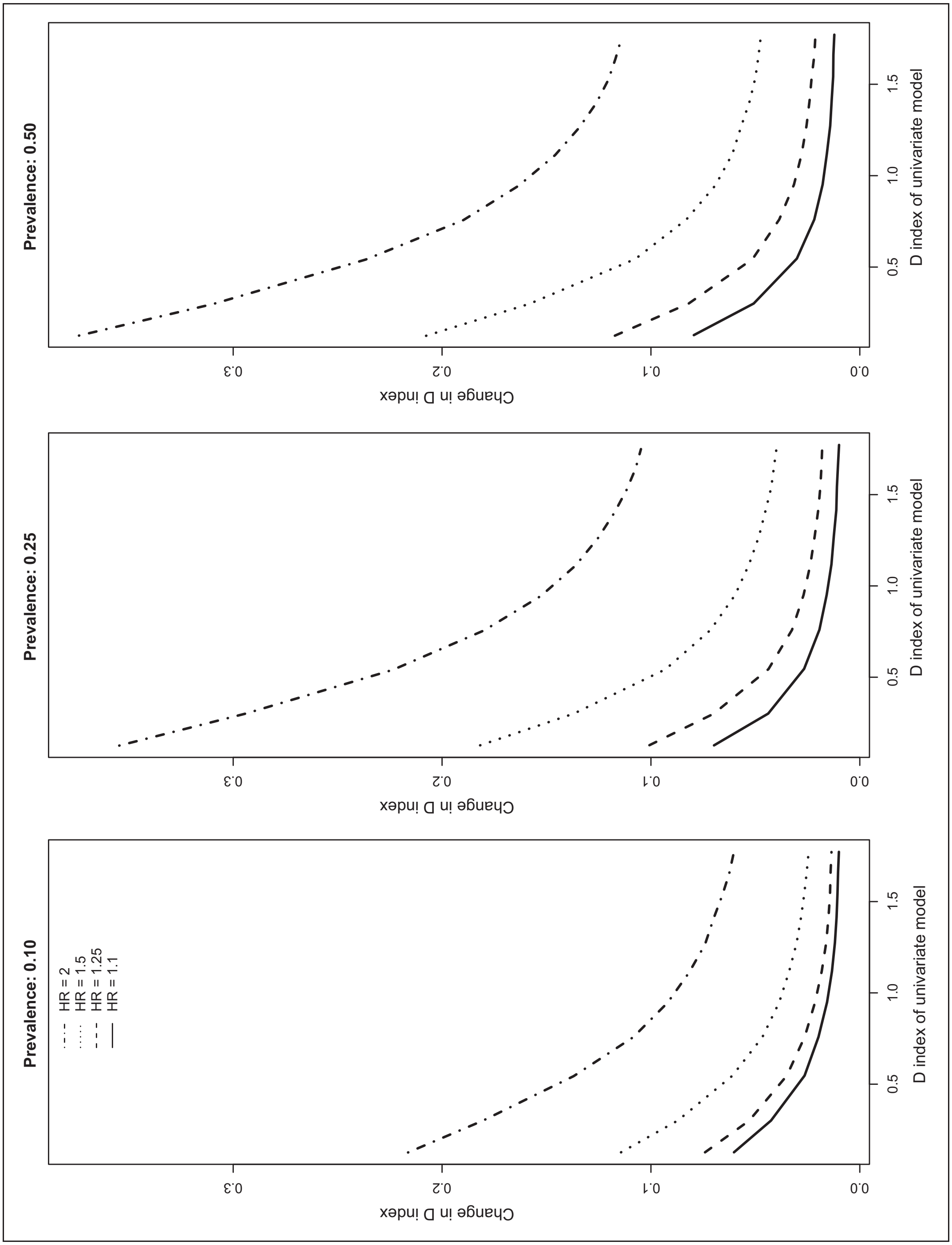

Figure 2. Relationship between change in D index and D index of univariate model (low event rate – uncorrelated binary).

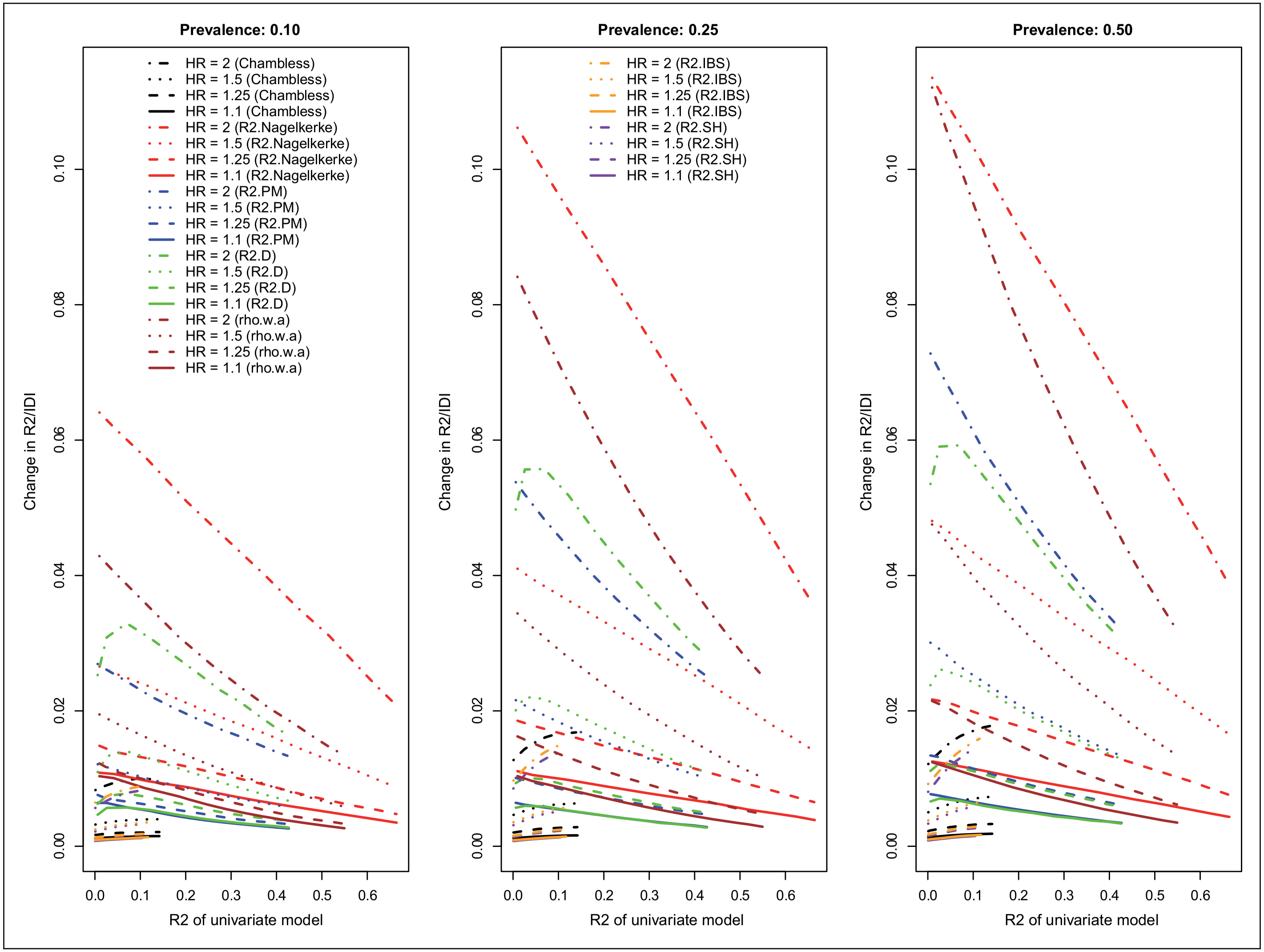

Figure 3. Relationship between change in R2 and R2 of univariate model (low event rate – uncorrelated binary).

the changes for the four concordance statistics (AUC(H), AUC(CD), AUC(U), and GHCI) when the novel binary risk factor was added to the regression model. Several observations merit comment. First, improvements in concordance due to the inclusion of the binary risk factor decreased as the concordance of the reference model increased. Second, improvements in concordance increased as the hazard ratio of the novel binary risk factor increased. Third, improvements in concordance due to the addition of the novel binary risk factor increased as the prevalence of the binary risk factor increased. Fourth, improvements in AUC(H) were essentially identical to improvements in AUC(U). Fifth, improvements in the AUC(CD) were modestly greater than improvements in the other three concordance measures while improvements in GHCI were modestly lower than improvements in the other three concordance measures. Changes in Royston's D index due to the inclusion of the novel binary risk factor displayed similar patterns (Figure 2).

The relationship between the change in each R^2-type measure and the R^2 of the univariate model is described in Figure 3. Because Chambless characterized the IDI as a difference in R^2-type measures, we have superimposed on this figure the relationship between Chambless' IDI and Chambless' estimate of R^2 for the reference model. Three of the R^2-type measures (the adapted Nagelkerke R_N^2 statistic, R_{PM}^2, and $\rho_{w,a}^2$) displayed similar patterns to those described earlier. A fourth measure, R_D^2, displayed an attenuation in the improvement in R^2 due to the inclusion of the novel risk factor when the R^2 of the univariate model was very low. The pattern of results for Chambless' adaptation of IDI and for R_{SH}^2 was less consistent than the results for the other R^2-type measures and the results described in the preceding paragraph. Furthermore, for both of these measures, improvements in model performance increased due the inclusion of the binary risk factor as the performance of the univariate model increased. For any given scenario, the adapted R_N^2 displayed greater increases due to the addition of the binary risk factor than did the other R^2-type measures. Both the IDI and the change in R_{IBS}^2 displayed less variability across the different scenarios than did the other R^2 measures. Finally, Chambless' R^2, R_{SH}^2, and R_{IBS}^2, displayed decreased variability across the different baseline models compared to that observed for the other R^2-measures.

The above findings pertain to the settings with a low observed event rate (i.e. the outcome was observed to occur for approximately 10% of subjects, with the remainder being subject to administrative censoring). Results for the settings with a high observed event rate (i.e. the outcome was observed to occur for approximately 75% of subjects, with the remainder being subject to administrative censoring) were similar to those observed in the setting with a low observed event rate (Figures 4 to 6). However, the absolute increase in concordance due to the inclusion of the novel binary risk factor tended to be greater when the observed event rate was high compared to when it was low. Second, differences between AUC(CD) and the other three concordance measures were amplified, while differences between GHCI and AUC(H) and AUC(U) were attenuated. The increase in the observed event rate tended to magnify the absolute increase in Royston's D index, albeit to a lesser degree than for the concordance indices. The differences between the previously observed results for R_{SH}^2 and Chambless' adaptation of the IDI from those of the other R^2-type measures (i.e. the previously observed increasing magnitude of improvements as the R^2 of the reference model increased) were no longer apparent in the settings with a high observed event rate. Differences between R_N^2 and $\rho_{w,a}^2$ were substantially diminished in the presence of a high event rate. Finally, Chambless' R^2, R_{SH}^2, and R_{IBS}^2 of the univariate model displayed increased variability in the presence of a high event rate compared to in the presence of a low event rate. However, this measure still displayed decreased variability compared to the other R^2-type measures. Changes in Chambless' R^2, R_{SH}^2, and R_{IBS}^2 due to the inclusion of a binary risk factor tended to be smaller in magnitude compared to changes in the other R^2-type measures.

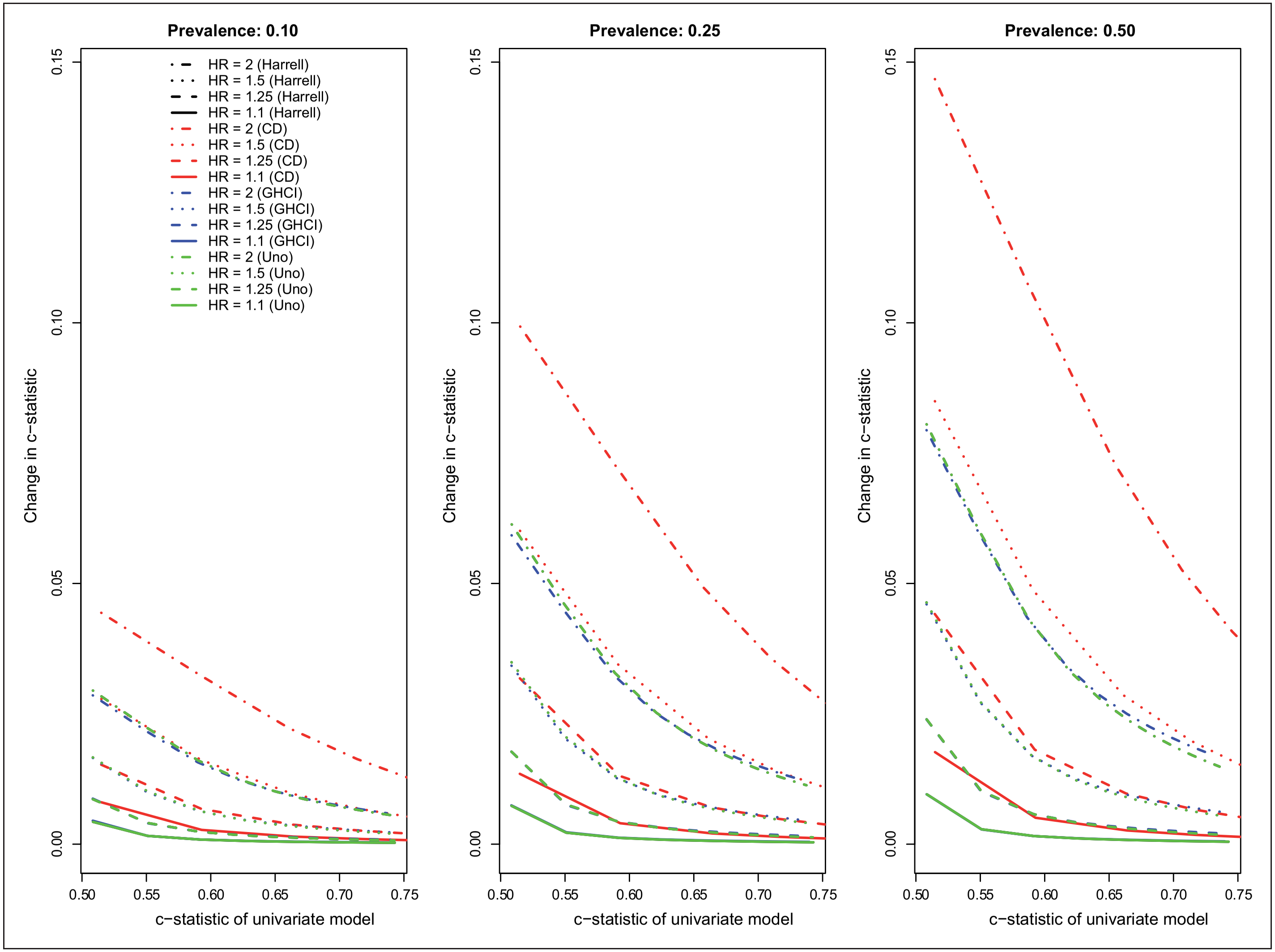

Figure 4. Relationship between change in c-statistic and c-statistic of univariate model (high event rate – uncorrelated binary).

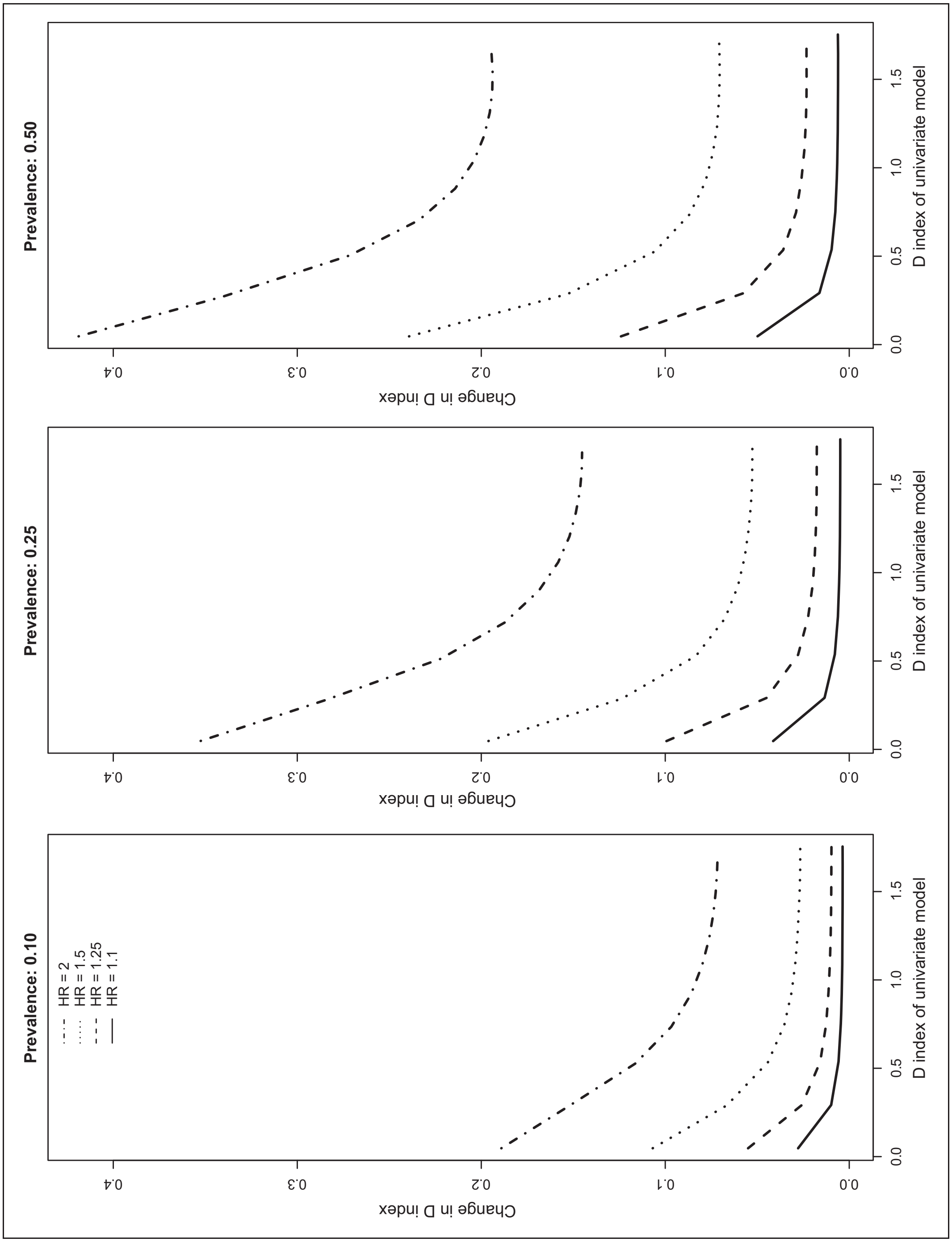

Figure 5. Relationship between change in D index and D index of univariate model (high event rate – uncorrelated binary).

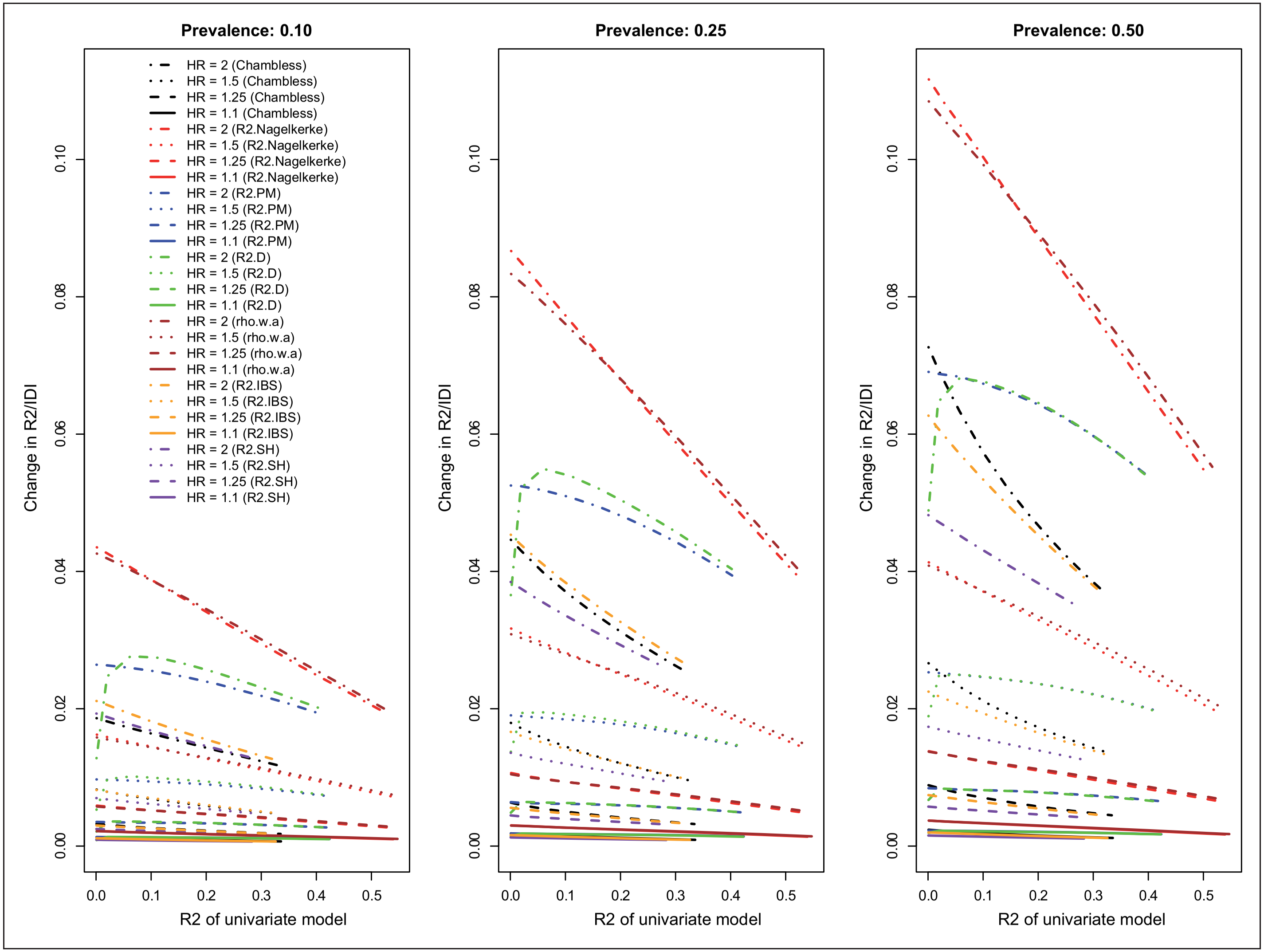

Figure 6. Relationship between change in R2 and R2 of univariate model (high event rate – uncorrelated binary).

When censoring was due to study dropout and event rates were high, results were qualitatively similar to those described earlier in the settings with administrative censoring. One noticeable change was that, in the presence of censoring due to study dropout, the effect of a novel marker was more pronounced on $\rho^2_{w,a}$ than it was on the modified Nagelkerke R^2. The R^2_{IBS} of the reference model displayed increased variability in the presence of censoring due to study dropout than in the presence of administrative censoring when the event rate was low. Furthermore, changes in R^2_{SH} and R^2_{IBS} due to the addition of the novel risk factor tended to decrease with increasing performance of the reference model, whereas the reverse was observed in the presence of administrative censoring when the event rate was low.

4.2 Binary risk factor correlated with the existing continuous risk factor

We considered four different sets of scenarios in which the novel binary risk factor was correlated with the existing continuous risk factor and in which subjects were subject to administrative censoring (low observed event rate versus high observed event rate; correlation of 0.5 versus correlation of 0.8). Results were qualitatively similar to those described above (results for the scenarios with a high observed event rate and a correlation of 0.5 are described in Figures A to C in the online supplemental material (available at http://smm.sagepub.com), the other results are not provided). The primary exception was that the changes in concordance and Royston's D index did not decrease smoothly as the predictive accuracy of the reference model increased. Instead, there was some jaggedness evident in the lines. In the presence of a low event rate, a low prevalence (10%) of the binary risk factor, and a high correlation between the binary risk factor and the continuous risk factor, then some of the R^2 measures (in particular R^2_{PM} and $\rho^2_{w,a}$) displayed some changes in behaviour, with the magnitude of increases in model performance due to the addition of the binary risk factor increasing as the performance of the reference model increased.

4.3 Continuous risk factor independent of existing continuous risk factor

Results for the setting with a continuous novel risk factor that is uncorrelated with the existing continuous risk factor, and in which there was a low observed event rate due to administrative censoring, were similar to those noted in Section 4.1 (Figure 7). For the IDI, R^2_{IBS} and R^2_{SH}, improvements in model performance due to the addition of the novel continuous risk factor increased as the performance of the reference model increased. With a high observed event rate in the presence of administrative censoring (Figure 8), all results were similar to those observed in Section 4.1. As with the binary risk factor, the differences in the observed behaviour between R^2_{SH}, R^2_{IBS}, and Chambless' IDI and the other R^2-type measures disappeared in the presence of a high event rate. For the concordance measures, the absolute increase in model accuracy was greater in the presence of a high observed event rate compared to in the presence of a low observed event rate.

When censoring was due to study dropout, rather than to administrative censoring, results tended to be qualitatively similar, with only minor deviations from those described in the previous paragraph. When censoring was due to study dropout, absolute increases in the concordance indices tended to be modestly greater compared to when censoring was due to administrative censoring. When censoring was due to study dropout, the performance of AUC(H), AUC(Uno), and the GHCI were similar to one another, while AUC(CD) tended to display greater changes in discrimination due to the addition of the continuous novel risk factor than did the other concordance measures. Furthermore, when censoring was due to study dropout, R^2_{SH}, R^2_{IBS}, and Chambless' IDI tended to have a behaviour that was more similar to that of the R^2-type measures.

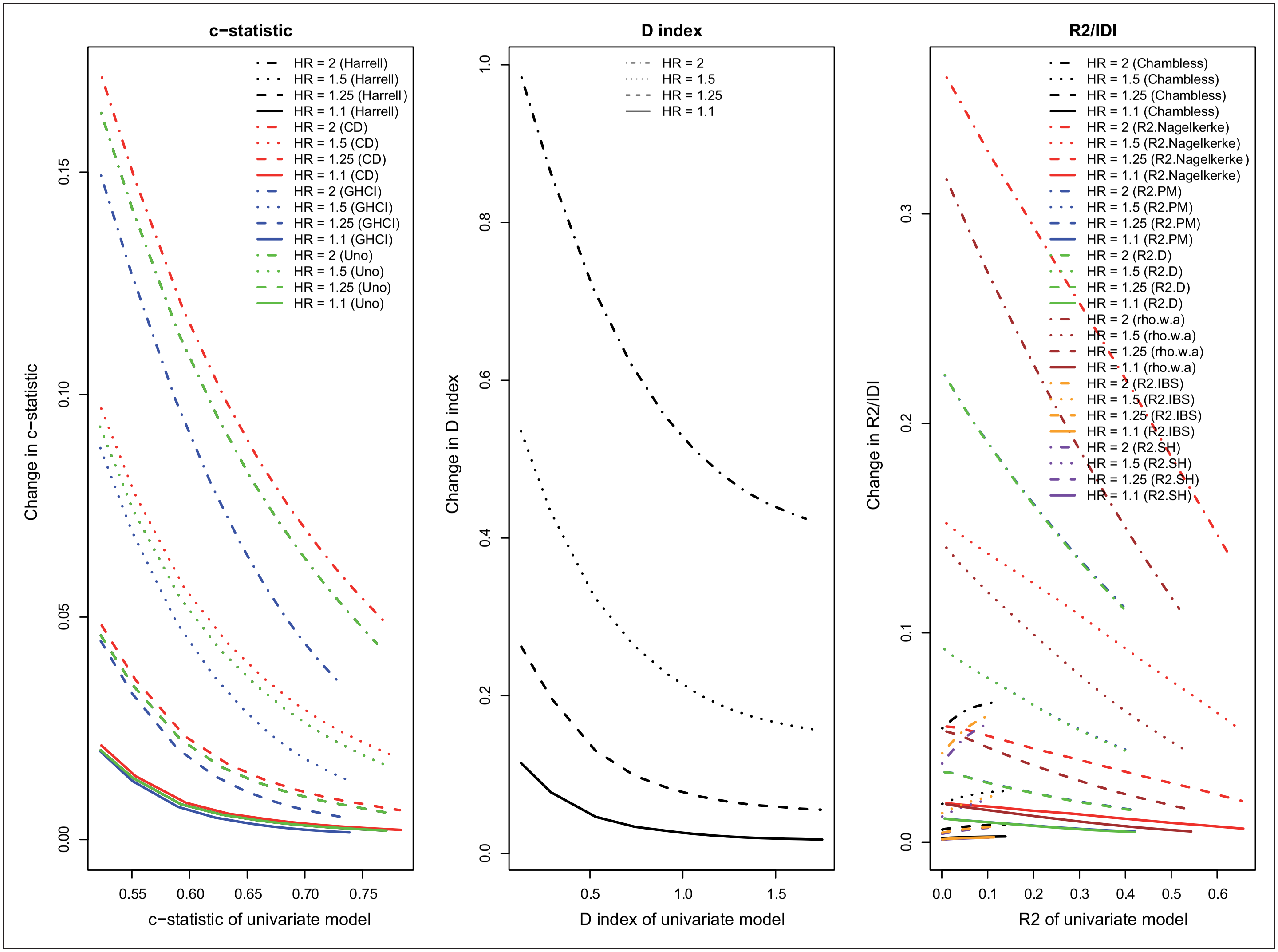

Figure 7. Relationship between change in model accuracy and model accuracy of univariate model (low event rate – uncorrelated continuous).

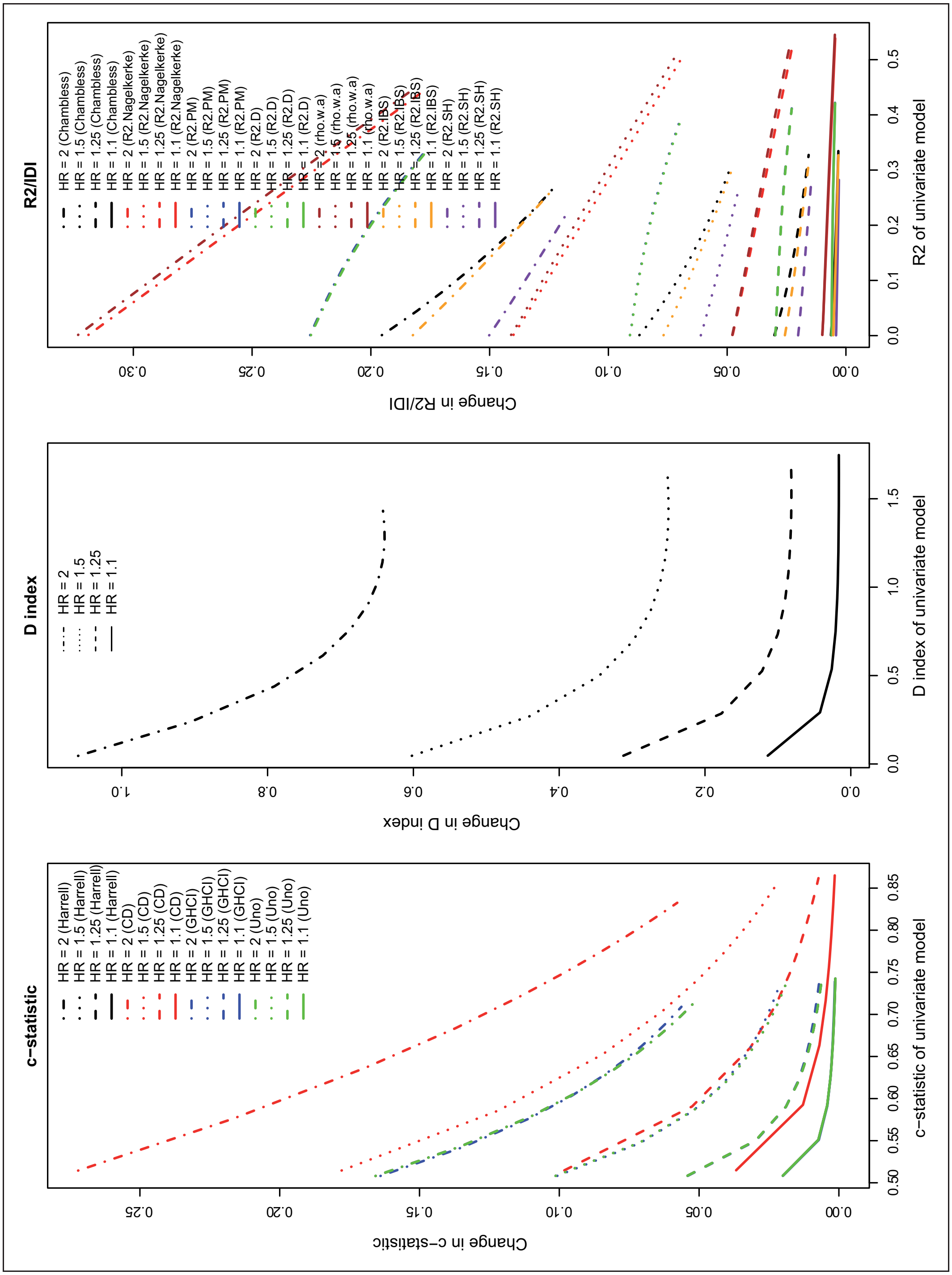

Figure 8. Relationship between change in model accuracy and model accuracy of univariate model (high event rate – uncorrelated continuous).

5 Case study

We provide a brief case study to compare the change in different measures of model performance when a risk factor is added to an existing Cox proportional hazards regression model. The sample consisted of patients hospitalized with heart failure and the survival outcome was time to death, with patients censored after 365 days of follow-up.

The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study was a cluster randomized trial intended to improve the quality of care for patients with cardiovascular disease in Ontario, Canada.[36,37] During the first phase of the study, detailed clinical data on patients hospitalized with heart failure between 1 April 1999 and 31 March 2001 at 103 hospitals in Ontario, Canada were obtained by retrospective chart review. Data on patient demographics, vital signs and physical examination at presentation, medical history, and results of laboratory tests were collected. Subjects with missing data on continuous baseline covariates necessary to estimate the risk score were excluded from the current case study, leaving 8635 patients for analysis.

We considered two variables for predicting the hazard of death over the 365 days subsequent to hospital admission. The first predictor variable was patient age. The second was the EFFECT-HF mortality prediction score, which is a point-based scoring system for predicting the risk of 30-day and 1-year mortality.[38] The score includes age, respiratory rate, systolic blood pressure, urea nitrogen, sodium concentration, cerebrovascular disease, dementia, chronic obstructive pulmonary disease, hepatic cirrhosis, cancer, and haemoglobin. Importantly, patient age is one of the components of the EFFECT-HF mortality prediction score. The correlation between these two variables was 0.63 in the study sample.

We considered two reference prediction models: the first contained only patient age as a linear variable, while the second contained only the EFFECT-HF score as a linear variable. To each model, we then added the other variable. Thus, we examined the effect of adding patient age to a model that initially consisted of only the EFFECT-HF score, and we examined the effect of adding the EFFECT-HF score to a model that initially consisted of only patient age. For each of the Cox proportional hazard regression models, we computed the performance measures described in Section 2. We determined the change in each measure of model performance when the second variable was added to the existing prediction model. All of the time-varying measures of model performance were assessed at 365 days (the time at which all subjects were subject to administrative censoring). Similarly, the IBS was determined over the 365 days of follow-up.

Within one year of hospital admission, 2825 (33%) patients died, while the remainder were subject to administrative censoring after one year of follow-up. Adding age to a prediction model that consisted of only the EFFECT-HF score resulted in negligible changes in predictive performance (range from –0.0003 to 0.0046, Table 1). The relative change in model performance due to the inclusion of age ranged from 0% to 1.6%. Thus, across all measures of model performance, there was a consistent conclusion that adding age to a model consisting of the EFFECT-HF score did not improve prognostic performance.

Adding the EFFECT score to a prediction model that consisted of only patient age resulted in substantially larger changes in model performance (range from 0.0748 to 0.5237). Royston's D displayed the greatest absolute increase when the EFFECT-HF score was added to the model consisting of age alone (0.5237). The changes in the four concordance indices due to the addition of the EFFECT-HF score were similar to one another. The ordering of the absolute change in the four concordance indices was similar to that observed in our simulations (Figure 7): the largest absolute change was observed for AUC(CD), while the smallest absolute change was observed for GHCI concordance index. There was greater disparity in the magnitude of the change in model R^2

Table 1. Performance measures for predicting mortality in patients hospitalized with heart failure.

Performance measure	Model: age only	Model: age + EFFECT score	Absolute change in performance	Relative change in performance (%)	Model: EFFECT score only	Model: EFFECT score + age	Absolute change in performance	Relative change in performance (%)
AUC(H)	0.6150	0.7002	0.0852	13.8	0.7004	0.7002	–0.0003	0.0
AUC(CD)	0.6220	0.7225	0.1004	16.1	0.7208	0.7225	0.0016	0.2
AUC(U)	0.6012	0.7001	0.0989	16.4	0.6998	0.7001	0.0003	0.0
GHCI	0.6012	0.6761	0.0748	12.4	0.6741	0.6761	0.0020	0.3
Royston's D index	0.6566	1.1804	0.5237	79.8	1.1798	1.1804	0.0006	0.1
Chambless R^2	0.0500	0.1456	0.0956	191.1	0.1448	0.1456	0.0008	0.5
R^2_{PM}	0.1074	0.2392	0.1317	122.6	0.2355	0.2392	0.0037	1.6
R^2_D	0.0933	0.2496	0.1563	167.4	0.2494	0.2496	0.0002	0.1
ρ^2_{wa}	0.1653	0.3408	0.1756	106.2	0.3362	0.3408	0.0046	1.4
R^2 (Nagelkerke)	0.1470	0.3894	0.2424	164.8	0.3891	0.3894	0.0003	0.1
R^2_{IBS}	0.0398	0.1256	0.0857	215.2	0.1254	0.1256	0.0002	0.1
R^2_{SH}	0.0459	0.1382	0.0923	201.3	0.1379	0.1382	0.0003	0.2

across the different R^2-type measures. The relative change in the four concordance indices ranged from 12.4 to 16.4%, while the relative change in Royston's D was 80%. The relative change in the different R^2-type measures ranged from 106% ($\rho^2_{w,a}$) to 215% (R^2_{IBS}), while Chambless' R^2-type measure displayed a relative change of 191%. While all measures allowed one to conclude that predictive accuracy increased with the inclusion of the EFFECT-HF score, the relative increase in model performance varied substantially across the different measures of model performance. In particular, the relative increase in the concordance statistics was substantially less than that of the R^2-type measures.

These results are concordant with the results of our simulations. In our simulations, we found that greater improvements in model performance were possible when the reference model had lower predictive accuracy. In examining Table 1, one notes that, across all measures of model performance, the reference model consisting of age had lower predictive accuracy than the reference model consisting of the EFFECT-HF score. Thus, one would anticipate greater improvements in model performance when adding the EFFECT-HF score to the reference model consisting of age alone, compared to when adding age to the reference model that consisted of the EFFECT-HF score. There is a simple explanation for these observations. The EFFECT-HF score incorporates 11 variables, one of which is age. Thus, adding the EFFECT-HF score to a model consisting of age is comparable to incorporating 10 additional covariates to the clinical prediction model. However, adding age to the model consisting of the EFFECT-HF score has a minimal effect since the existing score already incorporates the effect of age on the hazard of the outcome.

6 Discussion

We used an extensive set of Monte Carlo simulations to explore changes in measures of model performance for Cox proportional hazards models when a novel risk factor or marker was added to an existing regression model. We summarize our findings in two separate sets of conclusions.

The first set of conclusions pertains to factors that influence the sensitivity of a measure of model performance to the inclusion of a novel risk factor or marker. The second set of conclusions pertains to comparing the relative sensitivity of different measures of model performance to the inclusion of a novel risk factor or marker.

In examining factors that influence the sensitivity of a measure of model performance to the inclusion of a novel risk factor, several observations merit mention. First, the increase in predictive accuracy due to the inclusion of a novel risk factor tended to decrease as the predictive accuracy of the reference Cox proportional hazards regression model increased. Second, the magnitude of the increase in predictive accuracy due to the inclusion of a novel risk factor increased as the hazard ratio associated with this novel risk factor increased. Third, the increase in predictive accuracy due to the inclusion of a binary risk factor increased as the prevalence of the binary risk factor increased. The above findings were observed to hold for all measures of predictive accuracy examined: concordance indices, Royston's D index, and R^2-type measures. In some settings (low event rate and administrative censoring), Chambless' R^2, R^2_{IBS}, and R^2_{SH} exhibited some discrepant patterns, with improvements in model performance *increasing* as the performance of the reference model increased. For the concordance indices, the absolute increase in predictive accuracy due to the inclusion of a novel risk factor was greater when the observed event rate was higher compared to when the observed event rate was lower.

In comparing the relative sensitivity of different measures of model performance to the inclusion of a novel risk factor, several observations merit mention. First, we observed that improvements in AUC(H) were essentially identical to improvements in AUC(U). Thus, in settings reflective of those that we considered in our simulations, these two estimators should result in similar conclusions about the incremental benefit of novel risk factors or markers. Second, the behaviour for Chambless' adaptation of the IDI tended to be less consistent with those of the other measures of model performance. In particular, in some scenarios, the behaviour of Chambless' adaptation of IDI was different from that of the majority of R^2-type measures and from the concordance indices and from Royston's D. In some instances, the magnitude of the IDI increased as the performance of the reference model increased. This suggests that further attention needs to be focussed on Chambless' time-varying definition of R^2. Given differences in performance between this estimate and that of most of the other R^2-type measures, further research is required to explore its properties.

Some secondary observations include that the improvements in predictive accuracy and the accuracy of the reference regression model were no longer strictly decreasing when the novel binary risk factor was correlated with the existing continuous risk factor. Instead, the relations were somewhat jagged. A similar phenomenon was observed in our prior study.[3] In that prior study, when a novel binary risk factor was added to an existing logistic regression model, and when the novel risk factor was correlated with the existing continuous risk factor, then the relation between the change in accuracy (using either the c-statistic or Nagelkerke's R^2 statistic) and the accuracy of the reference logistic regression model displayed a similar jaggedness. The similarity of this finding in both in settings with binary outcomes and in settings with time-to-event outcomes suggests that this issue needs to be examined in greater detail in subsequent research on the mathematical properties of these estimators. However, this is beyond the scope of the current study. Finally, Chambless' R^2, R^2_{IBS}, and R^2_{SH} frequently displayed less variability across the different univariate or reference models compared to the other R^2-type measures.

A limitation of the current study is that we have focussed exclusively on the setting in which a single novel risk factor or marker is added to an existing reference clinical prediction model. This was done to reflect what we perceive to be the most common scenario in research on novel risk factors and markers. Our methods could easily be extended to examine multiple novel risk factors or markers. One would need to consider several factors in a multi-factorial design: (i) the correlation between the different novel risk factors or markers, (ii) the magnitude of the independent effect of each novel risk factor on the hazard of the outcome, and (iii) the prevalence of each of the novel risk factors or markers. We see no reason to anticipate why the sensitivity of the different performance measures would differ in the setting of multiple novel risk factors.

Many of the above observations have important consequences for researchers seeking to appraise novel risk factors or markers (e.g. genetic factors, biomarkers, lifestyle characteristics, or patient characteristics) that add prognostic information above and beyond that contained in conventional clinical prediction models when outcomes are time-to-event in nature. First, identification of risk factors that have a stronger etiological effect (i.e. that have a larger hazard ratio) will result in greater improvements in predictive accuracy, provided that these hazard ratios relate to predictor variables that have the same distribution. One way to achieve comparability in hazard ratios is by standardizing the predictor so that it has unit variance, or letting the hazard ratio refer to the 75 versus 25 percentile.[10,11] Second, identification of binary risk factors that are common or have higher prevalence in the population will result in greater improvements in predictive accuracy.[39] Third, greater improvements in absolute estimates of predictive accuracy can be expected in settings in which existing risk prediction models have low predictive accuracy. For fair comparisons, investigators should use the best existing model as the reference model to avoid overstating the incremental benefit of novel risk factors and markers.[40] Similarly, reviewers and readers should ascertain whether published studies have used the best available model as the reference model. Of note, studies should be sufficiently large for unbiased estimation of performance, since overfitting will cause upward bias in larger models where relatively many factors are included in the model.[10,41,42] Finally, studies with longer durations of follow-up (and thus with a higher observed event rate) will allow for greater increases in concordance due to the inclusion of a novel risk factor than studies with shorter duration of follow-up (and thus with a lower observed event rate).

One of the key conclusions is that the findings of the current study are largely similar to those of a previous study that compared changes in predictive accuracy when novel risk factors or markers are added to logistic regression models for predicting binary outcomes.[3] The similarities of the current findings with those from the prior study suggest that these findings describe underlying properties of the relations between the characteristics of different risk factors and improvements in predictive accuracy. It is important to note that consistent findings were observed regardless of the nature of the outcome (survival versus binary) and regardless of the measure of predictive accuracy that was used. Thus, the suggestions for biomedical researchers provided in the paragraph above are likely to be relevant regardless of the nature of the outcome of interest or of how predictive accuracy is quantified.

When selecting a measure of model performance, Harrell suggests that, while rank measures (such as concordance indices) may be useful for describing a given prediction model, they may not be very sensitive in choosing between competing models, and that this may be especially true when the models are strong[10] (page 78). Furthermore, he suggests that measures such as R^2 are more sensitive. However, he notes that an absolute change in R^2 may be difficult to interpret. Similarly, Uno et al. suggested that while c-statistics are commonly used to quantify the predictive ability of clinical prediction models, they are not sensitive for determining the incremental benefit of additional risk factors or markers[9] (page 1113). Furthermore, Uno et al. suggested that using differences in measures of explained variation may be more sensitive in detecting differences in predictive ability. Our findings provide support for these comments: greater increase in predictive accuracy is possible when the reference model has lower predictive accuracy. Concordance measures may, however, be slightly less sensitive to the inclusion of novel risk factors than are R^2-type measures. A related issue is the fact that the different measures have different ranges and interpretations. The concordance indices and the R^2-type measures are constrained to have values that lie between 0 and 1, but in most reasonable applications, the concordance indices will have values that lie between 0.5 and 1. Royston's D index does not have this constraint. The absolute increase in a given concordance index is, in most reasonable applications, bounded by 0.5, while there is no such constraint for D. Thus, if the reference model has a high degree of predictive accuracy, there is limited room for improvement with the addition of a new risk factor.

We have examined a wide range of different performance measures for assessing the performance of a Cox proportional hazards regression model: concordance-type statistics, R^2-type measures, and Royston's D. Several of these measures are analogues for survival outcomes of methods developed for continuous or binary outcomes. For instance, concordance-type measures were based upon the c-statistic for binary outcomes. The c-statistic is equivalent to using the area under the ROC curve for logistic regression models. Its use is not without controversy. Lobo et al. have criticized its use on several grounds, including that it ignores the goodness of fit of the model and that it summarizes the test performance over regions of the ROC space it which one would rarely operate.[43] Similarly, Hand criticizes the ROC curve area for ignoring different misclassification costs.[44] Despite these criticisms, we have considered the c-statistic for survival outcomes, as it is one of the most frequently used performance measures in this context. Furthermore, the relationship between the concordance-type measures and the ROC curve is different in the context of survival analysis than is the case for binary outcomes, where the two measures are identical.

We need to emphasize that performance measures such as concordance and explained variability do not reflect correctness of the underlying model.[45] Some may argue that the pragmatic behaviour of a prediction model is most relevant, i.e. its ability to discriminate events from non-events (c statistics) and provide (low and) high risk predictions for (non-)events. Furthermore, the degree of increase in model performance metric is only one of the many criteria that should be considered when deciding if the new variable should be added. Our examination of the relative sensitivity of the different performance measures should not be taken as a suggestion that the most sensitive measure is necessarily the best measure. We suggest that our findings of relative sensitivity be restricted to within-class comparisons. By doing so, one is restricting comparisons between measures that have the same interpretation and that use the same scale. Despite Harrell's caution described earlier, changes in concordance-type statistics are frequently used to assess the value/usefulness/clinical relevance of adding a novel risk factor to an existing clinical prediction model. Our results indicate that AUC(CD) is slightly more sensitive to inclusion of novel risk factors that is either AUC(H) or AUC(U), which both have nearly identical behaviour to one another, while AUC(GHCI) is the least sensitive of the concordance-type indices. These results suggest that if the focus is on detecting model improvement, then the use of AUC(CD) will have slight benefits. In reviews of different R^2-type measures, Choodari-Oskooei et al. recommended that, amongst the different measures of explained randomness, Kent and O'Quigley's $\rho^2_{w,a}$ be used. We note that Hielscher et al. described this as a measure of predictive accuracy. In the current study, we found that this measure was indeed one of the most sensitive of the R^2-type measures. This observation, together with Choodari-Oskooei et al.'s recommendation, provides further support for its more widespread use.

In summary, of the different concordance indices, Chambless and Diao's concordance index displays changes of a greater magnitude when a novel risk factor is added to an existing reference model. Furthermore, it displayed consistent and stable behaviour across the range of simulated settings. Of the different R^2-type measures, O'Quigley et al.'s modified Nagelkerke R^2 index and Kent and O'Quigley's $\rho^2_{w,a}$ index displays the greatest increase in model performance when a novel risk factor is added to an existing model.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results, and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). Dr Austin is supported in part by a Career Investigator award from the Heart and Stroke Foundation. Dr Steyerberg is supported in part by the Center for Translational Molecular Medicine (PCMM project) and a U-award (AA022802). The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) data used in the study was funded by a CIHR Team Grant in Cardiovascular Outcomes Research. These datasets were linked using unique, encoded identifiers and analysed at the Institute for Clinical Evaluative Sciences (ICES).

References

1. Austin PC, Manca A, Zwarenstein M, et al. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *J Clin Epidemiol* 2010; **63**: 142–153.
2. Chambless LE, Cummiskey CP and Cui G. Several methods to assess improvement in risk prediction models: extension to survival analysis. *Stat Med* 2011; **30**: 22–38.
3. Austin PC and Steyerberg EW. Predictive accuracy of risk factors and markers: a simulation study of the effect of novel markers on different performance measures for logistic regression models. *Stat Med* 2013; **32**: 661–672.
4. Choodari-Oskooei B, Royston P and Parmar MK. A simulation study of predictive ability measures in a survival model II: explained randomness and predictive accuracy. *Stat Med* 2012; **31**: 2644–2659.
5. Choodari-Oskooei B, Royston P and Parmar MK. A simulation study of predictive ability measures in a survival model I: explained variation measures. *Stat Med* 2012; **31**: 2627–2643.
6. Hielscher T, Zucknick M, Werft W, et al. On the prognostic value of survival models with application to gene expression signatures. *Stat Med* 2010; **29**: 818–829.
7. Schemper M and Stare J. Explained variation in survival analysis. *Stat Med* 1996; **15**: 1999–2012.
8. Pencina MJ, D'Agostino RB Sr and Song L. Quantifying discrimination of Framingham risk functions with different survival C statistics. *Stat Med* 2012; **31**(15): 1543–1553.
9. Uno H, Cai T, Pencina MJ, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011; **30**: 1105–1117.
10. Harrell FE Jr. *Regression modeling strategies*. New York, NY: Springer-Verlag, 2001.
11. Steyerberg EW. *Clinical prediction models*. New York: Springer-Verlag, 2009.
12. Pencina MJ and D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004; **23**: 2109–2123.
13. Chambless LE and Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Stat Med* 2006; **25**: 3474–3486.
14. Harrell FE Jr, Lee KL and Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; **15**: 361–387.
15. Gonen M and Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005; **92**: 965–970.
16. Royston P and Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004; **23**: 723–748.
17. Kent J and O'Quigley J. Measures of dependence for censored survival data. *Biometrika* 1988; **75**: 525–534.
18. Korn EL and Simon R. Measures of explained variation for survival data. *Stat Med* 1990; **9**: 487–503.
19. O'Quigley J and Flandre P. Predictive capability of proportional hazards regression. *Proc Natl Acad Sci USA* 1994; **91**: 2310–2314.
20. Akazawa K. Measures of explained variation for a regression model used in survival analysis. *J Med Syst* 1997; **21**: 229–238.
21. O'Quigley J and Xu R. *Handbook of statistics in clinical oncology*. New York: Marcel Dekker, 2001, pp.397–410.

22. Royston P. Explained variation for survival models. *Stata J* 2006; **6**: 1–14.
23. Nagelkerke NJD. A note on a general definition of the coefficient of determination. *Biometrika* 1991; **78**: 691–692.
24. Xu R and O'Quigley J. A measure of dependence for proportional hazards models. *J Nonparametr Stat* 1999; **12**: 83–107.
25. O'Quigley J, Xu R and Stare J. Explained randomness in proportional hazards models. *Stat Med* 2005; **24**: 479–489.
26. Schemper M. The explained variation in proportional hazards regression. *Biometrika* 1990; **77**: 216–218.
27. Graf E, Schmoor C, Sauerbrei W, et al. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999; **18**: 2529–2545.
28. Schemper M and Henderson R. Predictive accuracy and explained variation in Cox regression. *Biometrics* 2000; **56**: 249–255.
29. Schemper M and Kaider A. A new approach to estimate correlation coefficient in the presence of censoring and proportional hazards. *Comput Stat Data Anal* 1997; **23**: 467–476.
30. Harrell FE. The phglm procedure. *SUGI supplemental library users guide*. Cary, NC: SAS Institute, 1986, pp.437–466.
31. Allison PD. *Survival analysis using SAS®: a practical guide*. Cary, NC: SAS Institute, 2010.
32. Cragg JG and Uhler R. The demand for automobiles. *Can J Econ* 1970; **3**: 386–406.
33. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. 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.
34. Lusa L, Miceli R and Mariani L. Estimation of predictive accuracy in survival analysis using R and S-PLUS. *Comput Methods Programs Biomed* 2007; **87**: 132–137.
35. Bender R, Augustin T and Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med* 2005; **24**: 1713–1723.
36. Tu JV, Donovan LR, Lee DS, et al. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. *J Am Med Assoc* 2009; **302**: 2330–2337.
37. Tu JV, Donovan LR, Lee DS, et al. *Quality of cardiac care in Ontario*. Toronto, Ontario: Institute for Clinical Evaluative Sciences, 2004.
38. Lee DS, Austin PC, Rouleau JL, et al. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *J Am Med Assoc* 2003; **290**: 2581–2587.
39. Janssens AC, Moonesinghe R, Yang Q, et al. The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genet Med* 2007; **9**: 528–535.
40. Tzoulaki I, Liberopoulos G and Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *J Am Med Assoc* 2009; **302**: 2345–2352.
41. Austin PC and Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat Methods Med Res* 2017; **26**: 796–808.
42. Steyerberg EW, Harrell FE Jr, Borsboom GJ, et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001; **54**: 774–781.
43. Lobo JM, Jiménez-Valverde A and Real R. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol Biogeogr* 2007; **17**: 145–151.
44. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach Learn* 2009; **77**: 103–123.
45. Austin PC and Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol* 2012; **12**: 82. DOI: 10.1186/1471-2288-12-82.