Journal of Statistical Research *https://doi.org/10.3329/jsr.v58i2.80625*
2024, Vol. 58, No. 2, pp. 369-383 *ISSN 0256 - 422 X*

A SIMULATION STUDY TO ASSESS THE SENSITIVITY OF CONCORDANCE MEASURES TO THE ADDED PREDICTIVE ABILITY IN SURVIVAL PREDICTION MODELS

MALIHA BINTE ALAUDDIN*

Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh

Email: mbalauddin@isrt.ac.bd

M. SHAFIQUR RAHMAN

Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh

Email: shafiq@isrt.ac.bd

SUMMARY

Survival prediction models are often used in healthcare to estimate the prognosis of patients, guide treatment decisions, and allocate resources effectively. When developing a survival prediction model or updating the model with a new predictor or novel marker, it is important to evaluate their performance with measures that facilitate natural and intuitive interpretations and are sensitive to the correct value added by the new predictor. Concordance statistic (C-statistic) is frequently used to assess the predictive performance, especially discriminatory power of the models. Although multiple estimators for C-statistic, such as Harrell's, Uno's and Gonen & Heller's estimators, are available in literature, their performance under different survival data conditions, such as varying levels of censoring, and the added predictive value from a new predictor remains unclear. To address these aspects, this paper first showed an application of some popular C-statistics using two different datasets to describe how these C-statistics can be estimated and interpreted in practice, and secondly investigated their comparative performance using an extensive simulation study. The aim is to evaluate the robustness of these measures to varying degrees of censoring and their sensitivity to the added predictive value of a new predictor in the model, providing practical recommendations for their use. The findings revealed that Gonen & Heller's C-statistic was comparatively more robust to increasing levels of censoring than both Harrell's and Uno's estimators, with Uno's estimator performing moderately better than Harrell's. Additionally, Gonen & Heller's estimator proved to be more sensitive to the added predictive value of a new predictor, regardless of the type of predictor or the level of censoring. The paper concludes with recommendations for selecting the most effective C-statistics to evaluate the performance of survival prediction models across various real-world data scenarios.

Keywords and phrases: Survival predictions; Concordance probability; Cox model

* Corresponding author

1 Introduction

Survival prediction models are frequently used in healthcare research to predict the status of a patient's health outcome such as death, state of illness or recovery from disease in a given time and guide both clinicians and patients for taking joint decisions on the future course of treatment (Moons et al., 2012; Collett, 2014). For example, in oncology, clinicians are often interested to predict the 5-years survival probability of a cancer patient, and in cardiology, to predict the risk of developing cardiac disease etc (Omar et al., 2004). Given the importance of survival prediction models in clinical research, it is essential to evaluate the predictive ability of the models (Royston and Altman, 2013). In particular, it is often of interest to asses how well the predictions obtained from a model match with reality and how much additional value gained by inclusion of a new predictor when updating a model (Rahman and Rumana, 2019). This can be addressed using different arguments, resulting in different metrics for assessing model performance. For prediction models, it is common to assess its predictive performance by quantifying 'discrimination' and 'calibration' (Van Klaveren et al., 2023). Discrimination refers to the model's ability to correctly distinguish the two groups, i.e., subjects who experienced events have higher predicted probabilities compared to those who did not (D'Agostino and Nam, 2003). Calibration refers to "the agreement between the observed and predicted outcomes" (Steyerberg et al., 2010). Of these two aspects, assessing 'discrimination' of a model is relatively more important because re-calibration of the model is possible whereas improving its discrimination is not (Schmid and Potapov, 2012; Pencina et al., 2012b). In the context of developing and validating prediction models for survival data, where the outcome pertains to "time-to-event" and provides information about both the survival duration of subjects and their event status, the presence of censoring further complicates the scenario. Consequently, evaluating the performance of survival prediction models becomes even more challenging.

One commonly used metric to evaluate the performance of survival prediction models, especially for assessing the discriminatory power is the concordance statistic (C-index or C-statistic) (McLernon et al., 2023). The concordance statistic represents the proportion of concordant pairs (where the prediction and observed data are consistent) among all comparable pairs (pairs in which at least one subject experienced the event). A value of 0.5 indicates no discriminatory power, equivalent to random chance, while a value of 1 indicates perfect discrimination by the model (Heller and Mo, 2016). Given the widespread acceptance and straightforward interpretation of the C-statistic, it has been extensively studied in the literature as a measure of performance for survival prediction models, leading to the development of various estimation techniques. Notably, the estimator proposed by Harrell et al. (1982) is "a rank-correlation measure" that brings up a crucial concern about how to order survival times when censoring is present in the data. Later, Harrell's estimator was extensively studied and extended by Uno et al. (2011) incorporating "inverse probability weighting" to account for the effect of censoring, where weight is calculated from the Kaplan-Meier survival probability of censoring distribution. These rank-based approaches have the limitation of not considering the magnitude of the difference in survival times for pairwise comparisons. To address this, Gönen and Heller (2005) introduced an alternative method for estimating concordance probability, specifically for the Cox proportional hazard model (Cox, 1972) by using its properties and solely based on the estimated regression coefficients and the observed covariates.

Although a number of estimators of the concordance statistic have been proposed in the literature, it remains unclear which estimator is generally preferred across all aspects of survival prediction models. These aspects include developing a new prediction model, updating an existing model by adding a new predictor, or assessing the incremental value of a novel biomarker (Pencina et al., 2012a; Newson, 2010). Some studies have devoted considerable attention to examining the robustness of the concordance statistic to sample size and censoring, providing practical recommendations for its use in developing models for survival data (Rahman et al., 2017; Gerds et al., 2014). In prediction research, it is common practice to update existing models by integrating new predictors into them (Collett, 2014; Ohno-Machado, 2001), and as a result, the sensitivity of the concordance statistic to improvements in predictive ability has become an important area of concern (Pencina et al., 2012a; Mihaescu et al., 2010; Austin and Steyerberg, 2012).

However, the C-statistic has faced significant criticism for its lack of sensitivity to the inclusion of new predictors, particularly when updating models (Biswas et al., 2019; Newson, 2010). This insensitivity can lead to misleading conclusions, making the evaluation of the model's performance questionable (Seshan et al., 2013). Additionally, sample size and the presence of censoring in survival data may further distort the estimation of the C-statistic value (Wang and Long, 2016). With this background, the paper aims to evaluate the performance of three widely used C-statistic measures for survival prediction models developed in the Cox proportional hazard (PH) model framework (Cox, 1972). These measures are selected on the basis of their ease of interpretation, communication and availability or ease of implementation in commonly used statistical software. The C-statistic measures are evaluated with respect to the robustness to the degree of censoring and sensitivity to the added predictive value from the inclusion of a new predictor or risk factor and provide some practical recommendations. This is achieved by first applying the C-statistics to two real datasets with different levels of censoring to demonstrate how these estimators can be calculated and interpreted in practice. Subsequently, an extensive simulation study is conducted across diverse scenarios and discussed the recommendations for using them in practice.

2 Methodology

2.1 Notations and the model

Let $(t_i, \delta_i, \boldsymbol{x}_i)$ $(i = 1, 2, \ldots, n)$ represent the observed survival data for the i^{th} subject from a cohort of individuals, where $t_i = \min(T_i, C_i)$, with T_i denoting the failure time and C_i the censoring time, δ_i is the event indicator (1 for failure time and 0 for censoring time), and $\boldsymbol{x}_i$ is a vector of k predictors. The Cox PH model can be defined as

$$h(t_i|\boldsymbol{x}_i) = h_0(t_i)\exp(\boldsymbol{x}_i\boldsymbol{\beta}),$$

where the hazard $h(t_i|\boldsymbol{x}_i)$ at time t is a product of a baseline hazard $h_0(t_i)$ and the exponential of the linear predictor $\boldsymbol{x}_i\boldsymbol{\beta} = \beta_1 x_1 + \ldots + \beta_k x_k$. The predictive form of this model can be written in terms of the survival function as

$$S(t_i|\boldsymbol{x}_i) = S_0(t_i)^{\exp(\boldsymbol{x}_i\boldsymbol{\beta})},$$

where $S(t_i|\boldsymbol{x}_i)$ is the probability of surviving beyond time t given predictors $\boldsymbol{x}$, and $S_0(t_i) = \exp[-\int_0^{t_i} h_0(u)du]$ is the baseline survivor function at time t. To make predictions at time t, one uses estimates $\hat{\boldsymbol{\beta}}$ and $\hat{S}_0(t_i)$.

2.2 Concordance statistics for Cox PH models

Concordance probability is based on the property that a survival model should be able to predict a longer survival time for a subject who fails later in life than the subject who fails earlier (Pencina et al., 2012b). Based on the property, the concordance probability is essentially the fraction of concordant pairs among all comparable pairs. A necessary criterion for comparability is that the pair can be ranked, implying that the subject with shorter observed time can not be a censored observation, as in such scenario it will be ambiguous to determine which subject failed first. A comparable pair is also concordant if the prediction and the observed data go in consistent direction; that is, the model-based survival probability will be higher for the longer event-free individual. Estimating concordance for the Cox PH model requires ranking the survival functions for every comparable pair. This comes down to only comparing the linear predictor values in the case of the Cox PH model because one-to-one correspondence holds between the predicted survival time and the survival probability. For a randomly selected pair of subjects (i, j), the concordance probability can be defined as

$$\begin{aligned} C &= \Pr\big[S_i(t|\boldsymbol{x}_i) < S_j(t|\boldsymbol{x}_j)|t_i < t_j\big] \\ &= \Pr\big[(S_0(t|\boldsymbol{x}_i))^{\exp(\boldsymbol{x_i\beta})} < (S_0(t|\boldsymbol{x}_j))^{\exp(\boldsymbol{x_j\beta})}|t_i < t_j\big] \\ &= \Pr\big[\boldsymbol{x_i\beta} > \boldsymbol{x_j\beta}|t_i < t_j\big]. \end{aligned}$$

The following sub-sections describe the estimation of three concordance measures considered under study.

2.2.1 Harrell's C-statistic

Harrell's estimator involves calculating the proportion of concordance pairs among the comparable pairs (Harrell et al., 1982) and can be estimated as

$$\hat{C}_H = \frac{\sum_{i\neq j}^{n} I(\boldsymbol{x}_i\hat{\boldsymbol{\beta}} > \boldsymbol{x}_j\hat{\boldsymbol{\beta}}|t_i < t_j, \delta_i = 1)}{\sum_{i\neq j}^{n} I(t_i < t_j, \delta_i = 1)},$$

where $I(\cdot)$ denotes the logical indicator function and n is the number of observations. In the presence of censoring, not all subject pairs are comparable; a pair is said to be comparable if the shorter of the two survival times corresponds to an event. This implies that the value of the C_H depends on the censoring mechanism (Pencina et al., 2012b). Another limitation of this estimator is that it uses rank-based approach and disregards the magnitude of differences in survival times when conducting pairwise comparisons.

2.2.2 Uno's C-statistic

To address the shortcoming of Harrell's estimator, Uno et al. (2011) proposed a modified estimator for the C-statistic taking the censoring distribution into account by incorporating "inverse probability weighting" technique (Cheng et al., 1995). The proposed modified estimator can be expressed as

$$\hat{C}_U = \frac{\sum_{i \neq j}^{n} G(t_i)^{-2} I(\boldsymbol{x}_i\hat{\boldsymbol{\beta}} > \boldsymbol{x}_j\hat{\boldsymbol{\beta}} | t_i < t_j, t_i < \tau, \delta_i = 1)}{\sum_{i \neq j}^{n} G(t_i)^{-2} I(t_i < t_j, t_i < \tau, \delta_i = 1)},$$

where $G(t_i)$ is the Kaplan-Meier estimator of the censoring distribution, and τ is introduced to overcome instability in the tail part of the survival function and should be chosen as any time upto and including the last event time.

2.2.3 Gonen & Heller's C-statistic

Gonen & Heller proposed an alternative definition of concordance probability under Cox PH assumption. They argued that for a randomly chosen pair of subjects (i, j), the subject with the higher estimated log relative risk ($\boldsymbol{x}_i\boldsymbol{\beta}$) is expected to have the shorter survival time, assuming that the subjects are ordered according to increasing values of log relative risks derived from the model. Accordingly, the concordance probability can be defined as:

$$\begin{aligned} C(\boldsymbol{\beta}) &= \Pr\left[t_i < t_j | \boldsymbol{x}_i\boldsymbol{\beta} \geq \boldsymbol{x}_j\boldsymbol{\beta}\right] \\ &= \int_0^\infty S(t_j|\boldsymbol{x}_j, \boldsymbol{\beta}) dS(t_i|\boldsymbol{x}_i, \boldsymbol{\beta}) \\ &= \frac{1}{1 + \exp(\boldsymbol{x}_j\boldsymbol{\beta} - \boldsymbol{x}_i\boldsymbol{\beta})}. \end{aligned}$$

Based on the above definition, Gönen and Heller (2005) obtained an analytical formula for concordance probability under Cox PH model, and the estimator is defined as

$$\hat{C}_{GH} = \sum_{i<j} \frac{2}{n(n-1)} \left\{ \frac{I(\boldsymbol{x}_i\hat{\boldsymbol{\beta}} > \boldsymbol{x}_j\hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}_j\hat{\boldsymbol{\beta}} - \boldsymbol{x}_i\hat{\boldsymbol{\beta}})} + \frac{I(\boldsymbol{X}_j\hat{\boldsymbol{\beta}} > \boldsymbol{x}_i\hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}_i\hat{\boldsymbol{\beta}} - \boldsymbol{x}_j\hat{\boldsymbol{\beta}})} \right\}.$$

Unlike Harrell's C_H, Gonen and Heller's estimator, C_{GH}, does not directly use the observed event and censoring times; rather, it is a function of the model parameters and the predictor distribution, using all pairs of patients in its calculation. Since the partial likelihood estimator of $\boldsymbol{\beta}$ from Cox PH model is asymptotically unbiased even for high censoring, the C_{GH} is expected to provide unbiased estimate across different level of censoring.

3 Illustration Using Two Clinical Datasets

This section aims to illustrate an application of the concordance statistics under study using two clinical datasets with different level of censoring and risk profiles. For each dataset, several survival

prediction models were developed under the Cox PH framework with an aim to see whether there is any difference in the estimates across the measures and models with diverse predictive abilities and also to see how these estimates can be interpreted in the clinical contexts. The following subsections describe the illustrations, starting with primary biliary cirrhosis data, followed by prostate cancer data.

3.1 Primary biliary cirrhosis data

Primary biliary cirrhosis (PBC) is a chronic liver disease with limited treatment options, the most effective being liver transplantation. The data under consideration originates from the Mayo Clinic's trial on primary biliary cirrhosis conducted between 1974 and 1984 (Mayo Clinic's official web portal). This study considers the first 312 participants in the dataset who participated in the randomized trial. The outcome of interest is time to death or liver transplantation with 59% censoring. The details of the study design and variables available in the dataset can be found in the work of Murtaugh et al. (1994). For illustration purposes, a base model was first developed using age as the only predictor. Subsequently, the model was updated by sequentially adding the following predictors one by one: serum bilirubin, albumin, and prothrombin time. It is important to note that the predictors were selected based on the literature and exploratory analysis of the data. Since some of the continuous predictors exhibited a non-linear relationship with the outcome, they were log-transformed before being incorporated into the model. For each model, all types of concordance statistics were estimated. The sensitivity to the addition of a new predictor was assessed by calculating the relative changes in the estimates of the concordance statistics. These relative changes were computed for the updated model by comparing its estimates with those from the base model. To account for the effect of censoring, adjustments were made by balancing Harrell's and Uno's estimates, eliminating the difference between their estimates and Gonen and Heller's estimate for the base model.

Table 1: Relative change in C-statistic for inclusion of risk-factors in PBC data. The relative changes were computed by eliminating the difference between the two estimates for the base-model, thereby adjusting for the effect of censoring.

	Harrell's		**Uno's**		**Gonen & Heller's**	
Model	$\hat{C}_H$	Relative Change(%)	$\hat{C}_U$	Relative Change(%)	$\hat{C}_{GH}$	Relative Change(%)
Base model: Age	0.625	-	0.598	-	0.615	-
Added predictor						
Serum bilirubin	0.761	21.7	0.734	22.7	0.751	22.1
Albumin	0.779	24.6	0.752	25.7	0.769	25.0
Prothrombin	0.782	25.1	0.755	26.2	0.772	25.5

Note: Risk-factors were included 'one-at-a-time' to the model.

The results in Table 1 reveal a remarkable difference in the estimates of the three C-statistics for all models, indicating the influence of censoring on the estimates. Further analysis was conducted to assess the sensitivity of the concordance measures to the addition of a predictor to the base model. The results demonstrate differences in the relative changes in the C-statistic values, suggesting that not all C-statistics under study are equally sensitive to the inclusion of a predictor in the base model. Therefore, further investigation through a simulation study is warranted.

3.2 Prostate cancer data

The second illustration is based on the prostate cancer dataset, which is primarily available in the public domain (https://hbiostat.org/data/repo/prostate). The outcome of interest from 502 cancer patients is time-to-death with 29% censoring. The dataset contains a mixture of both demographic and clinical predictors. Similar to the previous illustration using the PBC data, to explore the sensitivity of the concordance measures to the addition of a predictor, we began with a base model containing a single predictor (stage). We then successively added more predictors and computed the relative changes.

Table 2: Relative change in C-statistic for inclusion of risk-factors in prostate cancer data. The relative changes were computed by eliminating the difference between the two estimates for the base-model, thereby adjusting for the effect of censoring.

	Harrell's		**Uno's**		**Gonen & Heller's**	
Model	$\hat{C}_H$	Relative Change(%)	$\hat{C}_U$	Relative Change(%)	$\hat{C}_{GH}$	Relative Change(%)
Base model: Stage	0.693	-	0.653	-	0.654	-
Added predictor						
Serum Hemoglobin	0.733	5.8	0.693	6.1	0.694	6.2
Size of Primary Tumor	0.763	10.1	0.723	10.7	0.724	10.7
Combined Index of Size & Grade	0.772	11.4	0.732	12.1	0.733	12.1

Note: Risk-factors were included 'one-at-a-time' to the model.

The results in Table 2 show that there is difference in the estimates of the three concordance measures for the same model, suggesting the influence of censoring. Furthermore, difference in the changes to adding a predictor across the measures is also observed. These results suggest further investigation through simulation study to identify which concordance measure is actually more sensitive to the inclusion of a new predictor.

4 Simulation Study

Two simulation series were carried out: (i) the first one to examine the impact of censoring under different level of sample size, and (ii) the second one to evaluate the sensitivity of concordance measures to the added predictive value gained by introducing a new predictor to an existing model.

4.1 Simulation design

In the first simulation series, several simulation scenarios were considered varying the level of censoring and sample size $n = 50, 100, 200, 500$. We considered a wide range of censoring rates: 0%, 20%, 50%, and 80% to assess their impact (i.e. for high censoring) on C-statistics for each of the sample size scenarios. For interpretation, we considered censoring rates below 50% as low and those are 50% or above as high, following the discussion in Hendry (2014). In the second simulation series, we evaluated the sensitivity of all the C-statistic measures under study by measuring the additional predictive value gained by increasing the effect size associated with the predictor of the model, following the work of Austin and Steyerberg (2013). While the sensitivity was assessed for one predictor, the value of the other predictor was fixed. The sensitivity was assessed for both the binary and continuous predictors under two different situations: one while they are independent and the other when they are correlated themselves. Further, the sensitivity was assessed for a discrete count-type predictor to see if there is any difference in the performance from the earlier two predictors.

For independent predictors scenario, we generated binary predictor Bernoulli distribution with probability of 0.5 and the continuous predictor was independently generated from standard normal distribution. For the correlated predictors, we generated two variates from a bivariate normal distribution with correlation coefficient $\rho = 0.2$, of which the first variate was considered as continuous predictor and the second variate is considered as a latent variable in the case of binary predictor. The binary predictor was then created from the latent variable, fixing its mean as the threshold value. For the count type predictor, it was generated from the Poisson distribution with mean = 5. For each of the scenarios, the survival times were generated from the Weibull distribution as follows

$$T_i = \Big(\frac{-\log u_i}{\lambda_i}\Big)^{1/\gamma} \quad i = 1, 2, \ldots, n,$$

where $\lambda_i = \exp(\boldsymbol{x}_i\boldsymbol{\beta})$ is a scale parameter with a vector of regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1)$ associated with the predictors and γ as the shape parameter of value 1.2, and u is a uniformly distributed random variable on (0, 1). To introduce random censoring, additional Weibull distributed censoring times were simulated using $C_i = (-\log u_i/\alpha)^{1/\gamma}$ where different choices of the scalar α were used to produce different proportions of censoring. A subject was then considered to be censored if their censoring time was shorter than their survival time. For the first simulation scenarios with different levels of censoring, the values of the regression coefficients were fixed at $\beta_1 = 1.5$ and $\beta_2 = 0.75$. In the second simulation scenarios when sensitivity was assessed for each predictor separately, the value of the regression coefficient associated with the predictor was varied in the interval $(-2, 2)$ fixing the value of the regression coefficients of the other predictors. For each scenario, we

simulated 1000 datasets using the above procedure. For each dataset, we fitted the Cox PH model using the maximum likelihood approach and calculated concordance measures for the fitted model. The average over 1000 simulations was reported for each measure. The sensitivity was measured as the mean change in concordance measures due to the added value of the regression coefficient from the previous value.

4.2 Simulation results

Table 3 shows the results of the first simulation series. The results revealed that all the concordance measures are affected by the degree of censoring, by increasing bias with the increasing degree of censoring. However, the amount of bias due to censoring is quite similar for any sample size scenario considered here. Of the measures, Gonen and Heller's C-statistic (C_{GH}) has shown to be more robust to the degrees of censoring by providing relatively less amount of bias compared to the other measures under study.

Figure 1 shows the sensitivity of the C-statistics for independent binary and continuous predictors separately under 0% and 30% censoring. The results demonstrate that Gonen and Heller's C-statistic appear to show greater sensitivity compared to the other measures for both binary and continuous predictors across all scenarios of censoring. However, Harrell's and Uno's C-statistics (C_H, C_U) showed very similar performance. Gonen and Heller's C-statistic (C_{GH}) also showed greater sensitivity for the scenario with correlated predictors under 0% and 30% censoring as shown in Figure 2, however, the performance of the C_{GH} almost close to the other measures. Similar results appear when sensitivity was assessed for an independent count-type predictor in Figure 3.

5 Discussion and Conclusion

The paper assessed the performance of some concordance metrics for survival prediction models, focusing on their robustness to the degree of censoring and sensitivity to the added predictive value due to the inclusion of a new predictor with the aim to provide some practical recommendations. The paper first showed an application of the concordance measures to two clinical datasets, each with different levels of censoring and predictive value, to illustrate the performance of these metrics. The results showed that these concordance measures performed differently for the same model developed from a given dataset. This led us to conduct a simulation study to determine which measures actually outperform others and under what conditions in censored survival data.

The simulation study revealed that all concordance measures were affected by censoring, showing increasing bias as the level of censoring increased. Among the measures, Gonen & Heller's estimator outperformed the others by exhibiting a negligible amount of bias compared to Harrell's and Uno's estimators, regardless of the level of censoring, indicating its robustness to censoring. These results held true for any sample size, and sample size did not affect the bias. When examining sensitivity, Gonen & Heller's estimator exhibited the highest sensitivity to the added predictive value of a new predictor. Harrell's and Uno's estimators were nearly identical at 0% censoring but showed slight differences at moderate levels of censoring (30%).

Table 3: C-statistic for varying percentage of censored data and sample size with true value of concordance statistic 0.791

		Harrell's			**Uno's**			**Gonen & Heller's**		
Sample size	Cens (%)	$\hat{C}_H$	SE	Rel. bias(%)	$\hat{C}_U$	SE	Rel. bias(%)	$\hat{C}_{GH}$	SE	Rel. bias(%)
50	0	0.792	0.034	0.126	0.792	0.034	0.126	0.794	0.030	0.379
	20	0.826	0.033	4.425	0.826	0.033	4.425	0.820	0.028	3.666
	50	0.887	0.030	12.137	0.892	0.030	12.769	0.851	0.026	7.585
	80	0.954	0.023	20.607	0.959	0.031	21.239	0.897	0.035	13.401
100	0	0.792	0.022	0.126	0.792	0.022	0.126	0.792	0.020	0.126
	20	0.822	0.024	3.919	0.823	0.023	4.046	0.814	0.020	2.908
	50	0.884	0.021	11.757	0.891	0.022	12.642	0.845	0.018	6.827
	80	0.951	0.016	20.228	0.957	0.024	20.986	0.884	0.023	11.757
200	0	0.792	0.016	0.126	0.792	0.016	0.126	0.792	0.014	0.126
	20	0.822	0.016	3.919	0.823	0.016	4.046	0.814	0.014	2.908
	50	0.883	0.015	11.631	0.891	0.015	12.642	0.842	0.013	6.448
	80	0.950	0.011	20.101	0.958	0.015	21.113	0.876	0.015	10.746
500	0	0.791	0.009	0.000	0.791	0.009	0.000	0.791	0.008	0.000
	20	0.822	0.010	3.919	0.823	0.010	4.046	0.813	0.009	2.781
	50	0.882	0.010	11.504	0.891	0.010	12.642	0.840	0.008	6.195
	80	0.949	0.007	19.975	0.959	0.009	21.239	0.870	0.010	9.987

Note: The reported results are mean of 1000 iterations.

The simulation findings align with the results from both illustrations using the two clinical datasets. For instance, Gonen and Heller's estimator demonstrated slightly greater sensitivity to the addition of new predictors when updating an existing model, which is consistent with the simulation findings under the scenario with 30% censoring, particularly in the case of correlated predictors. Both the simulation and real-data applications suggest that the level of censoring influences the degree of sensitivity in the concordance measures. Therefore, the level of censoring should be carefully investigated before selecting an appropriate concordance measure.

The reason for Gonen & Heller's outperformance is that it is a model-based estimator, i.e., it is a function of the regression coefficient, and retains all desirable statistical properties as long as

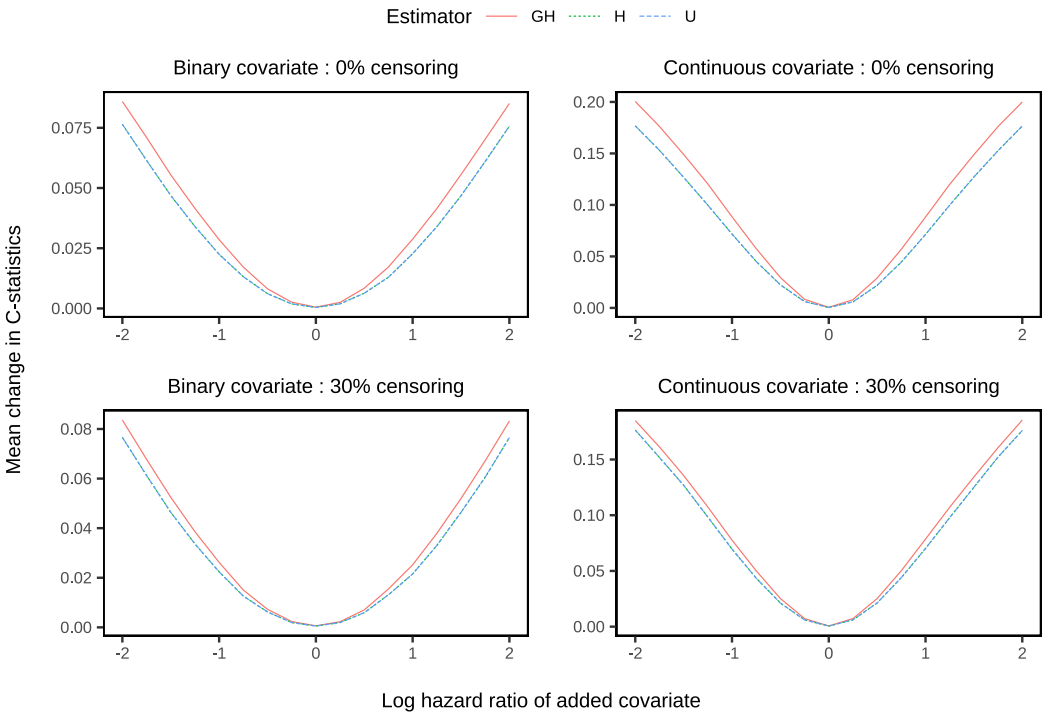

Figure 1: Sensitivity of C-statistics to the added value of the regression coefficient associated with a predictor that is independent of the other predictor in the model. The results are summarized for both 0% and 30% censoring.

the Cox PH model is correctly specified and estimated. As the effect of censoring on the regression coefficient is mediated through the partial-likelihood estimation, this concordance estimator is not affected by the censoring (Gönen and Heller, 2005). Moreover, since the estimator is completely dependent on the regression coefficient, larger changes in the coefficient result in greater sensitivity in the concordance estimator. In contrast, both Harrell's and Uno's estimators are based on the ranking of predicted risk given the observed data. Therefore, the rank order is less affected by changes in the predicted risk derived from the linear combination of the predictors and their coefficients(Rahman et al., 2017). However, the only common characteristic available among these concordance measures is that all these measures can be easily computed and programs or codes are available in the standard statistical softwares such as R and Stata.

Based on the findings of the study, some practical guidelines for selecting concordance measures when developing a survival prediction model are discussed as follows: Before selecting a concordance measure, it is recommended to assess the level of censoring in the data and determine whether a Cox proportional hazards (PH) model can be appropriately fitted while holding the PH assumption. If the Cox PH model is suitable for the data, we strongly recommend using Gonen & Heller's

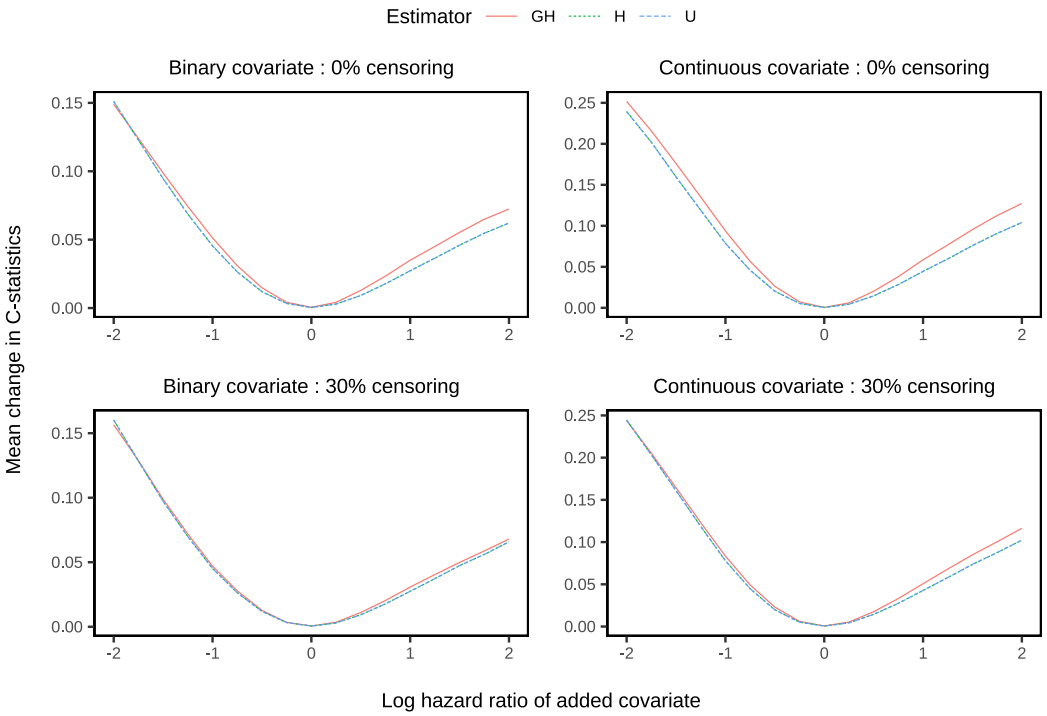

Figure 2: Sensitivity of C-statistics to the added value of the regression coefficient associated with a predictor that is correlated with the existing predictor in the model. The results are summarized for both 0% and 30% censoring.

C-statistic at any level of censoring. This recommendation is based on the estimator's robustness to censoring and its high sensitivity to the added predictive value from the model. A concordance measure with good sensitivity is particularly important when an existing model is being updated with new risk factors. In contrast, Uno's estimator is recommended if the Cox PH assumption does not hold and the predictive model is developed using a survival model framework other than the Cox PH framework.

Acknowledgements

The authors acknowledge Frank Harrell for making available both the primary biliary cirrhosis and prostate cancer datasets in the public domain under the department of biostatistics, Vanderbilt University, USA. Both datasets are available for free download at https://hbiostat.org/data/ under the authority of the Department of Biostatistics, Vanderbilt University, USA. The datasets are freely accessible in the public domain and may be used in research publications; the authority that made the data publicly available has approved the ethics approval and consent statement.

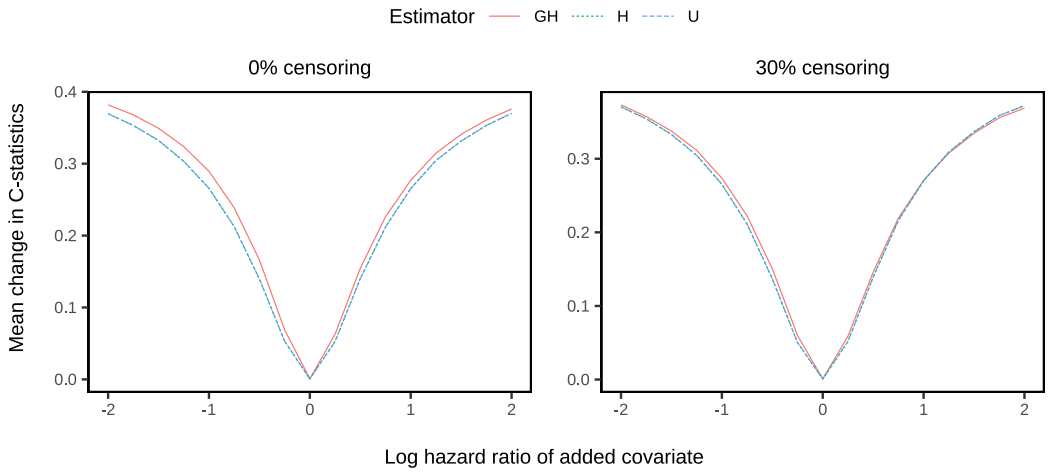

Figure 3: Sensitivity of C-statistics to the added value of the regression coefficient associated with a count-type predictor in the model. The results are summarized for both 0% and 30% censoring.

References

Austin, P. and Steyerberg, E. (2012), "Interpreting the Concordance Statistic of a Logistic Regression Model: Relation to the Variance and Odds Ratio of a Continuous Explanatory Variable," *BMC Medical Research Methodology*, 12, 82.

Austin, P. C. and Steyerberg, E. W. (2013), "Predictive accuracy of risk factors and markers: a simulation study of the effect of novel markers on different performance measures for logistic regression models," *Statistics in medicine*, 32, 661–672.

Biswas, B., Husain, M., and Rahman, M. S. (2019), "Review and evaluation of the concordance measures for assessing discrimination in the logistic regression methods," *Journal of Statistical Research*, 53, 63–77.

Cheng, S., Wei, L. J., and Ying, Z. (1995), "Analysis of transformation models with censored data," *Biometrika*, 82, 835–845.

Collett, D. (ed.) (2014), *Modelling Survival Data in Medical Research (3rd ed.)*, Chapman and Hall/CRC.

Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 187–220.

D'Agostino, R. B. and Nam, B.-H. (2003), "Evaluation of the performance of survival analysis models: discrimination and calibration measures," *Handbook of statistics*, 23, 1–25.

Gerds, T., Kattan, M., Schumacher, M., and Yu, C. (2014), "Concordance for Prognostic Models with Competing Risks," *Biostatistics*, 15, 526–536.

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

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982), "Evaluating the yield of medical tests," *Jama*, 247, 2543–2546.

Heller, G. and Mo, Q. (2016), "Estimating the concordance probability in a survival analysis with a discrete number of risk groups," *Lifetime data analysis*, 22, 263–279.

Hendry, D. J. (2014), "Data generation for the Cox proportional hazards model with time-dependent covariates: a method for medical researchers," *Statistics in Medicine*, 33, 436–454.

McLernon, D. J., Giardiello, D., Van Calster, B., Wynants, L., van Geloven, N., van Smeden, M., Therneau, T., Steyerberg, E. W., topic groups 6, and of the STRATOS Initiative, . (2023), "Assessing performance and clinical usefulness in prediction models with survival outcomes: practical guidance for Cox proportional hazards models," *Annals of internal medicine*, 176, 105–114.

Mihaescu, R., Van Zitteren, M., Van Hoek, M., Sijbrands, E. J., Uitterlinden, A. G., Witteman, J. C., Hofman, A., Hunink, M. M., Van Duijn, C. M., and Janssens, A. C. J. (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.

Moons, K. G., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G., and Woodward, M. (2012), "Risk prediction models: II. External validation, model updating, and impact assessment," *Heart*, 98, 691–698.

Murtaugh, P. A., Dickson, R. E., Van Dam, G. M., Malinchoc, M., Grambsch, P. M., Langworthy, A. L., and Gips, C. H. (1994), "Primary biliary cirrhosis: prediction of short–term survival based on repeated patient visits," *Hepatology*, 20, 126–134.

Newson, R. B. (2010), "Comparing the predictive powers of survival models using Harrell's C or Somers' D," *The Stata Journal*, 10, 339–358.

Ohno-Machado, L. (2001), "Modeling medical prognosis: survival analysis techniques," *Journal of biomedical informatics*, 34, 428–439.

Omar, R. Z., Ambler, G., Royston, P., Eliahoo, J., and Taylor, K. M. (2004), "Cardiac surgery risk modeling for mortality: a review of current practice and suggestions for improvement," *The Annals of thoracic surgery*, 77, 2232–2237.

Pencina, M. J., D'agostino, R. B., Pencina, K. M., Janssens, A. C. J., and Greenland, P. (2012a), "Interpreting incremental value of markers added to risk prediction models," *American journal of epidemiology*, 176, 473–481.

Pencina, M. J., D'Agostino Sr, R. B., and Song, L. (2012b), "Quantifying discrimination of Framingham risk functions with different survival C statistics," *Statistics in medicine*, 31, 1543–1553.

Rahman, M. S., Ambler, G., Choodari-Oskooei, B., and Omar, R. Z. (2017), "Review and evaluation of performance measures for survival prediction models in external validation settings," *BMC medical research methodology*, 17, 1–15.

Rahman, M. S. and Rumana, A. S. (2019), "A model-based concordance-type index for evaluating the added predictive ability of novel risk factors and markers in the logistic regression models," *Journal of Applied Statistics*, 46, 2145–2163.

Royston, P. and Altman, D. G. (2013), "External validation of a Cox prognostic model: principles and methods," *BMC medical research methodology*, 13, 1–15.

Schmid, M. and Potapov, S. (2012), "A comparison of estimators to evaluate the discriminatory power of time-to-event models," *Statistics in medicine*, 31, 2588–2609.

Seshan, V. E., Gönen, M., and Begg, C. B. (2013), "Comparing ROC curves derived from regression models," *Statistics in medicine*, 32, 1483–1493.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010), "Assessing the performance of prediction models: a framework for some traditional and novel measures," *Epidemiology (Cambridge, Mass.)*, 21, 128.

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.

Van Klaveren, D. et al. (2023), "Methodological Concerns about Concordance-Statistic for Benefit as a Performance Measure for Treatment Benefit Prediction Algorithms," *Diagnosis and Prognosis*, 4, 45–55.

Wang, M. and Long, Q. (2016), "Addressing issues associated with evaluating prediction models for survival endpoints based on the concordance statistic," *Biometrics*, 72, 897–906.

Received: October 20, 2024

Accepted: March 7, 2025